# Models Versus Satisfaction: Towards a Better Understanding of Evaluation Metrics*

Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, Shaoping Ma†
Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

## ABSTRACT

Evaluation metrics play an important role in the batch evaluation of IR systems. Based on a user model that describes how users interact with the rank list, an evaluation metric is defined to link the relevance scores of a list of documents to an estimation of system effectiveness and user satisfaction. Therefore, the validity of an evaluation metric has two facets: whether the underlying user model can accurately predict user behavior and whether the evaluation metric correlates well with user satisfaction. While a tremendous amount of work has been undertaken to design, evaluate, and compare different evaluation metrics, few studies have explored the consistency between these two facets of evaluation metrics. Specifically, we want to investigate whether the metrics that are well calibrated with user behavior data can perform as well in estimating user satisfaction. To shed light on this research question, we compare the performance of various metrics with the C/W/L Framework in estimating user satisfaction when they are optimized to fit observed user behavior. Experimental results on both self-collected and public available user search behavior datasets show that the metrics optimized to fit users' click behavior can perform as well as those calibrated with user satisfaction feedback. We also investigate the reliability in the calibration process of evaluation metrics to find out how much data is required for parameter tuning. Our findings provide empirical support for the consistency between user behavior modeling and satisfaction measurement, as well as guidance for tuning the parameters in evaluation metrics.

## CCS CONCEPTS

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

## KEYWORDS

evaluation metrics, user models, user satisfaction

**ACM Reference Format:**
Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, 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 (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401162

## 1 INTRODUCTION

As batch evaluation plays a central part in IR research, how to design and meta-evaluate different evaluation metrics have been widely studied for many years. Originated from Cranfield paradigm [12], a lot of evaluation metrics are proposed to compare the effectiveness of different systems based on test collections with relevance judgments. With the wide usage of Web search engines in our daily life, user-centric evaluation has also drawn much attention recently. This is mainly reflected in two aspects. On one hand, different user models which describe how simulated users interact with search systems are developed for evaluation metrics. Moffat and Zobel [32] first encodes an explicit user model in Rank-Biased Precision (RBP) which assumes that users will examine the next result with a certain probability. Considering different hypotheses and constraints, more sophisticated metrics have been proposed in this line of research. Further, Moffat et al. [30, 31] introduce the C/W/L Framework to characterize user models with the probabilities of three interrelated behaviors: viewing, continuing, and stopping behaviors. This formalize the connection between evaluation metrics and user models. On the other hand, the concept of satisfaction has also become one of the major concerns in search evaluation. User satisfaction can be regarded as the fulfillment of a specified desire or goal [22]. Although some researchers argue that user satisfaction alone does not provide reliable evaluation [16, 26], it provides a measurement of users' subjective feelings and experiences about their search processes. Recent studies investigate the relationship between evaluation metrics and user satisfaction [1, 17, 21]. Their findings demonstrate that the performance of a system measured by evaluation metrics has a strong correlation with user satisfaction. Therefore, user satisfaction is widely regarded as the gold standard in search evaluation related work.

From the perspective of user-centric evaluation, a valid evaluation metric should not only correspond to a user model which can accurately predict user behavior, but also correlate well with

*This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011, 61902209) and Beijing Academy of Artificial Intelligence (BAAI). Dr Weizhi Ma has been supported by Shuimu Tsinghua Scholar Program.
†Corresponding Author: Yiqun Liu


*SIGIR '20, July 25–30, 2020, Virtual Event, China*

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401162

user satisfaction. However, to our best knowledge, few studies have explored the consistency between these two facets of evaluation metrics. A recent work [43] attempts to construct a novel framework which considers the accuracy of user model and the correlation with user satisfaction for meta-evaluation of metrics. They show how the accuracy of user model and the correlation between metric and satisfaction change with parameter values in session level. Nevertheless, further analysis and deeper insight should be provided to better understand the relationship between the two facets of evaluation metrics. Considering that most existing metrics are designed in query level, in this paper, we investigate whether the metrics that are calibrated with user behavior data based on a number of queries can simultaneously perform well in estimating user satisfaction for a broader set of queries. Specifically, we attempt to answer the following research questions:

- **RQ 1:** Are the metrics calibrated with user behavior data consistent with those optimized to fit user satisfaction feedback in evaluating the performance of search systems?
- **RQ 2:** Are the parameters of metrics stable while being calibrated with different samples of user behavior data?
- **RQ 3:** How much data is enough for an evaluation metric to tune its parameters to correlate well with user satisfaction?

To shed light on the above research questions, we compare the performance of various metrics with the C/W/L Framework in estimating user satisfaction when they are optimized to fit observed user behavior. To meta-evaluate metrics in a more realistic search scenario, besides a public available search user behavior dataset, we also conduct a field study which collects daily search logs and explicit feedbacks from 30 participants for one month. Experimental results on both two dataset show that the metrics calibrated with users' click behavior can perform as well as those optimized to fit user satisfaction feedback. We further investigate the reliability and data requirements for the calibration of evaluation metrics. Our findings demonstrate that calibrating evaluation metrics with an appropriate scale of user behavior data is an effective and feasible way to evaluate search systems. In summary, we make the following contributions:

- We thoroughly study the consistency between accuracy of user model and correlation with user satisfaction in a more fine-grained manner.
- We delve into the stability and reliability of parameters encoded in different user models and formally study the scale of user behavior data used to optimize these parameters.
- We conduct a field study from which we obtain a practical dataset that will be publicly available upon publication of the paper.

The remainder of this paper is organized as follows. In Section 2, we review a broad range of related studies about user models of metrics, user satisfaction, and meta-evaluation of metrics. Then we describe our methods to meta-evaluate user models and user satisfaction for metrics in Section 3. Section 4 and Section 5 shows some details of our data collection process and experimental settings, respectively. The experiments and results are shown in Section 6 to investigate the relationship between user models and user satisfaction. Finally, we discuss the conclusions and limitations of our work in Section 7.

