Constructing Better Evaluation Metrics by Incorporating the Anchoring Effect into the User Model

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

Fan Zhang
fan.zhang@whu.edu.cn
Wuhan University
Wuhan, China

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

ABSTRACT

Models of existing evaluation metrics assume that users are rational decision-makers trying to pursue maximised utility. However, studies in behavioural economics show that people are not always rational when making decisions. Previous studies showed that the anchoring effect can influence the relevance judgement of a document. In this paper, we challenge the rational user assumption and introduce the anchoring effect into user models. We first propose a framework for query-level evaluation metrics by incorporating the anchoring effect into the user model. In the framework, the magnitude of the anchoring effect is related to the quality of the previous document. We then apply our framework to several query-level evaluation metrics and compare them with their vanilla version as the baseline in terms of user satisfaction on a publicly available search dataset. As a result, our Anchoring-aware Metrics (AMs) outperformed their baselines in term of correlation with user satisfaction. The result suggests that we can better predict user query satisfaction feedbacks by incorporating the anchoring effect into user models of existing evaluating metrics. As far as we know, we are the first to introduce the anchoring effect into information retrieval evaluation metrics. Our findings provide a perspective from behavioural economics to better understand user behaviour and satisfaction in search interaction.

CCS CONCEPTS

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

KEYWORDS

anchoring effect, cognitive bias, user behaviour, information retrieval, evaluation metrics

ACM Reference Format:
Nuo Chen, Fan Zhang, and Tetsuya Sakai. 2022. Constructing Better Evaluation Metrics by Incorporating the Anchoring Effect into the User Model. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531953

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531953

1 INTRODUCTION

Models of existing evaluation metrics are based on the assumption that users are rational decision-makers trying to pursue maximised utility. However, studies in behavioural economics show that one's judgement can be influenced by cognitive biases, which deviates their decisions from what is expected given rational decision-making models [29–31]. Since the process of the user interacting with a search engine can also be viewed as a decision-making process, cognitive bias can also take place in search interaction [14]. The anchoring effect is one of the cognitive biases. It is a phenomenon that judgements and decisions are "biased toward the initial values" and "different starting points yield different estimates" [29]. To date, the knowledge about the anchoring effect in search interaction is still limited, let alone incorporating the anchoring effect into models of evaluation metrics. Shokouhi *et al.* [26] showed that the relevance judgment of documents can be affected by the anchoring effect. They observed that when the previous document is highly relevant or somewhat relevant, the relevance label assigned to the next document is more likely to be affected by the previous document. However, they did not further apply their findings to modify existing user models of evaluation metrics.

To address the research gap discussed above, we issue the following research questions:

- **RQ1**: How can we incorporate the anchoring effect into the user model based on existing work and apply it to the framework of current metrics?
- **RQ2**: How do our Anchoring-aware Metrics (AMs) compare with the state-of-the-art metrics in terms of correlation with user satisfaction?

In this paper, we propose a framework to enhance the performance of evaluation metrics for query-level search by incorporating the anchoring effect into the document utility derivation model. We test the effectiveness of our framework in terms of correlation with user satisfaction on the THUIR1 dataset [7]. Our experiments demonstrate the strong influence of the anchoring effect on users' utility derivation. The result also shows that our Anchoring-aware Metrics outperform the baseline metrics in term of correlation with user satisfaction. This suggests that we can predict users' query satisfaction feedbacks more effectively by incorporating the anchoring effect into user models of existing evaluating metrics.

As far as we know, we are the first to propose a metric framework based on a user model incorporating the anchoring effect. This serves as an initial step towards developing more effective evaluations metrics by incorporating cognitive effects into the user model.

2 PRIOR ART

The last two decades have seen a range of evaluation metrics based on explicit or implicit user behaviour models being proposed. These metrics include Discounted Cumulative Gain (DCG) and its variants [10], Rank-Biased Precision (RBP) [22], Expected Reciprocal Rank (ERR) [6], Expected Browsing Utility (EBU) [35], Time-Biased Gain (TBG) [27], U-measure [23], INSQ [21], INST [4], Bejeweled Player Model (BPM) Metrics [36], and so forth. Moffat *et al.* [20, 21] introduced the C/W/L Framework to provide a uniform ground for comparing different metrics by formalising the user browsing behaviour into three different but interrelated aspects: Continuation probability (C), Weight function (W) and Last probability (L). In recent years, studies regarding user models and user model-based evaluation metrics have been extended to the session-level [11, 16, 33], but this is beyond the scope of our study.

