Chapter 3
From Rational Agent to Human with Bounded Rationality

Abstract To clarify and address the errors that occur in model parameter estimations and behavior predictions, researchers may need to start with investigating the hidden gaps between rational agent and human that are ignored or covered by oversimplified model assumptions. These gaps could occur in both factual, ad hoc retrieval and whole-session interactive retrieval and involve multiple aspects of search interactions, including not only user characteristics and their search strategies but also search task features, search interfaces, as well as situational factors. In this chapter, we summarize and briefly discuss the gaps we identified between simplified rational assumptions and empirically confirmed human biases and then propose a preliminary bias-aware evaluation framework to describe the connections between different stages of search sessions and diverse types of biases. The identified gaps will serve as the basis for developing bias-aware user models, search systems, and evaluation metrics.

3.1 Background

Formally modeling users often serves as a fundamental step toward predicting users' search activities and evaluating varying aspects of search system performances. Building formal models also facilitates the simulation of user actions and associated system responses, which supports the generation of synthetic evaluation data and enhances the reproducibility and reusability of offline IR evaluation materials. However, from a user-oriented perspective, as discussed in Chap. 2, previous research from both IR and other related fields (e.g., information seeking, human-computer interaction, behavioral economics, and decision-making) calls into question the fundamentals of existing IR user models of varying types (e.g., Agosto, 2002; Azzopardi, 2021; Barnes, 1984; Charness & Dave, 2017; Eickhoff, 2018; Kahneman, 2003; Liu & Han, 2020) and demands revisiting the implicit assumptions upon which formal models and evaluation measures were built. In general, boundedly rational users may not be able to perform accurate computation tasks and complex comparisons among available options due to limited cognitive resources and insufficient information regarding the problematic situation. As a result, users

J. Liu, *A Behavioral Economics Approach to Interactive Information Retrieval*, The Information Retrieval Series 48, https://doi.org/10.1007/978-3-031-23229-9_3

under the impacts of multilevel biases and situational restrictions usually rely on certain *mental shortcuts* for addressing most of the tasks and do not always pursue theoretically optimal outcomes as it is assumed in most formal models and offline evaluation measures. Moreover, these mental shortcuts or biased decision-making strategies may be affected and even reinforced by diverse components of search systems (e.g., ranked search result lists, query recommendations based on different rules, vertical search results) and algorithmic biases in IR, especially in search result ranking (Diaz et al., 2020; Ekstrand et al., 2019; Gao & Shah, 2019).

To clarify and address the errors that occur in model parameter estimations and behavior predictions, researchers need to start with investigating the gaps between rational agent and human that are ignored or covered by oversimplified model assumptions. These gaps could occur in both factual, ad hoc retrieval and whole-session interactive retrieval and involves multiple aspects of search interactions, including not only user characteristics and their search strategies but also search task features, search interfaces, as well as situational factors. Although investigating these implicit gaps alone do not guarantee improved results in behavior prediction and user-oriented search evaluation, it serves as a critical starting point toward developing more accurate, behaviorally realistic formal user models. Building these bias-aware user models may also increase the transparency of advanced machine-learning (ML)-based user models trained based on large-scale behavioral logs and help explain the hidden behavioral traces behind improved performances in relevance estimation, behavioral prediction, and IR evaluation.

3.2 Gaps Between Biased Users and Formal User Models

Regarding click models, although different models have been developed based on diverse assumptions, user models, and parameters, most of the assumptions are associated with the two widely examined components: attractiveness and examination (Chuklin et al., 2015; Zhang et al., 2022). Many of these assumptions, especially the ones regarding the impacts of rank positions, document features and browsing sequences and share similar characteristics with that of offline evaluation metrics (Zhang et al., 2017). For instance, multiple click models were built upon examination hypothesis (e.g., cascade model, position-based model, user browsing model) and assumes that the probability of examination is largely affected by the rank position of retrieved search results. This assumption echoes the implicit user models behind many offline metrics, such as nDCG and ERR, which assign lower weights (or higher stopping or skipping rate) to lower ranked documents. The prediction target of click models, clickthrough events, is associated with clicking activities as online evaluation metrics. Thus, the biases and cognitive limits that affect the robustness of evaluation metrics may also generate impacts on the performance of click models through creating unexpected variations in levels of attractiveness and examination probabilities in SERP browsing. Given these overlaps and similarities, we combined clicks and offline evaluation metrics into the same section of gap

analysis to avoid unnecessary duplications. Specifically, we include the discussion on clickthrough events, levels of attractiveness, and examination probabilities under the broader scope of online behavioral metrics, which also include dwell time features, cursor movement, as well as other browsing-related user search activities.