# 2 RELATED WORK

## 2.1 User Models of Metrics

Evaluation metrics encapsulate assumptions about user behavior [14, 31]. It has been shown that most advanced evaluation metrics are fundamentally related and are underlied by different user behavior models [9, 30]. Further, within the C/W/L framework proposed by Moffat et al. [31], these user behavior models can be described by the continuation probability / weight function / last probability of search users while the probabilities will be affected by different aspects [31, 32]. These metrics mainly follow an assumption that users scan ranked results from top to bottom before they stop [13]. Järvelin and Kekäläinen [18] propose normalized discounted cumulative gain (NDCG), which formalizes user gain from a result list as a discounting process where the continuation probability declines along with the increasing ranks of results. Arguing that the discount function of NDCG only implicitly connects to a user model and do not characterize how users actually interact with search systems, Moffat and Zobel [32] propose rank-biased precision (RBP), which explicitly encodes a user stopping model. It assumes that users examine the $(i+1)$-th result after examining the $i$-th result with persistence $\theta$ and will end their examination with probability $1-\theta$. Besides considering the position impact, Expected Reciprocal Rank (ERR) [10] takes result relevance into consideration and defines the probability that a user is satisfied with a document to be related with relevance of the document. Rather than assuming that users continue to examine the next result with a fixed probability distribution of user behavior, recently developed adaptive metrics assume that the continuation/stopping probability is influenced by the previous gains and expectations of users. Examples include INST [5], Bejeweled Player Model (BPM) [49] and Information Foraging Based Measure (IFT) [3].

While the aforementioned metrics are effective, it is essential to determine practical parameters for the underlying user behavior models (e.g., persistence $\theta$ in RBP) [42, 46]. In this paper, we investigate parameters that calibrated with underlying user behavior models in terms of stability and reliability.

## 2.2 User Satisfaction

User satisfaction can be understood as the fulfillment of a specified desire or goal in information retrieval [22]. It measures users' actual feelings about a system and can be considered as the golden standard in search performance evaluation [1, 17]. Ali and Beg [2] indicate that explicit judgments of actual users can lead to a more realistic evaluation of system performance.

Existing work delves into the relationship between user satisfaction and user behavior [23, 28, 41, 47]. For example, Kim et al. [23] demonstrate that click-level satisfaction has a strong correlation with click dwell time with which user satisfaction can be predicted. [28] investigate relationship between mouse movement information on search engine result pages (SERPs) and user satisfaction. They also build a model to predict user satisfaction based on the motif which is the frequently-appeared sequence of mouse positions. In addition, it has been found that duration of completing a search task has a negative correlation with user satisfaction [47]. Correlation with actual user satisfaction is often considered to be the ultimate test for newly proposed evaluation metrics. [1] show

that user satisfaction is strongly correlated with some evaluation metrics such as CG and DCG. There exists a number of studies investigating different evaluation methods and the correlation between these methods and satisfaction [28, 31, 38].

Our contributions in this paper complement existing work on thoroughly investigating the consistency between two facets of evaluation metrics, i.e., user behavior prediction and user satisfaction reflection.

### 2.3 Meta-Evaluation of Metrics

The meta-evaluation of evaluation metrics has been widely studied in recent years and different criteria (e.g., error rate, discriminative power and etc.) have been proposed to compare different evaluation metrics. Buckley and Voorhees [7] adopt "error rate" which is the likely error of concluding "System A is better than system B", to compare between different metrics. Sakai [35] propose "Discriminative power" which shares similar intuitive with "fuzziness value" proposed in [7]. "Discriminative power" refers to the power of a measure to discriminate among systems. Another criteria to compare metrics is to calculate the correlations between system orderings generated by different metrics [37]. Other methods such as statistical ability [48], judgment cost [8] have also been studied. Chen et al. [11] adopt user satisfaction as the ground truth for evaluating different evaluation metrics, that is, examining whether the metric under consideration has a strong positive correlation with user satisfaction, which are usually generated after the users have completed a search. Wicaksono and Moffat [42] also consider metrics used in evaluation as a surrogate for user satisfaction. They investigate the relationship between model accuracy and user satisfaction to explore effectiveness metrics.

In this paper we follow the same principle introduced in [42] and investigate relationship between model accuracy and correlation with satisfaction. What we add on top of prior work on meta-evaluation of metrics is that we conduct a more fine-grained analysis of consistency between accuracy of user model and user satisfaction in query-level evaluation.

## 3 METHODOLOGY

As we have mentioned before, the validity of an evaluation metric reflects in two facets, which are user behavior modeling and satisfaction measurement. Therefore, in this section, we describe how to evaluate these two facets respectively.

### 3.1 Accuracy of User Models

Following previous work [3, 43], we measure the accuracy of an underlying user model by estimating how well the model predicts users' actual behavior. In the view of the C/W/L Framework, user models are characterized with any of three different but interrelated ways: Continuation (C) probability, Weight (W) function and Last (L) probability. Therefore, an intuitive method to measure the accuracy of a user model is to calculate the distances between the probability distributions of $C(\cdot)$, $W(\cdot)$, and $L(\cdot)$ defined by the user model and their corresponding observed distributions, denoted as $\hat{C}(\cdot)$, $\hat{W}(\cdot)$, and $\hat{L}(\cdot)$.

$C(\cdot)$, $W(\cdot)$, and $L(\cdot)$ are usually determined given a user model. For some adaptive metrics, the computation of $C(\cdot)$, $W(\cdot)$, and $L(\cdot)$ is related to gains and costs of results that users have encountered so far. Therefore, we average the values of $C(\cdot)$, $W(\cdot)$, and $L(\cdot)$ across all rank lists in the dataset for adaptive user models. In this paper, we use relevance judgments to measure the gains of results. In addition, we assume that the costs of different results are the same and omit this factor for simplicity. These simple assumptions are also adopted by some previous studies [43, 49].

Different from probability distributions defined by user models, $\hat{C}(\cdot)$, $\hat{W}(\cdot)$, and $\hat{L}(\cdot)$ need to be estimated from observed user behavior. Following previous research [43], we estimate $\hat{C}(\cdot)$, $\hat{W}(\cdot)$, and $\hat{L}(\cdot)$ as follows:

$$\hat{C}(i) = \frac{\sum_{u\in U}\sum_{q\in Q(u)}\hat{P}(view = i+1|u,q)}{\sum_{u\in U}\sum_{q\in Q(u)}\hat{P}(view = i|u,q)} \tag{1}$$

$$\hat{W}(i) = \frac{\sum_{u\in U}\sum_{q\in Q(u)}\hat{P}(view = i|u,q)}{\sum_{u\in U}\sum_{q\in Q(u)}\sum_{j=1}^{N}\hat{P}(view = j|u,q)} \tag{2}$$

$$\hat{L}(i) = \frac{\sum_{u\in U}\sum_{q\in Q(u)}\hat{P}(view = i|u,q) - \hat{P}(view = i+1|u,q)}{\sum_{u\in U}\sum_{q\in Q(u)}\hat{P}(view = 1|u,q)} \tag{3}$$