Despite the multifarious metrics, the models behind them can all be viewed as the simulation of the process of a user interacting with a system under operational settings [24], and the metric score can be viewed as a simulation of the user's *utility* (also referred to as *gain* or *benefit*) accumulated during that process. Carterette [5] summarised three underlying models comprising model-based measures: (1) a browsing model that describes how a user interacts with results; (2) a model of document utility, describing how a user derives utility from individual relevant documents; (3) a utility accumulation model that describes how a user accumulates utility in the course of browsing. Moffat *et al.* [19] classified the user browsing model into two categories: static and adaptive according to whether user behaviour will be affected by the benefits derived from the document.

Implicit in the above study is the assumption that users are rational decision-makers trying to pursue the optimal solution. However, evidence from behavioural economics suggests that it is not always the case in reality. Tversky and Kahneman's work [29–31] showed that one's judgement can be influenced by cognitive biases, which deviates their decisions from what is expected given rational decision-making models. These biases arise from people's limited ability to attend to and correctly process all the information available to them [13]. Since the process of a user interacting with a search engine (like making relevance judgement, deciding to continue or to leave, evaluating satisfaction etc.) can also be viewed as a decision-making process, cognitive bias can also take place in search interaction [14]. Azzopardi [2] summarised the ways in which cognitive biases affect the search process, including: (1) Querying, (2) Document examination, (3) Relevance judgment, and (4) Satisfaction perception.

To date, a range of studies in IR discipline have shed light on cognitive biases in search interactions and crowdscourcing [9, 12, 15], like the recency effect [18], the ordering effect[8, 25], and the reference dependence effect [17]. Zhang *et al.* [37] further modified existing metrics by incorporating the factor of recency effect. However, the above studies did not specifically address the anchoring effect. The anchoring effect is one of the cognitive biases. It is a phenomenon that judgements and decisions are "biased toward the initial values" and "different starting points yield different estimates" [29]. Shokouhi *et al.* [26] showed that the relevance judgment of documents can be affected by the anchoring effect. They observed that *when the previous document is highly relevant or somewhat relevant, relevance labels assigned to the next document is more likely to be affected the previous document.* Thomas *et al.* [28] also observed the anchoring effect in relevance judgements of crowdsourced labels. However, the above works did not further apply their findings to modify existing user models of evaluation metrics.

Compared to these studies, our work sheds light on users' search interaction under the anchoring effect and is the first to propose a metric framework based on a user model incorporating the anchoring effect.

3 FRAMEWORK OF ANCHORING-AWARE METRICS

To answer **RQ1**, we propose the framework of Anchoring-aware Metrics (AMs) with the document utility model incorporating the anchoring effect. Before introducing the modification, we start by a generic formulation of existing metrics as follows. The score of most evaluation metrics can be calculated from the inner-product of a gain vector and a weighting (or discount) vector [38]:

$$f(q) = \sum_{n=1}^{N} g_n(q) \cdot d_n \tag{1}$$

where $f(q)$ is the score of query q, $g_n(q)$ denotes the gain users obtain from the n-th result returned by the system given the query q, and d_n denotes the relative importance to the user of encountering a relevant document at position n. For metrics compatible with the C/W/L framework, the discount vector d can be considered as the weight vector W in the C/W/L framework [20, 21]. However, for ERR, its discount vector should be specially treated since Azzopardi [3] argued that ERR is not C/W/L. For most metrics, we can let $g_n(q) = r_n(q)$, where $r_n(q)$ ranging in $[r_{\min}, r_{\max}]$ denotes the relevance level of the n-th result.

We then modify the document utility model base on the work of Shokouhi *et al.* [26]: (1) The user's utility derived from the first document depends only on the quality of itself, since there is no reference. (2) For documents at rank 2 or below, the user's utility depends not only on the quality of the current document, but also on the quality of the previous document. In other words, the quality of the previous document is the anchor. We define the relevance level users perceived under anchoring effect as the following:

$$r'_n(q) = \begin{cases} r_n(q) & \text{for} \quad n = 1, \\ \alpha \cdot r_{n-1}(q) + (1-\alpha) \cdot r_n(q) & \text{otherwise} \end{cases} \tag{2}$$

where α is a weight for controlling the anchoring effect. The larger the α is, the greater the anchoring effect of the previous document is on the user's utility derived from the current document. To take the anchoring effect into account when computing evaluation measures, we can simply let $g_n(q) = r'_n(q)$ in Eq. 1. Note that here we suppose users can still judge the quality of documents without bias, but their gain derived from the current document is eventually affected by the quality of the last document seen. We made this assumption because: *document quality* should be considered as an objective concept and, in contrast, the *utility* (or *gain*) that users derive from a document should be subjective. Note that for metrics like ERR and INST, their models are *adaptive*, which means that the probability whether users decide to leave or continue is not only related to the