Table 3.1 summarizes the gaps between user biases reported in Chap. 1 and the mainstream behavioral (including clickthrough activities by users) and offline evaluation metrics. By summarizing these gaps, this chapter reveals the inconsistencies between simplified user models and real-life bounded rationality users revealed (both explicitly and implicitly) by existing user studies and evaluation experiments. The first column lists the widely studied human biases, and we interpreted the bias-aware user models in the context of information searching in the second column. We also provide references to related empirical research or conceptual papers for each identified bias so that readers can refer to the original definitions of user biases and bounded rationality through these citations.

Nevertheless, the exploration on boundedly rational users and bias-aware evaluation by us as a research community is still far from complete. Also, at the operationalization level, how to represent different factors associated with boundedly rational search strategies and estimate their impacts in evaluation remains to be a major challenge. Before addressing these broader challenges at both empirical and methodological levels, we need to first review and synthesize the related research progresses made by information seeking, IR, as well as behavioral science researchers on scattered topics and individual specific problems.

To facilitate the analysis, we present the user models derived from the empirical research on user biases and explain how these derived user behavioral models contradict with the existing metrics. The metrics analyzed in the table include not only the basic metrics, such as *nDCG*, *rank-biased precision* (RBP), as well as *metrics@K* but also more recent metrics and evaluation frameworks, such as C/W/L (Azzopardi et al., 2021; Moffat et al., 2017). To cover a broad range of evaluation measures, our analysis includes both process-oriented evaluation measures, which focus on search process and behaviors (e.g., querying, clicking, dwell time measures) (Hofmann et al., 2016), and outcome-oriented evaluation measures, which characterize search result features (e.g., relevance-based metrics, usefulness) (Clarke et al., 2020; Harman, 2011; Sanderson, 2010; White, 2016). Note that process-oriented measures (third column) also include assumptions of click actions, which also involve and overlap some components of click models (e.g., attractiveness, examination probabilities) discussed in the previous chapter.

In the third and fourth columns, we explain how each user bias and the associated user model contradict with the assumptions and parameter setups of different evaluation measures. These gaps are identified based upon both Zhang et al. (2017)'s general evaluation framework and IR studies on individual metrics (e.g., Azzopardi et al., 2018; Chapelle et al., 2009; Moffat & Zobel, 2008). Under each type of user models and metrics, we explain how they conflict with each individual user biases identified in behavioral experiments and what would be the possible changes and adjustments we could make on existing model components or under current frameworks (e.g., rank-based discount rate) to incorporate user biases into

Table 3.1 Gaps between user biases and IR metrics

Empirically confirmed user biases	Derived bias-aware user model/feature	Process-oriented measure (e.g., click, browsing, and dwell time)	Outcome-oriented measure
Reference dependence (Tversky & Kahneman, 1991)	Users evaluate their relative search gains and losses (e.g., increased search efforts, decreased search efficiency) according to certain reference points or pre-search expectations, not merely final outcomes	All process-oriented/behavioral measures: users may evaluate different search actions based on losses and gains with respect to a reference point or expectation developed in previous interactions, rather than final outcome values. For instance, initially encountered high-quality results may lead to higher thresholds of *attractiveness*, *examination*, as well as *clicks* for following results both within the same SERP and in other query segments	Rank-biased precision (RBP), expected reciprocal rank (ERR): stopping and skipping rate may vary across different ranks due to the changes in references and expectations Normalized discounted cumulated gain (nDCG), utility accumulation model of C/W/L: different users may have different utility discount rates at different moments or states of search sessions due to the variations in reference points. Thus, it may be of help to design and test different utility discount models based upon users' task states and state transition patterns within sessions (Liu et al., 2020a; Liu & Yu, 2021)
Loss aversion (Tversky & Kahneman, 1991; Kahneman, 2003)	Users' evaluations of different search results and search strategies are more sensitive to the variations in perceived or estimated search losses than to gains, which may lead to changes in subsequent search and evaluation tactics	All process-oriented/behavioral measures: users tend to be more sensitive to perceived losses and try to avoid search actions or results that are likely to result in search time losses and reduced cognitive resources (e.g., increased search efforts, limited useful information)	RBP, ERR: users may have a higher stopping and skipping rate at a rank where they perceive a relatively loss (e.g., less relevant title, confusing search snippet). nDCG, document utility model of C/W/L: a perceived loss at a rank or a search iteration may lead to an increased gain discount rate
	Users' judgments on different search actions	The specific forms and narratives of search	RBP, nDCG, metrics@K,

(continued)

Table 3.1 (continued)