where $\hat{P}(view = i|u,q)$ is the estimated probability that user $u$ views the item listed at rank $i$ for query $q$. $U$ is the set of users and $Q(u)$ is the set of queries associated with user $u$. To estimate $\hat{P}(view = i|u,q)$, we need to know examination sequences of users. Since it is difficult and expensive to collect eye-tracking signals from users, click-through logs are usually taken as a surrogate to infer users' gaze distributions. Inspired by a recent work [44], $\hat{P}(view = i|u,q)$ is estimated as follows:

$$\hat{P}(view = i|u,q) = \begin{cases} 1 & i \le DC(u,q) \\ e^{-(i-DC(u,q))/g(K(u,q))} & otherwise \end{cases} \tag{4}$$

where $K(u,q) = w_0 + w_1 \cdot DC(u,q) + w_2 \cdot NC(u,q)$. $DC(u,q)$ is the deepest rank position clicked and $NC(u,q)$ is the number of distinct items clicked. $g(x) = \ln(1+e^x)$ is a softplus function. Using eye-tracking data with the J&A dataset [19], Wicaksono et al. [43] fit the function of $K(u,q)$ and show $K(u,q) = 3.48 - 0.46 \cdot DC(u,q) + 0.20 \cdot NC(u,q)$. In our experiments, we also use this function to estimate $\hat{P}(view = i|u,q)$ and the observed probability distributions $\hat{C}(\cdot)$, $\hat{W}(\cdot)$, and $\hat{L}(\cdot)$. Further, considering that the parameters of $K(u,q)$ calibrated with the J&A dataset may not be suitable for other datasets and sometimes we do not have eye-tracking data to fit the function, we employ last click signal as an alternative way to estimate $\hat{P}(view = i|u,q)$. Similar to another work [3], given the last click position, $\hat{P}(view = i|u,q)$ is set to be 1 for previous results and 0 otherwise. We denote this as a hard (H) method to estimate $\hat{P}(view = i|u,q)$ and Equation 4 as a soft (S) method.

Given the model-derived probability distributions $C(\cdot)$, $W(\cdot)$, and $L(\cdot)$, as well as the estimations of the observed probability distributions $\hat{C}(\cdot)$, $\hat{W}(\cdot)$, and $\hat{L}(\cdot)$, we can measure the accuracy of user models by calculating mean squared error (MSE) or weighted mean squared error (WMSE) functions of the corresponding probability distributions. Note that the weighting is required for $C(\cdot)$ because it is a set of independent values at different rank positions. We compare the effectiveness of different estimation methods with different probability distributions and different signals, which are denoted as S_C, S_W, S_L, H_C, H_W, and H_L, respectively. For example,

S_C (Soft and Continuation) means we measure the accuracy of user models by calculating WMSE of $C(\cdot)$ and $\hat{C}(\cdot)$ with estimation of view probability in Equation 4.

### 3.2 Correlation between Evaluation Metrics and User Satisfaction

Regarding user satisfaction as the gold standard in search evaluation, the correlation between scores of evaluation metrics and user satisfaction is widely used to compare the performance of different metrics by a range of studies [11, 20, 25, 50]. Among these studies, Pearson correlation coefficient (Pearson's $r$ [33]) is the most popular to be used to measure the correlation. Besides Pearson's $r$, Chen et al. [11] also conduct concordance test [36] to compare different metrics and conclude that a small number of data pairs may yield discrete and unreliable results for concordance test. Given user satisfaction feedback is discrete in our experiments, in this paper, we mainly use Spearman correlation coefficient (Spearman's $\rho$ [39]) to measure the relationship between evaluation metrics and user satisfaction. Spearman's $\rho$ is a nonparametric measure of rank correlation which assesses monotonic relationships rather than linear relationships as Pearson's $r$. We also compare the results by using Spearman's $\rho$ and Pearson's $r$ in our experiments and draw similar conclusions with respect to our research questions. Therefore, we only report the results with Spearman's $\rho$ in our experiments due to the limitation of space.

## 4 DATA COLLECTION

To meta-evaluate metrics in terms of the consistency between the accuracy of user models and correlation with satisfaction, we conduct a field study which has been proposed to overcome the limitations of lab studies and large-scale log analysis [15, 45]. 30 participants are involved in our field study for one month. These participants include 13 females and 17 males whose ages range from 18 to 41. Realistic search data including search behavior and user feedback are collected. In what follows, we provide the details of data collection procedure as well as data used for following experiments.

### 4.1 Field study

We follow the same principle introduced in [15, 45] and form our dataset using the following procedure:

(1) *Introduction stage*: Participants are instructed about the requirements of our field study. They are asked to install a browser extension on their laptops, which records their daily web search activities while they can turn it on or off anytime. To make participants familiar with concepts and process used in this study, we provide two training sessions after a detailed introduction of our field study.
(2) *Data collection stage*: After the training process, participants can use their laptops to perform searching for daily purposes as usual. During a one-month period, We collect two search data including user behavior data and user feedback data using the pre-installed extension. **User behavior data**: We record their issued queries and corresponding SERPs and landing pages. Specifically, URLs and HTML content of aforementioned pages are recorded. Besides content data, participants' mouse interactions such as mouse movement, clicks and scrolling and temporal information including dwell time on SERPs and landing pages are also recorded. **User feedback data**: They are asked to review their search history and provide feedbacks. To protect the privacy, they are allowed to freely discard any log that they are not willing to share with us. Since we seek to investigate the relationship between the accuracy of user models and correlation with user satisfaction at query level, we ask participants to provide query-level search satisfaction. We also ask them to provide underlying search goals which are used for further annotations. The overall search feedbacks we collected in the field study are shown in Table 1. Compared to previous field study [45], we focused more on users' query-level feedbacks.
(3) *Summarization stage*: After the data collection stage, participants were paid based on their contributions: about $5 for participating our field study and $0.15 for each valid search query log they provided. We also collect their suggestions and feedback for our field study with a post-experiment questionnaire. Most participants were satisfied with the design of our field study and felt free for providing search logs and feedbacks because they were allowed to remove any logs that they were not willing to share.

### 4.2 Collected Data

After filtering invalid data (parts of user behaviors or page contents are not recorded successfully), we obtain a field study dataset where 3,875 queries and corresponding user behavior and user feedback data are available.