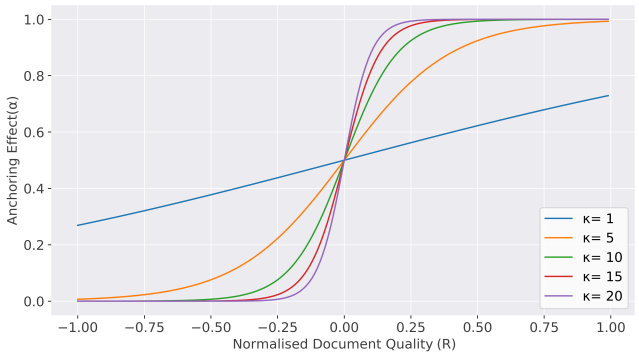

Figure 1: The curve of the anchoring effect when $\lambda = 1$ and κ = 0.05, 0.15, 0.25, 0.35, 0.45 respectively.

ranking position n but also related to the relevance of the top n documents. Therefore, the d_n in Eq. 1 may also be a function of the relevance of the top n documents. For example, ERR is given by [6]:

$$\sum_{k=1}^{n} \frac{1}{k} P_s(k) \tag{3}$$

Here, $P_s(k)$ denotes the probability the user abandons the ranked list at position k. Here, $\frac{1}{k}$ and $P_s(k)$ are viewed as g_n and d_n respectively. $P_s(k)$ is given by:

$$\prod_{i=1}^{k-1}(1 - p_s(i))p_s(k)$$

where

$$p_s(i) = \frac{2^{r_i(q)} - 1}{2^{r_{\max}}}$$

is a function of $r_i(q)$. The probability that the user continues browsing from position n to position $n+1$ in INST is related to the total accumulated relevance, and the d_n of INST [4] ($W(n)$) is a function of: $\sum_{i=1}^{n} r_i(q)$. After considering the anchoring effect, both probabilities above are also computed by replacing $r_n(q)$ with $r'_n(q)$, since our fundamental assumption is that users' decisions in search interaction are affected by cognitive biases.

The work of Shokouhi *et al.* [26] further suggested that the factor of anchoring effect is not set in stone: when the quality of the previous document is low, the factor of anchoring effect should be trivial. On the other hand, when the quality of the previous document is fair or high, the factor of the anchoring effect should be substantial. We use a sigmoid function to simulate this:

$$\alpha = \frac{\lambda}{1 + \exp(-\kappa R_{n-1}(q))} \tag{4}$$

where

$$R_{n-1}(q) = \frac{r_{n-1}(q) - (r_{\max} + r_{\min})/2}{r_{\max} - (r_{\max} + r_{\min})/2}$$

is a normalisation of the relevance level r_{n-1} ranging in $[-1, 1]$, λ stands for the upper bound of the factor of the anchoring effect, and κ stands for how fast the anchoring effect grows as the document quality increases. The smaller the κ, the faster the growth. Figure 1 shows the curve of the anchoring effect in our framework when $\lambda = 1$ and $\kappa = 1, 5, 10, 15, 20$ respectively.

4 EXPERIMENT

To answer **RQ2**, we apply our AM framework to several query-level evaluation metrics and compare them with their vanilla version as the baseline in terms of user satisfaction on the THUIR1 [7] dataset.

The THUIR1 dataset is developed from a laboratory experiment. It involves 2, 435 single-SERP sessions, along with click-through logs and query-level satisfaction feedbacks from users, as well as 4-level graded relevance labels for all the 10 results on each SERP. We excluded 44 records which can not be parsed due to various reasons and eventually get 2, 391 records.

In our experiment, we apply the framework of AM to several query-level metrics by replacing the relevance $r_n(q)$ in $g_n(q)$ and d_n in Eq. 1 with $r'_n(q)$ as we mentioned in Section 3. These metrics include: (1) ERR, (2) Precision, (3) scaled DCG [22], (4) RBP, (5) INSQ, and (6) INST. We use scaled DCG for its convenience for calculating the probability the user continues examining.

To compare the effectiveness of our AMs and the baseline metrics, we randomly split the data into 5 folds, using 4 folds of the data as the training set to tune the parameters, and using 1 fold as the test set to compare the performance of different metrics. We repeat this process 10 times.