Empirically confirmed user biases	Derived bias-aware user model/feature	Process-oriented measure (e.g., click, browsing, and dwell time)	Outcome-oriented measure
Framing effects (Nelson et al., 1997)	and results are affected not only by the nature of the options but also the ways in which they are framed and presented	result snippets (e.g., organic search results, images, news verticals) with similar or same contents may result in different levels of attractiveness, examination probability, clickthrough rate, as well as document dwell time (if clicked)	document utility model of C/W/L: a result framed or perceived as a loss (relative drops in result quality and clarity, increased difficulty in comprehension) may incur significantly higher stopping and discount rates in browsing processes
Salience bias (Tiefenbeck et al., 2018)	Users often focus on and are more likely to be attracted by the information objects that are especially remarkable and more salient than other objects	All behavioral measures: users may spend longer dwell time and have higher examination and click probabilities on visually salient items and objects	RBP, nDCG, metrics@K, document utility model of C/W/L: salient items (e.g., vertical results, knowledge cards, organic search results near vertical blocks) may have higher click rates and lower stopping or discount rates; salient items, with similar contents to others, may have a higher estimated relevance. Note that these assumptions associated with salience bias partially echo that of vertical click models
Peak-end rule, position bias, order effects; primacy and recency (Kahneman, 2003)	In listwise and session evaluations, a user's overall experience is significantly affected by peak and end/recent points of local experiences	Session behavioral measures: the local search behavior and experience measures at peak and end search moments can better represent session-level experience than traditional sum and average-value-based measures. Knowledge of this bias conflict with the assumption that all moments or search iterations are	Session-level measures, utility accumulation model of C/W/L: the local search result metrics at peak and end search moments can better represent session experience than sum and average-value-based measures, e.g., mean average precision (MAP); users' in situ perception (e.g., query-level search

(continued)

Table 3.1 (continued)

Empirically confirmed user biases	Derived bias-aware user model/feature	Process-oriented measure (e.g., click, browsing, and dwell time)	Outcome-oriented measure
		equally important for a session	satisfaction) at peak and end search moments can better represent session experience
Decoy effect (Zhang & Zhang, 2007)	Users change their preference between different search results when presented with a third option (the decoy) that is asymmetrically dominated	Clicks, browsing (e.g., scrolls, mouse, and eye movements), and dwell time measures: users' implicit feedback (e.g., dwell time) on two similar results could be affected by an implicit decoy option in decision-making	RBP, nDCG, metrics@K, user stopping model of C/W/L: a decoy search result may affect gain discount and stopping rates at adjacent rank positions. Researchers need to look at the implicit connections among different results on a SERP
Priming effect (Tipper, 1985)	A user's exposure to a search result subconsciously affects their evaluation of a subsequent result or recommendation	Previously encountered search result snippets may affect the probability of attractiveness and examination on subsequent search result snippets presented on SERPs. The changes in attractiveness and examination probabilities may also result in variations in clickthrough rates	All relevance- and usefulness-based measures: the relevance and usefulness levels of an encountered landing page may affect the user's evaluation criteria (e.g., thresholds for relevant and usefulness judgment) in following search interactions
Confirmation bias, anchoring bias (Nickerson, 1998)	Users tend to accept the search results that are consistent with their prior beliefs, expected conclusions, and/or the initially encountered search results or documents	Clicks, browsing, and dwell time measures: users tend to spend more time and attention on results that confirm their existing beliefs and expectations; results that echo existing beliefs and in situ search expectations may enjoy a higher clickthrough rate and dwell time	RBP, nDCG, metrics@K, user stopping model and document utility model of C/W/L: lower ranked results and/or later reviewed results that confirm existing beliefs or initially encountered results may be associated with a lower discount rate and skip rate. Thus, researchers may need to measure

(continued)

Table 3.1 (continued)

Empirically confirmed user biases	Derived bias-aware user model/feature	Process-oriented measure (e.g., click, browsing, and dwell time)	Outcome-oriented measure
			the relevance of current result to both the overall topic or query and the user's anchoring point
Ambiguity effects, risk aversion (Pratt, 1978)	Users prefer search results and recommendations with low uncertainty or ambiguity (e.g., Web pages that present "clear facts" or direct answers to queries)	Clicks: users may have a lower click rate on results that seem to be uncertain or ambiguous (although these results may be useful for completing open-ended, intellectually challenging search tasks) Dwell time measures: users may tend to spend less time, have a higher skip rate, or underestimate relevance on seemingly ambiguous results	RBP, nDCG, ERR, metrics@K, document utility model of C/W/L: users may have a higher skip rate and gain discount rate on search results that seem to be ambiguous or uncertain. Thus, the specific discount rate could be written as a function of document relevance, rank position, and content ambiguity of both document itself and the corresponding search result snippet on the SERP
Theory of satisficing (Simon, 1955)	Users tend to stop at satisficing or "good enough" search results, rather than keeping exploring potentially better search results or seeking for theoretically optimized search outcomes	Clicks, browsing, and dwell time measures: users' criteria for satisficing results are affected by their prior interactions and in situ search expectations. Increased search efforts or frustrations may lower the threshold of satisficing	RBP, nDCG, ERR, metrics@K, user stopping model of C/W/L: Instead of having a preexisting cost budget in mind, a user may stop searching once a satisficing result is encountered during SERP browsing. The specific satisficing threshold, however, may vary across different search sessions, and individuals and may be related to both pre-search expectation and in situ outcomes and estimated difficulty