To calculate metric score of a given evaluation metric, absolute judgments (i.e., relevance judgments or usefulness feedbacks) for search results are essential. Although we have collected usefulness feedbacks from users in our field study, it is usually infeasible and expensive to collect them in realistic scenarios to evaluate the performance of search systems. To this end, we also recruit nine external assessors to make relevance judgments for search results in our field study. For each "query-document" pair, assessors are provided with the underlying search goals (collected during the field study) for the query to help them better understand the information need behind the given query. We adopt a 4-point relevance annotations, following the rating criteria used in [27]:

- *Irrelevant.* The result is not relevant or a spam page.
- *Somewhat relevant.* The result only provides minimal information about the query.
- *Fairly relevant.* The result provides substantial information about the query.
- *Highly relevant.* The result is dedicated to the query, it is worthy of being a top result in a web search engine.

Given that the user study dataset [11] only contains relevance judgments, in this paper, we only utilize relevance judgments for experiments on both two datasets. Usefulness feedbacks are left for future research.

In all, we collect relevance judgements for all 37,550 search results of the given 3,875 queries. Each "query-document" pair is annotated by three assessors and a median score is used to aggregate relevance scores provides by different assessors. The Fleiss's $\kappa$

**Table 1: Search feedback information collected in the field study.**

| | Attribute | Description | Value |
|---|---|---|---|
| **Task** | Background | Please describe the time, location and your intention for this search task. | open-ended question |
| | Satisfaction | Were you satisfied with the process of searching for completing this task? | (0) unsatisfied - (4) very satisfied |
| | Success | How much useful information have you found for this task? | (0) not any - (4) all you want |
| | Difficulty | How do you feel about the difficulty of this task to find relevant information? | (0) easy - (4) extremely difficult |
| **Query** | Expectation | What information did you expect to find for this query? | open-ended question |
| | Relation | What is the relation between this query B and the last query A? | (0) initial query; (1) substitute/(2) add/(3) delete terms for the same topic; (4) B is a subtopic of A; (5) B and A are two subtopics of a same topic; (6) B is a new topic related to A. |
| | Satisfaction | Were you satisfied with the search results for this query? | (0) unsatisfied - (4) very satisfied |
| | Reason for Leaving | Why do you reformulate this query or end your search? | (A) have found enough information; (B) come up with a better query; (C) cannot find useful information; (D) other reason___ |
| **Result** | Usefulness | For each result, how do you rate its usefulness for completing this search task? | (0) useless - (4) highly useful |

of relevance judgments is 0.693, showing a substantial agreement between assessors [24]. All the collected data used in this paper, including the relevance judgments, is available online for academic research.[1]

## 5 EXPERIMENTAL SETTINGS

In this section, we introduce the datasets and evaluation metrics we used in our experiments. We also show the settings of parameter calibration for evaluation metrics.

### 5.1 Datasets

Wicaksono et al. [43] have summarized four datasets [11, 19, 29, 40] for comparing correlation coefficients between metric scores and satisfaction ratings. Among these datasets, we use THUIR1 [11] dataset in our experiments because it involves click-through logs and query-level satisfaction feedbacks from users as well as 4-level graded relevance judgments for all the ten results on each SERP. Besides this public available dataset, we also conduct experiments on the field study dataset we construct in Section 4. Note that we only focus on query-level evaluation in this paper. In addition, for the field study dataset, we find that participants examined more than one SERP for only 60 (about 1.5%) queries. Considering only the top-ten results on the first SERP in our experiments, we filter out these queries and their corresponding sessions. Some statistics of these two datasets are shown in Table 2.

### 5.2 Evaluation Metrics

We investigate the relationship between the accuracy of user models and correlation with user satisfaction for a number of representative evaluation metrics with the C/W/L Framework.

[1] http://www.thuir.cn/tiangong-ss-fsd/

**Table 2: Statistics of the datasets used in our experiments.**

| | User Study [11] | Field Study |
|---|---|---|
| # pairs of $(u, q)$ | 2, 685 | 3, 535 |
| # results per SERP | 10 | $\sim 10$ |
| relevance judgments | 4-level | 4-level |
| query-level satisfaction | 5-level | 5-level |

**Traditional metrics.** First, we select some traditional metrics which have been commonly used in IR evaluation for several years. Precision (Prec) is usually compared as a baseline metric for its simplicity. Taking recall into account, Average Precision (AP) is regarded as the expected utility for the user population [34] and widely used for ranked retrieval. Besides Precision and AP, DCG and RBP are also involved in the set of traditional metrics. DCG introduces the notion of graded gain and discounting the gain depending on rank position, while RBP explicitly encodes a user model defined by the continuation probability.

**Adaptive metrics.** Rather than pre-defining a fixed probability distribution of user behavior, adaptive metrics assume that the continuation/stopping probability is influenced by the previous gains and expectations of users. Therefore, we also consider some adaptive metrics which encode dynamic user models, including INST, BPM, and IFT.

We utilize an evaluation tool called cwl_eval[2] [4] to instantiate the above evaluation metrics. Upon the C/W/L Framework, this tool provides straightforward interfaces to select and configure metrics. Given gain vectors and cost vectors for a metric, cwl_eval can output a series of measurements as well as the corresponding $C(\cdot)$, $W(\cdot)$, and $L(\cdot)$ vectors. In our experiments, we consider the

[2] https://github.com/ireval/cwl

top-10 results for all the metrics. Gains are generated by mapping 4-level graded relevance judgments to scores in {0.0, 0.25, 0.5, 1.0}. For convenience, we adopt the default setting for costs in cwl_eval, which assumes the cost of each result to be one (a unit cost). As for measurements, we report the expected utility (EU) as the outputs of metric scores.