We first calibrate the parameters of the baseline metrics except for ERR and Precision by user behaviour. Here we choose the browsing behaviour as the surrogate for user behaviour. We suppose that the probability that users examine the first result is 1, and we compute the probabilities that users examine the result at rank i using the Continuation probability $C(i)$ of the C/W/L framework [3]:

$$\hat{V}_M(i) = \begin{cases} 1 & \text{for} \quad i = 1, \\ \hat{V}_M(i-1) \cdot C(i-1) & \text{otherwise} \end{cases} \tag{5}$$

Here, $\hat{V}_M(i)$ denotes the estimated probability that users examine the result at rank i given by metric M. $C(j)$ is the probability that the user will progress from rank j to rank $j+1$. As there is no eye-tracking-based evidence in our dataset to show whether a user examined the document, we used the click-through records in the logs to infer the probability that a particular document is examined. We compute the probability based on the model proposed by Wicaksono *et al.* [34]:

$$\hat{V}(i \mid u, q) = \begin{cases} 1 & \text{for} \quad i \le DC(u,q), \\ e^{n/g(K(u,q))} & \text{otherwise} \end{cases} \tag{6}$$

Here, $\hat{V}(i \mid u, q)$ is the estimated probability that user u views the item listed at rank i for query q. $g(x) = \ln(1 + e^x)$ is a *softplus* function. $K(u, q)$ is a linear combination of the deepest rank position clicked $DC(u, q)$ and the number of distinct items clicked $NC(u, q)$:

$$K(u, q) = w_0 + w_1 \cdot DC(u, q) + w_2 \cdot NC(u, q)$$

where (w_0, w_1, w_2) are parameters estimated from the data.

Wicaksono *et al.* [32] fitted $K(u, q)$ on the THUIR1 dataset and showed that $w_0 = 3.48$, $w_1 = 0.46$, and $w_2 = 0.20$. Here we also use this setting to estimate $\hat{V}(i \mid u, q)$. We compute the Total Squared Error (TSE) between $\hat{V}_M(i)$ and $\hat{V}(i \mid u, q)$ over the top-30 ranks of all the queries in the training set, and use it to measure the goodness of the fit. We then perform a grid search to find the optimal parameter setting that minimise TSE. After finding the optimal

Table 1: The mean and standard deviation of each parameter optimised for each trial

	AM-ERR		AM-Precision		AM-scaled-DCG			AM-RBP			AM-INSQ			AM-INST		
	λ	κ	λ	κ	b	λ	κ	p	λ	κ	T	λ	κ	T	λ	κ
Mean	1.00	19.10	1.00	17.10	1.90	1.00	13.30	0.85	1.00	12.20	12.00	1.00	6.70	12.60	1.00	5.20
σ	0.000	0.316	0.000	1.853	0.037	0.000	6.325	0.000	0.000	8.230	9.298	0.00	5.165	6.667	0.00	1.033

Table 2: Spearman's correlations between metrics and user satisfaction.

Metric	UB	US
ERR	0.311^{**}	
AM-ERR	**0.315**	
Precision	0.241^{***}	
AM-Precision	**0.306**	
scaled-DCG	0.322^{***}	0.324^{***}
AM-scaled-DCG	**0.348**	
RBP	0.313^{***}	0.331^{***}
AM-RBP	**0.347**	
INSQ	0.290^{***}	0.326
AM-INSQ	**0.333**	
ISNT	0.282^{***}	0.330
AM-INST	**0.332**	

values of parameters related to user browsing models, parameters of metrics without the anchoring effect are determined.

For our AMs, they have additional parameters λ and κ. To tune these parameters with the training data, we compute the correlation between the scores of metrics and the scores of query-level user satisfaction feedback. We then perform a grid search to find the optimal parameter setting that maximise Spearman's correlation coefficient ρ. We search the values for λ in range $[0, 1]$ with step of 0.1 and the values for κ in range $[0.05, 0.5]$ with step of 0.05. Since our AMs are calibrated with user satisfaction, for fair comparison, we compare them not only with the baseline metrics calibrated with user behaviour but also with those calibrated with user satisfaction.