(continued)

Table 3.1 (continued)

Empirically confirmed user biases	Derived bias-aware user model/feature	Process-oriented measure (e.g., click, browsing, and dwell time)	Outcome-oriented measure
Bandwagon effect (Schmitt-Beck, 2015)	Users tend to seek for and accept certain search strategies and search results simply because other users are using them	Click, browsing, and dwell time measures: in search and evaluation contexts where users can observe other users' reactions (e.g., ratings, retweets, and comments in social information seeking), a user may be more likely to react to (e.g., click) or given a higher rating on results and recommendations that are broadly accepted by other users	In social information seeking and search, offline evaluation metrics may need to take into account the impacts of social factors presented in retrieval process, in addition to the widely studied factors, such as features of search result snippets, rank positions, and document relevance. This effect is less relevant in traditional Cranfield experiments, where researchers treat searches as individual, separate events

formal models. More detailed discussions on the extension of model assumptions based on the knowledge of bounded rationality and the development of user models beyond current structures would be provided in the following chapters. In this chapter, our goal is to provide an overview of the major gaps between identified biases and existing formal models, rather than examining the specific metric revision or model extension plan associated with each user bias.

Research on boundedly rational users should not be treated as an independent research topic that is separated from traditional formal models. Similar to the aim of behavioral economics within broader economics research problem space (cf. Kahneman, 2003), our goal behind emphasizing user biases and identifying gaps between rational agents and human is not to replace existing user models or negate the value of widely applied offline evaluation metrics. Instead, as it is presented in Fig. 3.1, the knowledge of these gaps will allow researchers to extend and further generalize existing formal models and metrics in a user-oriented, bias-aware manner, with existing metrics being a special simplified or ideal application scenario (with no or minimized impacts from human biases and situational factors). In other words, the existing formal models and simplified assumptions could be used as the computational basis for incorporating new parameters and representations of human and situational factors and for developing more sophisticated user models. The extent to which the enhanced bias-aware user models could capture the search and judgment strategies of real users depends on both the empirical knowledge of

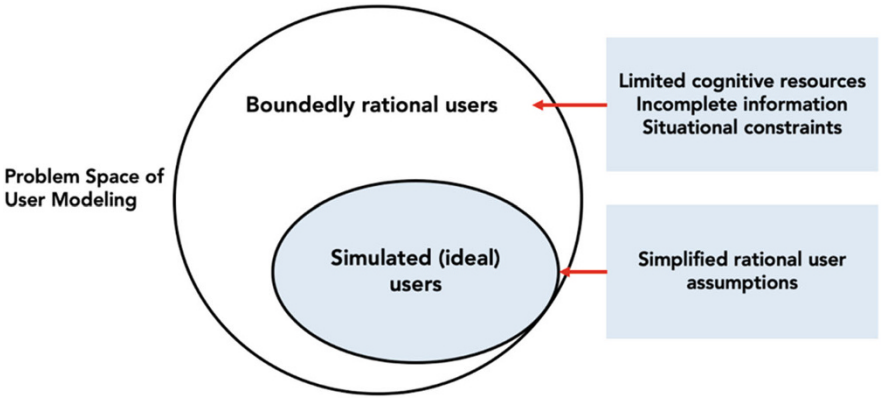

Fig. 3.1 Simulated ideal user and boundedly rational user in IR experiments

human bounded rationality from multiple disciplines and the available resources for model training and validation.

In addition, at practical level, identifying these implicit gaps in research will also inform the design of adaptive, personalized search and recommendation systems that can take into account both algorithmic biases (e.g., Ekstrand et al., 2019; Gao & Shah, 2019) and human biases in search interactions (e.g., Azzopardi, 2021; Liu & Han, 2020). Taking this bias-aware perspective into user modeling, we could further enhance the effectiveness and *multi-dimensional fairness* of existing adaptive and interactive information systems (e.g., Liu & Shah, 2022; Luo et al., 2014; Voorhees, 2008; Zhang et al., 2020a, b).

The summary presented In Table 3.1 focuses on a series of widely used behavioral and offline evaluation metrics proposed and tested in previous research and is by no means exhaustive. Instead, our goal in this chapter is to present and illustrate major *metric-bias gaps* based on the discussions on assumptions and formal user models in Chap. 2 and inform the design of bias-aware user modeling and evaluation framework built upon existing behavioral measures and offline evaluation metrics. The descriptions of derived user models (second column) are developed based on the definitions and empirical evidences on each identified user biases in the first column. More details regarding findings from behavioral experiments, related concepts, and theories, as well as similar cognitive or perceptual biases will be provided and discussed in Chap. 4. In this chapter, our hope is that the readers can have a flavor of existing research on human biases that lead to boundedly rational decisions, as well as their conflicts with the assumptions of formal IR models.