Note that we can set the parameters of metrics when we import them from cwl_eval. We fix the value of $k$ to be 10 for Precision and AP does not contain any parameter, so metric scores and probability distribution vectors of these two metrics are calculated directly. For the remaining metrics (DCG, RBP, INST, BPM, and IFT), we first calibrate their parameters by grid search on training sets and then compare their performance on test sets. In the implementation of cwl_eval, DCG contains a parameter $k$ which is the rank cut off, while fixes the base of the log for discounting to be 2.0. Considering that it is the base, rather than $k$, that is associated with the underlying user model, we let $k = 10$ and search the value of the base in range (1.0, 5.0] with step of 0.1. Larger value of the base is used to model more patient users who are willing to examine more results on the SERP. We also modify the function of continuation probability of DCG in cwl_eval from $\log_{base}(i+1)/\log_{base}(i+2)$ to $\left(1+\log_{base} i\right) / \left(1+\log_{base}(i+1)\right)$. This modification can yield different continuation probabilities with different values of the base. For RBP, the persistence parameter $\theta$ also denotes the patience of a user. We search the value of $\theta$ in range [0, 1] with step of 0.05. All of the three adaptive metrics, INST, BPM, and IFT, have a same parameter $T$, which is the desired amount of gain. We search the value of $T$ for these metrics in range [0.5, 5.0] with step of 0.5. Besides $T$, BPM and IFT have other parameters. For BPM, we only consider its static version which contains another parameter $K$. $K$ is the amount of cost that a user is willing to spend, and we search the value of $K$ in range [2, 10] with step of 2. For IFT, we consider the version which takes both Goal and Rate into account. Note that IFT contains six parameters in total. To reduce the parameter space, in our experiments, we only consider the parameter $T$, which is the target gain, and the parameter $A$, which is the expected average rate of gain. We search the value of $A$ in a set of 0.05, 0.1, 0.2, 0.5, 1. Other parameters of IFT are set as suggested by Azzopardi et al. [3] ($b_1 = b_2 = 0.25$, $R_1 = R_2 = 10$).

## 6 EXPERIMENTS AND RESULTS

To answer the three research questions, we conduct a series of experiments. In this section, we describe the details of our experiments and report the results for each research question.

### 6.1 Models Versus Satisfaction

To investigate whether the metrics calibrated with user behavior data can perform as well as those optimized to fit user satisfaction in evaluating the performance of search systems (**RQ 1**), we compare Spearman's $\rho$ between user satisfaction and evaluation metrics calibrated with different signals.

**Taxonomy of queries.** To evaluate the performance of evaluation metrics in different search scenarios, the experiments are conducted on different categories of queries, following the taxonomy proposed by Broder [6]. This taxonomy classify search queries into informational queries, navigational queries, and transactional queries. With respect to the two datasets shown in Table 2, information of query category is also provided. For the user study dataset, Chen et al. [11] classify the search queries into different categories based on the queries and corresponding descriptions. Similarly, for our field study dataset, two experts annotate query categories at first and a third annotation is required when there is a disagreement. Finally, we get 1, 240 informational queries and 1, 445 navigational/transactional queries in the user study dataset. Our field study dataset contains 3, 163 informational queries and 372 navigational/transactional queries.

**Bootstrap samples of datasets.** To give a fair comparison in our experiments, for each dataset, we use bootstrapping to generate a training set and a corresponding test set and repeat this process one hundred times. Each time the training set is a bootstrap sample (by random sampling with replacement) which has the same sample size as the original data while the queries that are not involved in the training set then form a corresponding test set. Note that the sampling process is employed for different query categories individually.

**Calibration with the accuracy of user models.** As we have mentioned in Section 3.1, the accuracy of user models can be measured by calculating MSE or WMSE of the model-derived and observed probability distributions with different methods, including S_C, S_W, S_L, H_C, H_W, and H_L. For each bootstrap sample, we calibrate the parameters of a metric to optimize the accuracy of its underlying user model on the training set and then calculate Spearman's $\rho$ between user satisfaction and metric scores with the calibrated parameters on the corresponding test set. The average value of Spearman's $\rho$ on one hundred test sets of bootstrap samples is reported in our experiments.

**Calibration with the correlation with user satisfaction.** Besides calibrating with the accuracy of user models, we also compare the performance of different metrics calibrated with Spearman's $\rho$ between user satisfaction and metric scores. One method, denoted as SAT, is to calibrate the metrics with Spearman's $\rho$ on the training set for each bootstrap sample. On the contrary, another method is to calibrate the metrics with Spearman's $\rho$ on the test set for each bootstrap sample. It is denoted as UB, which means this method shows the upper bound of the performance of metrics.

**Results.** Table 3 and Table 4 report Spearman's $\rho$ between different metrics and user satisfaction on different categories of queries in the field study dataset. Comparing different calibration methods with user behavior data, H_L performs best for almost all the metrics except DCG. The difference between DCG and other metrics may be due to the implicit rather than explicit user model of DCG. Note that Precision and AP do not involve any parameter to be calibrated, thus have the same result for different methods. Although there is a difference between different calibration methods with user behavior data, some of them can have the same or even better performance compared with SAT method, which calibrates metrics with user satisfaction feedbacks on the training set. For example, when RBP and BPM are calibrated with H_L method, their Spearman's $\rho$ are 0.282 and 0.298 for informational queries (the same as they are calibrated with SAT method), while 0.332 and 0.350 for navigational/transactional queries (better than SAT method). Comparing different evaluation metrics, the performance is different. Generally, metrics with more sophisticated user models

**Table 3: Spearman's $\rho$ between different metrics and user satisfaction on informational queries in the field study dataset. The first row shows different methods with which metrics are calibrated. (* indicates that SAT/UB significantly performs better than the best calibration method with user behavior data at a level of $p < 0.001$ using two-tailed pairwise t-test)**

| | S_C | S_W | S_L | H_C | H_W | H_L | SAT | UB |
|---|---|---|---|---|---|---|---|---|
| Prec | | | | 0.240 | | | | |
| AP | | | | 0.291 | | | | |
| DCG | 0.257 | 0.264 | 0.261 | 0.266 | **0.275** | 0.269 | 0.279* | 0.282* |
| RBP | 0.268 | 0.274 | 0.268 | 0.280 | 0.280 | **0.282** | 0.282 | 0.285* |
| INST | 0.247 | 0.263 | 0.249 | 0.269 | 0.271 | **0.271** | 0.285* | 0.286* |
| BPM | 0.247 | 0.252 | 0.267 | 0.249 | 0.267 | **0.298** | 0.298 | 0.299* |
| IFT | 0.226 | 0.239 | 0.248 | 0.240 | 0.248 | **0.284** | 0.296* | 0.296* |

**Table 4: Spearman's $\rho$ between different metrics and user satisfaction on navigational and transactional queries in the field study dataset. The first row shows different methods with which metrics are calibrated.**

| | S_C | S_W | S_L | H_C | H_W | H_L | SAT | UB |
|---|---|---|---|---|---|---|---|---|
| Prec | | | | 0.188 | | | | |
| AP | | | | 0.327 | | | | |
| DCG | 0.225 | 0.248 | 0.237 | 0.278 | **0.299** | 0.282 | 0.316* | 0.318* |
| RBP | 0.252 | 0.263 | 0.249 | 0.314 | 0.314 | **0.332** | 0.329 | 0.339* |
| INST | 0.227 | 0.278 | 0.229 | 0.306 | 0.307 | **0.326** | 0.327 | 0.340* |
| BPM | 0.196 | 0.222 | 0.291 | 0.351 | 0.298 | **0.350** | 0.339 | 0.361* |
| IFT | 0.158 | 0.209 | 0.231 | 0.231 | 0.246 | **0.350** | 0.347 | 0.356* |

have larger upper bound of Spearman's $\rho$ between metric scores and user satisfaction. For example, BPM, which takes both desired gain and cost into consideration, has the largest upper bound for either informational queries or navigational/transactional queries. However, there may be a gap between the performance of metrics calibrated with user behavior data and their upper bounds. Note that the performance of IFT is not as well as that of BPM. One possible reason is that we only tune two parameters of IFT.