5 RESULTS AND DISCUSSIONS

5.1 Analysis of the Parameters

Table 1 reports the mean and standard deviation of each parameter optimised for each of the 10 trials. From table 1, we can observe that the two parameters of the Anchoring-aware Metrics differ substantially in terms of stability. For λ, which stands for the upper bound of the factor of the anchoring effect, it is very stable and it almost never varies among different metrics. The large value of λ also shows the strong influence of the anchoring effect on users' utility derivation. However, regarding the case of κ, it is unstable and varies among different metrics. A possible reason is that our function regarding α is not sufficient for explaining the anchoring effect well. More user studies are required to further scrutinise the mechanism of the anchoring effect in search interaction.

5.2 The Effectiveness of AMs

Table 2 reports the results of Spearman's correlations between different metrics and user satisfaction. Note that *UB* means calibrated by User Behavior and *US* means calibrated by User Satisfaction. Bold font indicates the strongest correlation in each block.*, ** and *** indicates the difference between an AM and its baseline(s) within a block is significant at $p < 0.05$, $p < 0.01$ and $p < 0.001$ level with a Bonferroni correction respectively. Each block compares an AM metric and its baseline. As we can see from Table 2, all of our AMs outperformed their baselines, no matter whether they are calibrated by user behaviour or user satisfaction. The difference(s) between the AM and its baseline(s), except for INSQ (US) vs. AM-INSQ and INST (US) vs. AM-INST, are statistically significant. These results suggest that after incorporating the anchoring effect into the user model, the metric scores better reflect users' satisfaction feedbacks, and we can better predict user query satisfaction feedbacks by incorporating the anchoring effect into user models of existing evaluating metrics.

Recall that for ERR and INST, we compute the probability whether users decide to leave or continue using the user perceived relevance under the anchoring effect $r'_n(q)$. Our AM-ERR and AM-INST correlate better than their baselines with user satisfaction feedbacks, indicating that users' browsing behaviours can be better predicted after incorporating the anchoring effect into the user model.

5.3 Further Investigation

To investigate to what extent the proposed AMs change the preference for SERPs compared with their baselines, we conduct an additional experiment on the THUIR1 dataset. For each query in the dataset, we compute the score of SERP given by the above baseline metrics and their AM versions respectively, with parameters using the value of the mean of each parameter optimised in the previous trials (refer to Table 1). We obtain $2,391 * 2,390/2 = 2,857,245$ different SERP pairs in total. We then collect the count of SERP pairs where the preferences given by the baseline metric and its AM version are different.

As a result, the number of inconsistent pairs for ERR and AM-ERR is 67,314 (2.36% of the total), for Precision and AM-Precision is 327,639 (11.47% of the total), for scaled DCG and AM-scaled-DCG is 254,464 (8.91% of the total), for RBP and AM-RBP is 296,241 (10.37% of the total), for INSQ and AM-INSQ is 313,125 (10.96% of the total), and for INST and AM-INST is 312,804 (10.94% of the total). This suggests that around 2% to 11% of the conclusions drawn by current metrics are overridden after considering the anchoring effect. Figure 2 shows the count and the proportion of inconsistent pairs for each metric group.

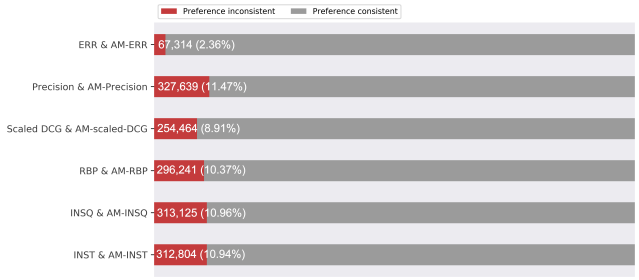

Figure 2: The count and proportion of inconsistent pairs for each metric group

6 CONCLUSIONS AND LIMITATIONS

In this paper, we challenge the rational user assumption that underlies the models of existing evaluation metrics. We propose a framework for query-level evaluation metrics by incorporating the anchoring effect into the user model. As far as we know, this is the first in the IR discipline. To test the effectiveness of our framework, we apply it to several query-level evaluation metrics and compare them with their vanilla version as the baseline in terms of user satisfaction on a publicly available search dataset.

The analysis of the parameters of our framework shows the strong influence of the anchoring effect on users' utility derivation. The comparison between our Anchoring-aware Metrics (AMs) and their baselines shows that most of our AMs outperformed their baselines. This indicates that we can better predict user query satisfaction feedbacks by incorporating the anchoring effect into user models of existing evaluating metrics. Further investigation suggests that around 2% to 11% of the conclusions drawn by current metrics are overridden after considering the anchoring effect. Our findings provide a perspective from behavioural economics to better understanding user behaviour and satisfaction in search interaction and to construct better evaluation metrics by incorporating cognitive effects into the user model.