Table 3.1 presents a series of basic boundedly rational user models derived from the knowledge of user biases and bounded rationality in decision-making (e.g., Kahneman, 2003) and points out the ways in which they may conflict with existing components of click models, online behavioral metrics, as well as offline outcome-oriented metrics. The identified gaps between user models associated with existing click models and evaluation metrics and knowledge of human bounded rationality

pave new paths toward developing more behaviorally accurate and practically useful prediction models and evaluation measures.

In general, Table 3.1 can serve as a checklist or initial research agenda for graduate students and young researchers to explore available research topics and develop models of critical user biases in the context of interactive IR. For instance, given the knowledge about *reference dependence biases* from behavioral economics research, researchers could redesign the utility discount rates and formulate it as a function of not only query and rank positions but also the dynamic reference level within the current search session. In addition, with respect to *salience bias*, researchers and system designers need to go beyond traditional rank position factor and take into account the impacts of other system output factors and items and examine their levels of saliency compared to adjacent search results and model the possible perceptual biases associated with the visually more salient items. Similar to these biases, exploring the impacts of other cognitive and perceptual biases can also enhance our understanding of search decision-making and potentially pave ways toward more useful user models.

Beyond individual human biases and cognitive limits, it is also critical to explore the in situ interactions among different types of human biases and investigate the ways in which they are affected by algorithmic biases reflected in ranked result lists and jointly decide local search decisions (e.g., query reformulation, clicking, search stopping) and global perceptions and judgments (e.g., whole-session user satisfaction, perceived level of search and task success).

3.3 Hidden Problems Behind Metric-Bias Gaps

Exploring and clarifying the gaps between formal user models (especially the associated implicitly made assumptions) and human biases can help researchers understand and explain different aspects and types of bounded rationality in search-related decision-making. Also, our investigation on the basic assumptions and hidden gaps offers an opportunity to revisit and reflect on the fundamentals of the established IR models, metrics, and the ranking algorithms designed and trained based upon them. Although different user biases, user models, and metrics take different forms and are applied in varying ways, they share many similarities in behavioral and perceptual origins and can be grouped into a small set of gap categories. Specifically, most of the metric-bias gaps (especially the ones related to evaluation and judgment) discussed above are associated with *three main problems*:

- *Problem 1*: dynamic and subjective nature of users' *perceived* rewards and costs, which usually deviate from actual behavior-based events and simulated rewards and costs in click models and evaluation metrics
- *Problem 2*: changing evaluation criteria and thresholds on document relevance, usefulness, and other related dimensions of evaluation across different moments and states of interactive search sessions

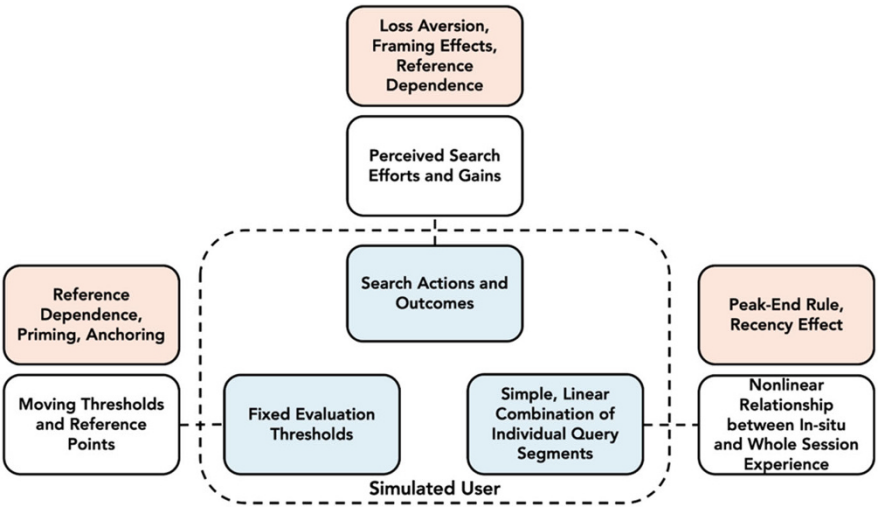

Fig. 3.2 Some gaps between simulated user and boundedly rational user

- *Problem 3*: nonlinear relationship between in situ local evaluation (e.g., query-level search gains, perceived cognitive loads and efforts) and whole-session search evaluation (e.g., session-level user engagement, perceived level of success, and search satisfaction)