The results in the user study dataset are shown in Table 5 and Table 6. Compared with results in the field study dataset, there are some differences. First, the performance of calibration methods is different. On one hand, the difference between these methods is smaller than that in the field study dataset. For example, the performance of DCG on informational queries with different calibration methods is almost the same. On the other hand, Soft (S) methods perform better on informational queries while Hard (H) methods perform better on navigational/transactional queries. It suggests that users who search for navigational/transactional queries tend to end their search process, rather than view more results with a discounting probability, after last click. Second, comparing different metrics, it is found that the metrics have similar and better performance on informational queries, while various and worse performance on navigational/transactional queries. This may due to some inherent features of the user study dataset. In laboratory user study, participants usually required to finish several tasks pre-defined by researchers, which may have an influence on their search behavior.

**Table 5: Spearman's $\rho$ between different metrics and user satisfaction on informational queries in the user study dataset. The first row shows different methods with which metrics are calibrated.**

| | S_C | S_W | S_L | H_C | H_W | H_L | SAT | UB |
|---|---|---|---|---|---|---|---|---|
| Prec | | | | 0.473 | | | | |
| AP | | | | 0.482 | | | | |
| DCG | **0.487** | 0.487 | 0.487 | 0.487 | 0.486 | 0.486 | 0.486 | 0.488* |
| RBP | **0.488** | 0.473 | 0.488 | 0.482 | 0.472 | 0.481 | 0.488 | 0.490* |
| INST | **0.474** | 0.468 | 0.474 | 0.474 | 0.468 | 0.473 | 0.472 | 0.475* |
| BPM | **0.484** | 0.481 | 0.476 | 0.484 | 0.481 | 0.396 | 0.485 | 0.495* |
| IFT | **0.484** | 0.473 | 0.484 | 0.469 | 0.469 | 0.418 | 0.493* | 0.498* |

**Table 6: Spearman's $\rho$ between different metrics and user satisfaction on navigational and transactional queries in the user study dataset. The first row shows different methods with which metrics are calibrated.**

| | S_C | S_W | S_L | H_C | H_W | H_L | SAT | UB |
|---|---|---|---|---|---|---|---|---|
| Prec | | | | 0.262 | | | | |
| AP | | | | 0.342 | | | | |
| DCG | 0.303 | 0.303 | 0.304 | 0.303 | 0.312 | **0.314** | 0.332* | 0.335* |
| RBP | 0.312 | 0.320 | 0.312 | 0.319 | **0.324** | 0.320 | 0.322 | 0.328* |
| INST | 0.311 | 0.329 | 0.311 | 0.313 | **0.329** | 0.317 | 0.330 | 0.334* |
| BPM | 0.276 | 0.281 | 0.302 | 0.276 | 0.282 | **0.304** | 0.309 | 0.329* |
| IFT | 0.284 | 0.287 | **0.291** | 0.280 | 0.287 | 0.283 | 0.303* | 0.322* |

Although the performance of calibration methods to optimize the accuracy of user models in the user study dataset is different from that in the field study dataset, most metrics such as RBP, INST, and BPM, when calibrated with user behavior data, can have a competitive performance compared with SAT method.

To summarize the results in this section, we conclude the following findings:

- Most metrics calibrated with user behavior data (by appropriate measurements) can perform as well as those optimized to fit user satisfaction in evaluating the performance of search systems.
- Without eye-tracking data, employing last click signal is enough to provide information for estimating the accuracy of user models.
- RBP and BPM have relatively better performance with respect to the consistency between two facets of metrics.

## 6.2 Stability of Parameters

To verify whether the parameters are stable while being calibrated with different samples of user behavior data (**RQ 2**), we make a sensitivity analysis of the parameters. Recall that we generate one hundred bootstrap samples in Section 6.1, in this section, we take RBP and BPM as two examples and depict their calibrated parameters on the training set and the value of Spearman's $\rho$ between metric scores and user satisfaction on the test set for all the samples in the field study dataset. We compare the performance of H_L and SAT calibration methods. The results are shown in Figure 1 and Figure 2.

From the figures, we can find some differences between different calibration methods and query categories. Comparing different

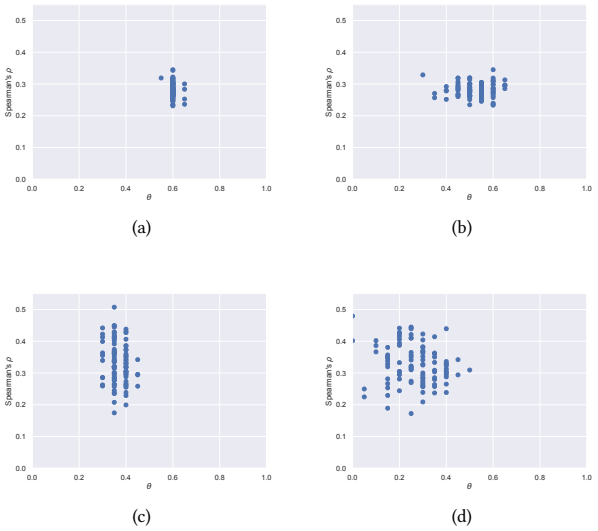



**Figure 1: Sensitivity analysis of parameter $\theta$ and Spearman's $\rho$ between metric score and user satisfaction for RBP on the field study dataset. (a) and (b) compare the performance of RBP calibrated with H_L and SAT calibration methods on informational queries while (c) and (d) compare that on navigational and transactional queries.**