Nevertheless, there are also some limitations in our work which we would like to tackle for future work. (1) The parameter κ in our framework is unstable and varies among different metrics. This suggests that our framework does not perfectly simulate the mechanism of the anchoring effect. In the future, a more elaborated model should be proposed based on further user studies. (2) The provider of the THUIR1 dataset did not provide details of the relevance assessment, so these relevance scores might have been already affected by the anchoring effect. In this case, our method might be in a sense doubling the effect. This could affect the effectiveness of our results. To obtain unbiased "pure" judgements, a new dataset with an appropriate arrangement in assessment to avoid the odering effect and the anchoring effect is needed. (3) Our framework is for query-level search interactions, and it is still unknown whether the anchoring effect occurs in a multi queries search session. It would be interesting to discuss what would be the impact of the anchoring effect in other more complex information retrieval tasks, such as diversity or dynamic search [1]. (4) The ultimate target for IR community is to help users succeed in tasks instead of merely making the users *feel* satisfied. Future studies should focus on how to help users succeed through the understanding of cognitive biases.

REFERENCES

[1] Ameer Albahem, Damiano Spina, Falk Scholer, and Lawrence Cavedon. 2019. Meta-evaluation of Dynamic Search: How Do Metrics Capture Topical Relevance, Diversity and User Effort? *Advances in Information Retrieval 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I*, 14.

[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] 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

[5] 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* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 903–912. https://doi.org/10.1145/2009916.2010037

[6] 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

[7] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of Online and Offline Web Search Evaluation Metrics. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017).

[8] 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

[9] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654

[10] 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

[11] Kalervo Jarvelin, Susan L Price, Lois L. M. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted Cumulated Gain based Evaluation of Multiple-Query IR Sessions. In *Advances in Information Retrieval: 30th European Conference on IR Research, Ecir 2008, Glasgow, UK, March 30 – April 3, 2008*. 4–15.

[12] Diane Kelly, Chirag Shah, Cassidy R. Sugimoto, Earl W. Bailey, Rachael A. Clemens, Ann K. Irvine, Nicholas A. Johnson, Weimao Ke, Sanghee Oh, Anezka Poljakova, Marcos A. Rodriguez, Megan G. van Noord, and Yan Zhang. 2008. Effects of Performance Feedback on Users' Evaluations of an Interactive IR System. In *Proceedings of the Second International Symposium on Information Interaction in Context* (London, United Kingdom) *(IIiX '08)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/1414694.1414712

[13] 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 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420130102

[14] 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

[15] Annie Y.S. Lau and Enrico W. Coiera. 2009. Can Cognitive Biases during Consumer Health Information Searches Be Reduced to Improve Decision Making? *Journal of the American Medical Informatics Association* 16, 1 (01 2009), 54–65. https://doi.org/10.1197/jamia.M2557 arXiv:https://academic.oup.com/jamia/article-pdf/16/1/54/2572282/16-1-54.pdf

[16] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2019. From a User Model for Query Sessions to Session Rank Biased Precision (SRBP). In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (Santa

Clara, CA, USA) *(ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 109–116. https://doi.org/10.1145/3341981.3344216

[17] 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

[18] 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* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 923–931. https://doi.org/10.1145/3292500.3330981

[19] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2015. INST: An Adaptive Metric for Information Retrieval Evaluation. In *Proceedings of the 20th Australasian Document Computing Symposium* (Parramatta, NSW, Australia) *(ADCS '15)*. Association for Computing Machinery, New York, NY, USA, Article 5, 4 pages. https://doi.org/10.1145/2838931.2838938

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 arXiv:https://www.science.org/doi/pdf/10.1126/science.185.4157.1124

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

[31] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5 (1992), 297–323.

[32] 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

[33] Alfan Farizki Wicaksono and Alistair Moffat. 2021. Modeling search and session effectiveness. *Information Processing Management* 58, 4 (2021), 102601. https://doi.org/10.1016/j.ipm.2021.102601

[34] Alfan Farizki Wicaksono, Alistair Moffat, and Justin Zobel. 2019. Modeling User Actions in Job Search. In *ECIR*.

[35] 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

[36] 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

[37] Fan Zhang, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Cascade or Recency: Constructing Better Evaluation Metrics for Session Search. 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, 389–398. https://doi.org/10.1145/3397271.3401163

[38] Yuye Zhang, Laurence Anthony F. Park, and Alistair Moffat. 2009. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval* 13 (2009), 46–69.