The three problems that illustrate the some of the major gaps and conflicts between simulated, formal users and real-life boundedly rational users are summarized in Fig. 3.2. Regarding *Problem 1*, according to findings from behavioral research on reference dependence and loss aversion, users' perceived rewards and costs, which usually serve as the basis for subsequent decision-making, are formed based on dynamic reference points. This thesis conflicts with the final-value-based measures (e.g., total number of clicks and dwell time on SERPs, browsing and cursor movement distances, average nDCG) and assumptions on static costs and gains in search interactions (Azzopardi, 2011, 2014; Azzopardi & Zuccon, 2016). Also, the idea and findings regarding the impacts of reference dependence, anchoring, and framing also challenge the commonly used assumption on cost budget (e.g., metrics@K) as users' in situ perceptions of search cost and implicit acceptable gain-cost ratio may vary over time due to the changes in references and expectations. Built upon the common reference points identified in behavioral economics research (e.g., Kahneman, 2003; Markle et al., 2018; Martin, 2017; Tversky & Kahneman, 1991), Liu and Han (2020) developed a variety of estimated reference-dependent measures and demonstrated their contributions to predicting user behavior and satisfaction. However, how users actually evaluate costs and rewards (e.g., how much time cost equals to or is comparable to the benefits from relevant results) still remains ambiguous and would require further user study efforts to address. To extend existing formal models and evaluation measures, researchers may need to

examine more possible reference points or learn personalized reference levels and in situ expectations regarding information gains and efforts from individual users' search interactions and relevance judgments (Liu & Han, 2020).

Taking a step back from diverse evaluation measures, *Problem 2* calls for a revisit on the fundamental ground-truth measures (e.g., relevance, usefulness, user satisfaction) based upon which a large body of outcome-oriented evaluation measures were constructed, especially in the context of whole-session search evaluation. For instance, according to Scholer et al. (2013), users' relevance criteria vary over time and are largely affected by the quality of documents they evaluated in prior search iterations (i.e., threshold priming). Users who are exposed to only non-relevant documents in early sessions tend to assign significantly higher average relevance scores to the documents in later sessions, compared to the users who are exposed to highly relevant documents in early sessions. Thus, to obtain more balanced, unbiased assessment results and avoid the impacts of priming, researchers should expose expert assessors to multiple levels or broader range of relevance score levels early in the evaluation process. This early exposure to diversified documents will help assessors better calibrate the relevance thresholds for judgments. Although the original experiments on this relevance threshold priming effects were conducted in controlled evaluation-only settings, it is possible that threshold priming, as a form of reference dependence in evaluation, also exists in real-life search interactions and may affect not only explicit judgments and evaluations but also implicit feedbacks (e.g., dwell time, cursor movement, browsing, and examination on search results). Also, as an evidence of the complexity of user evaluation, Scholer et al. (2013) also found that users struggled to base their judgments merely on topical relevance and clearly block out the effects of cognitive, situational, and affective relevance. This result indicates that the implicit changes of evaluation thresholds is multidimensional in nature and different dimensions of judgment may interact with each other.

In addition to controlled evaluation-oriented settings, some researchers also examined in situ relevance judgment (e.g., Jiang et al., 2017) and explored the interactions between relevance judgment and usefulness annotations during Web search sessions (Mao et al., 2016). These studies demonstrate the variations in user perceptions and dynamic evaluation thresholds and thereby partially explain why traditional relevance-based evaluation metrics built upon simplified models are not always well aligned with in situ user satisfaction, especially in complex, intellectually challenging search tasks. Compared to document relevance, usefulness as a ground-truth label has the potential to achieve better performance in estimating users' actual search experience and properly evaluating system performances in a user-oriented manner (Cole et al., 2009). External assessors are capable of annotating document usefulness when offered more information about the search context (Mao et al., 2016). However, understanding the nature of usefulness (especially its connections to individual differences and preferences), developing standard and unbiased usefulness-based measures, and applying them in large-scale reproducible experiments are still open challenges to the IR research community (Liu, 2022).

These fundamental challenges in IR evaluation are often bypassed in controlled relevance-based evaluation experiments and user simulations.

Due to the mixed effects generated from multiple sources (e.g., individual human biases, task characteristics, search states, and in situ search session experience), the annotation-based ground-truth measures in IR may act as *moving* targets, rather than *fixed* optimal points that often assumed to be consistent across different search sessions and experimental settings. This dynamic nature introduces fundamental challenges to user-aware IR evaluation and may cause systematic errors in search system evaluation across different contexts. As a result, research on standardizing the documentation and reuse of interactive IR evaluation resources (e.g., tasks, search interaction logs, user judgments, as well as trained models) still face various challenges (Gäde et al., 2021; Liu, 2022). To address this issue and enhance the robustness of IR evaluation, researchers need to further explore the role and impacts of individual differences, especially the systematic user biases and situational limits (e.g., time limit, available system support, quality of information) and capture the systematic effects hidden in seemingly random errors for achieving more accurate user modeling and realistic search system evaluation.