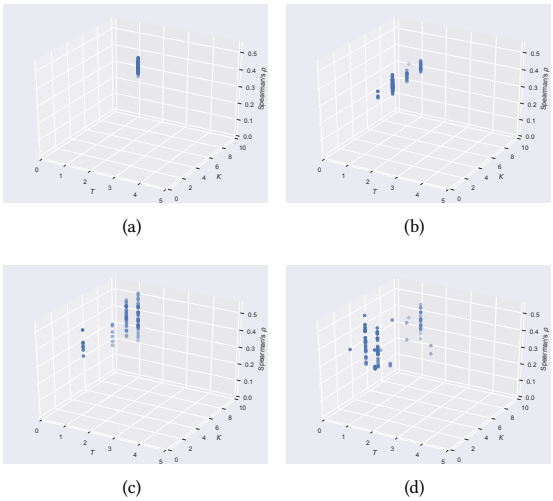



**Figure 2: Sensitivity analysis of parameter $T$, $K$ and Spearman's $\rho$ between metric score and user satisfaction for BPM on the field study dataset. (a) and (b) compare the performance of BPM calibrated with H_L and SAT calibration methods on informational queries while (c) and (d) compare that on navigational and transactional queries.**

calibration methods, the values of calibrated parameters with H_L method lie in a smaller range than those with SAT method. For example, considering informational queries, the optimal values of $\theta$ are from 0.55 to 0.65 when calibrated with H_L method, while from 0.30 to 0.65 when calibrated with SAT method. Given that RBP and BPM calibrated with H_L method have the same or even better performance compared with SAT method in Section 6.1, the parameters learned with H_L method are more stable than SAT method. When evaluation metrics are calibrated to optimize the correlation between metric scores and user satisfaction feedbacks, it may overfit the parameters of metrics, which makes the value of parameters and the performance of metrics unstable. On the contrary, user models characterize search behavior of simulated users, which makes the relationship between metrics and user satisfaction easily explicable. Therefore, calibrating metrics with user behavior data may lead to more stable values of parameters and performance of metrics.

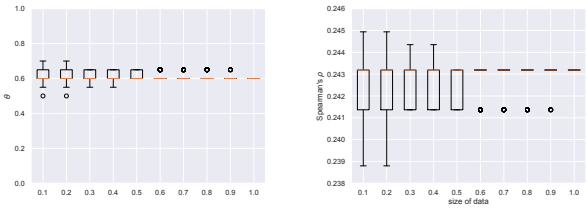
On the other hand, comparing different query categories, the optimal values of parameters are different between informational queries and navigational/transactional queries. For example, $\theta$ of RBP for informational queries (from 0.55 to 0.65) is larger than that for navigational/transactional queries (from 0.30 to 0.45). Note that $\theta$ measures the persistence of simulated users. It is intuitive that $\theta$ is larger for informational queries since more information is required to fulfill the needs for informational queries. It indicates that it is necessary to calibrate the parameters of metrics and may lead to different results for different query categories.

In summary, this section suggests that the parameters of evaluation metrics are stable while being calibrated with different samples of user behavior data, which demonstrates the effectiveness of calibrating evaluation metrics with user behavior data to evaluate search systems.

## 6.3 Data Requirements

Finally, we investigate how much data is required for an evaluation metric to tune its parameters to correlate well with user satisfaction (**RQ 3**). Slightly different from previous experiments, in this section, we first randomly sample 30% of the original dataset as our fixed test set. The remaining queries are used for training and we compare the performance of calibrating metrics with different scales of the training data. The range of the scales is from 10% to 100%, with step of 10%. Except using 100% of the training data, for each scale, we randomly sample corresponding size of queries and calibrate metrics with these queries. Then we compute Spearman's $\rho$ between metric scores and user satisfaction feedbacks on the fixed test set. We repeat this process one hundred times and show the distributions of optimized values of parameters and performance of metrics. In this section, we also take RBP and BPM as two examples. The experiments are conducted on informational queries in the field study dataset. The test set contains 949 queries while the overall training set contains 2, 214 queries. In addition, the calibration method we select is H_L.

The results are shown in Figure 3 and Figure 4. For both metrics, it is found that either the values of parameters or the performance of metrics become less stable as we utilize smaller size of the training data. For RBP, the value of $\theta$ and the performance of metric nearly remains the same when we utilize 60% (about 1328 queries) or




(a) (b)

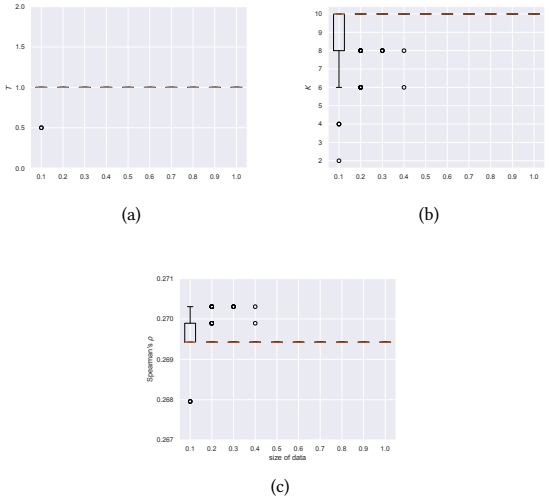
**Figure 3: The influence of data size on parameter $\theta$ and performance of RBP.**

larger scale of the training data. While for BPM, even we utilize only 20% (about 443 queries) of the training data, the values of its parameters and performance are quite stable, as same as those by using 100% of the training data. It suggests that a small size of user behavior data can be representative to characterize general behavior of the population in a certain search scenario. For example, the optimized parameters of BPM in Figure 4 reveal some behavior of the participants in our field study dataset when they search for informational queries. Given the underlying user model of BPM, users expect to find information which can provide gains measured by $T$. Simultaneously, they are willing to spend limited effort to examine a number of results measured by $K$. Figure 4 show that $T = 1.0$ and $K = 10$ for most queries. Therefore, we can construct the following user model of the participants in our field study. They use a search engine to find some useful information about their information needs. If they find a perfect result (gain is 1) which provides all the information they require, or they find several results which provide complementary information and jointly fulfill their needs, they will end their search process. Otherwise, they usually examine all the top-ten results on the first SERP and then reformulate a new query or try other ways to solve their problem.

In this section, we take deep insight to investigate data requirements for the calibration of evaluation metrics. The results show that it is feasible to calibrate evaluation metrics to perform well in estimating the performance of search systems by utilizing a small scale of user behavior data.

## 7 CONCLUSION

User-centric evaluation raises the validity of an evaluation metric in two facets. On one hand, the underlying user model of the metric should accurately predict user behavior. On the other hand, the score of the metric should correlates well with user satisfaction. In this paper, we take a further step toward better understanding the relationship between these two facets of evaluation metrics. In particular, we investigate whether the metrics that are calibrated with user behavior data can simultaneously perform well in estimating user satisfaction. To shed light on this research question, we compare the performance of a broad of metrics defined by the C/W/L Framework in terms of estimating user satisfaction when they are calibrated with user behavior data. Besides a previous public resource, we construct a field study dataset where daily search




(a) (b)

(c)

**Figure 4: The influence of data size on parameter $T$, $K$ and performance of BPM.**

logs and explicit feedbacks of 30 participants are collected for one month.

Comparing the performance of different evaluation metrics and different calibration methods, we find that most metrics calibrated with user behavior data which contains only users' last click signals can perform as well as those optimized to fit user satisfaction feedbacks in evaluating the performance of search systems. We further make a sensitivity analysis of parameters to investigate the reliability in the calibration process of evaluation metrics. The results indicate that the parameters calibrated with user behavior data are stable while feeding with different samples of data. It is also found that the values of calibrated parameters may be different for different query categories. Finally, we also explore data requirements for the calibration of evaluation metrics. Our findings demonstrate that calibrating evaluation metrics with a small scale of user behavior data is an effective and feasible way to evaluate search systems.

Focusing on the consistency between user behavior modeling and satisfaction measurement of evaluation metrics is an important new way of thinking about meta-evaluation of metrics [43]. Our work provides empirical support for the consistency between these two facets, as well as guidance for tuning the parameters of evaluation metrics in this line of research. Nevertheless, to develop a fundamental framework for this meta-evaluation method, more research questions need to be addressed in the future work. (1) We adopt the methods proposed in some previous studies to measure the accuracy of user models and the correlation between evaluation metrics and user satisfaction. However, wherever these methods are best appropriate to the measurements remains to be explored. (2) In this paper, we mainly focus on query-level evaluation metrics. As session search evaluation has been paid more attention recently, it is also important to construct effective user behavior models and verify the consistency of two facets for session-level evaluation metrics.

## REFERENCES


[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th SIGIR Conference*. ACM, 773–774.

[2] Rashid Ali and MM Sufyan Beg. 2011. An overview of Web search evaluation methods. *Computers & Electrical Engineering* 37, 6 (2011), 835–848.

[3] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *Proceedings of the 41st SIGIR Conference*. ACM, 605–614.

[4] Leif Azzopardi, Paul Thomas, and Alistair Moffat. 2019. cwl_eval: An evaluation tool for information retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1321–1324.

[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*. ACM, 625–634.

[6] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.

[7] Chris Buckley and Ellen M Voorhees. 2017. Evaluating evaluation measure stability. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 235–242.

[8] Stefan BÃijttcher, Charles L. A. Clarke, Peter C. K. Yeung, and Ian Soboroff. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*.

[9] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 903–912.

[10] 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*. 621–630.

[11] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th SIGIR Conference*. ACM, 15–24.

[12] Cyril Cleverdon, Jack Mills, and Michael Keen. 1966. ASLIB Cranfield Research Project: factors determining the performance of indexing systems. (1966).

[13] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. 87–94.

[14] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2014. Injecting user models and time into precision via Markov chains. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 597–606.

[15] Jiyin He and Emine Yilmaz. 2017. User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 67–76.

[16] Charles R Hildreth. 2001. Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. *Information research* 6, 2 (2001), 6–2.

[17] Scott B Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction?. In *Proceedings of the 30th SIGIR Conference*. ACM, 567–574.

[18] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[19] Jiepu Jiang and James Allan. 2016. Correlation between system and user metrics in a session. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 285–288.

[20] Jiepu Jiang and James Allan. 2017. Adaptive persistence for search effectiveness measures. In *Proceedings of the 2017 CIKM Conference*. ACM, 747–756.

[21] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth WSDM Conference*. ACM, 57–66.

[22] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

[23] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 193–202.

[24] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[25] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, and Shaoping Ma. 2018. Towards designing better session search evaluation metrics. In *Proceedings of the 41st SIGIR Conference*. ACM, 1121–1124.

[26] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Satisfaction with Failure or Unsatisfied Success: Investigating the Relationship between Search Success and User Satisfaction. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1533–1542.

[27] Mengyang Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Cognitive Effects in Session-level Search User Satisfaction. KDD.

[28] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 493–502.

[29] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does relevance mean usefulness and user satisfaction in Web search?. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 463–472.

[30] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 24.

[31] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd CIKM conference*. ACM, 659–668.

[32] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2.

[33] Karl Pearson. 1896. VII. Mathematical contributions to the theory of evolution.-III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 187 (1896), 253–318.

[34] Stephen Robertson. 2008. A new interpretation of average precision. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 689–690.

[35] 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*. 525–532.

[36] Tetsuya Sakai. 2013. How intuitive are diversified search metrics? Concordance test results for the diversity U-measures. In *Asia Information Retrieval Symposium*. Springer, 13–24.

[37] Mark Sanderson et al. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[38] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 555–562.

[39] C Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (1904), 72–101.

[40] Paul Thomas, Alistair Moffat, Peter Bailey, Falk Scholer, and Nick Craswell. 2018. Better Effectiveness Metrics for SERPs, Cards, and Rankings. In *Proceedings of the 23rd Australasian Document Computing Symposium*. ACM, 1.

[41] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ahmed Hassan, and Ryen W White. 2014. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 123–132.

[42] Alfan Farizki Wicaksono and Alistair Moffat. 2018. Empirical evidence for search effectiveness models. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1571–1574.

[43] Alfan Farizki Wicaksono and Alistair Moffat. 2020. Metrics, User Models, and Satisfaction. In *Proceedings of the Thirteenth WSDM Conference*. ACM.

[44] Alfan Farizki Wicaksono, Alistair Moffat, and Justin Zobel. 2019. Modeling User Actions in Job Search. In *European Conference on Information Retrieval*. Springer, 652–664.

[45] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The influence of image search intents on user behavior and satisfaction. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 645–653.

[46] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Yunqiu Shao, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Grid-based Evaluation Metrics for Web Image Search. In *The World Wide Web Conference*. 2103–2114.

[47] Ya Xu and David Mease. 2009. Evaluating web search using task completion time. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 676–677.

[48] Emine Yilmaz and Stephen Robertson. 2010. On the choice of effectiveness measures for learning to rank. *Information Retrieval* 13, 3 (2010), 271–290.

[49] 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*. 425–434.

[50] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. How Well do Offline and Online Evaluation Metrics Measure User Satisfaction in Web Image Search?. In *Proceedings of the 41st SIGIR Conference*. ACM, 615–624.