According to findings on anchoring bias (e.g., Chen et al., 2022), users' existing beliefs, biases, and initially encountered information have significant impacts on their subsequent judgments of document usefulness and credibility and information use behavior in search interactions (White, 2013). The variations in in situ evaluation thresholds and users' search expectations also affect users' search stopping and skipping strategies, which calls for revisits and adjustments on related metrics and parameters (e.g., nDCG, ERR, user stopping model of C/W/L), especially in user-oriented session evaluation. Given these findings, researchers can assign adjusted *weights* to documents judged at different points of search sessions to mitigate the impact of user biases on relevance and usefulness labeling and reliability of IR evaluations. To facilitate user-aware reproducible IR evaluation experiments, the possible changes of judgment thresholds, users' references, as well as other impacts associated with bounded rationality should be considered and properly represented as part of the interactive IR test collections (Liu, 2022).

Problem 3 goes beyond individual result evaluation and focuses on the connection between a sequence of local, in situ experience and whole-session evaluation. According to recent research on user biases in search evaluation (Liu & Han, 2020; Liu et al., 2019b), users' experiences at peak and end points usually have higher impacts on session-level evaluations than other search moments. In addition, users' overall experience has no significant correlation with other intuitive search effort measures, such as total dwell time and total number of clicks. This result echoes the findings on peak-end rule from behavioral experiments (e.g., Kahneman, 2003; Sels et al., 2019). Thus, similar to within-SERP evaluation, in interactive session evaluation, the weights of search outcomes and experiences at different moments or under different search states may have largely different impacts on whole-session search experience due to multiple cognitive effects and biases, such as reference dependence biases, peak-end evaluation rule, anchoring biases, and recency effects (Brown & Liu, 2022; Chen et al., 2022; Liu & Han, 2020; Liu et al., 2019b;

Zhang et al., 2020a, b). The specific weight distributions applied in different sessions and evaluation contexts may need to be tailored to different user populations, search task types, as well as distributions of task states and query-level search intentions (Liu et al., 2020a; Mitsui et al., 2017).

In addition to individual click models and evaluation metrics, researchers have also developed a series of formal models to characterize information search interaction. For instance, information foraging theory (IFT) depicts and predicts online information seeking activities based on the assumption that users always try to maximize the rate at which they collect useful information (Pirolli & Card, 1999). This assumption of IFT could be traced back to the economic man assumption behind classical microeconomics theories: individuals have complete, unbiased knowledge about their search costs and gains, and they always seek to optimize the allocation and consumption of limited resources available in order to obtain optimal outcomes. This assumption allows researchers to model the changes of cost-gain ratio in search sessions and calculate expected utility as the basis for evaluating and prediction next-step search decisions, such as continuing browsing, clicking, and search stopping.

Similarly, other researchers have also followed these assumptions in IFT and applied economic models in developing formal models of search gain, cost, and user actions in IR (e.g., Azzopardi, 2011, 2014). Azzopardi (2014) extends search economic theory built in previous works by developing a more comprehensive interaction cost model and derived eight interaction-based hypotheses regarding search behavior. These hypotheses jointly cover different aspects of the interactions among query cost, page cost, assessment cost, snippet cost, assessment probability, and search performance or efficiency. The experimental results obtained on TREC Aquaint Collection show that the economic models of interaction can to some extent predict the observed search behaviors and that the economic approach could provide credible explanations for users' search actions. While the adoption of economic models and assumptions reduce the computational complexity of formal user models in these studies, the assumptions of always maximizing utility contradict with the knowledge of multiple empirically confirmed human biases, such as theory of satisficing, loss aversion, and reference dependence biases (Agosto, 2002; Liu & Han, 2020). Beyond offline system evaluation experiments, researchers also need to examine the components of existing interaction cost models (e.g., costs of formulating queries, reading content pages, browsing search result snippets, and transiting different subtopics) and their deviations from users' perceptions and estimations. Also, when modeling and evaluating search interactions in sessions, researchers need to pay attention to the dynamic gaps between user perception (e.g., perceived time length) and search activities (e.g., actual dwell time on Web pages) (Luo et al., 2017). The outcome-perception gap is associated with both individual differences (e.g., users' tolerance of information uncertainty and tendencies of risk aversion) and in situ changes of search gains, efforts (e.g., relevance of previously examined documents, total elapsed time), as well as search intentions.

Beyond examining specific measures and user models, it is also critical to rethink and revisit the *ground-truth measures* based upon which we evaluate systems and

meta-evaluate evaluation metrics in light of individual users' differences and biases (Liu, 2022). *User satisfaction* as a self-reported measure has been widely applied in interactive IR evaluation experiments (e.g., Chen et al., 2017; Liu & Shah, 2019a; Mao et al., 2016) and information systems evaluation in general (Gatian, 1994; Wixom & Todd, 2005; Zviran & Erlich, 2003). According to the empirical findings on peak-end rule and recency effects, researchers need to re-examine the relationship between in situ *experienced satisfaction* and session-level retrospective *remembered satisfaction* (which may significantly deviate from average or total value of in situ satisfaction scores). Besides, user satisfaction as a multifaceted concept may subject to the influence of multiple interrelated factors, such as document relevance, information understandability, emotional state, and task state in information seeking and retrieval (Liu, 2021). Deconstructing user satisfaction measure into separate dimensions may allow researchers to better capture the dynamic nature of user satisfaction and evaluate the multifaceted contributions of IR systems to users and their search tasks in a more accurate manner.

Apart from user biases, there are also other practical challenges associated with above evaluation measures. For instance, it might be reasonable to assume that users have an implicit or subconscious "cost budget" (e.g., the maximum number of clicks and/or time spent) for a search interaction. As discussed in previous chapters, the idea of cost budget serves as an implicit basis for multiple offline metrics (Zhang et al., 2017). However, it is difficult to accurately estimate users' cost budgets, mainly for three reasons:

1. Different users have different levels of topic familiarity, task urgency, and search literacy.
2. Same user may have different cost budgets under varying search intents. For instance, users may have more flexible budgets under exploration stage but become stricter when they have a well-defined target item for search.
3. A user's perceived cost is not always consistent with objectively measured costs and is subject to the influence of contextual factors, such as time pressure, task difficulty, and users' emotional states.

Luo et al. (2017) found that there are gaps between users' perception of time and actual dwell time and that document relevance can significantly affect users' perception of time and their satisfaction feedbacks.

Also, in relevance and usefulness estimation, researchers usually assume a landing page to be useful if a user spent more than 30 s on reading the page (Chen et al., 2017; White & Huang, 2010) or if the page is clicked by two different users under similar search tasks (Hendahewa & Shah, 2017; Shah & González-Ibáñez, 2011). However, depending on the nature of the motivating task, users' topic familiarity and domain knowledge, and the availability of "direct answers" on SERPs, this assumption, which could be established in laboratory settings, may not always be tenable in real-life search scenarios (Liu & Shah, 2022).

Beyond individuals' information seeking and search contexts, a user's search and evaluation activities are also affected by the information generated and decisions made by other users and social interactions (i.e., Bandwagon effect; Barnfield,

2020). For instance, in the context of social information seeking, users tend to accept information that are widely accepted or used by other users (e.g., tweets that receive a large number of retweets, answers that receive a large number of votes and follow-up comments in social Q&A sites) (Asghar, 2015; Kim et al., 2013). Information from social networks, Q&A sites, and discussion forums is playing an increasingly important role in everyday life tasks and decision-making events (Kairam et al., 2013; Oeldorf-Hirsch et al., 2014). Investigating the role of Bandwagon effect would allow researchers to better understand individuals' information use and decision-making activities.

Although the popularity of information objects may reflect certain aspects of information quality, they may be caused by certain information source exposure biases on the algorithmic side (cf. Diaz et al., 2020), which may end up increasing or reinforcing users' existing biases toward certain pre-search beliefs, perspectives, or political views. To address this issue, next-generation search systems should not only address the task of algorithmic debiasing in ranking and information exposure but also provide *cognitive debiasing* for addressing users' current anchoring and reference dependence biases, which may lead to undesired decision-making outcomes (Croskerry, 2003).

3.4 Preliminary Bias-Aware Interactive User Modeling and Evaluation Framework

This section proposes a general, preliminary bias-aware framework to facilitate the integration of insights regarding user biases with IR research, especially user-oriented search evaluation. Our discussion on the framework includes user behavioral models and assumptions as the foundation, bias-aware extension of online and offline metrics, ground-truth labels and assessors, levels of evaluation (i.e., single-query-level and session/task level), as well as evaluation settings and environments. In Fig. 3.3, we seek to comprehensively cover the overall broad picture and depict the vision of bias-aware user modeling and evaluation. We leave further discussion on the role of each bias and model specifications (e.g., operationalization of costs and rewards, hyperparameters for model learning, structure of loss functions, optimization rules) for the following chapters as well as future research works and experiments. Based on the above discussion on the gaps between formally simulated users and boundedly rational users, the framework presented in Fig. 3.3 can serve as a preliminary work or initial structure within which more detailed user models focusing on different levels and components could be better defined and tested in individual experiments. Chapter 4 will further explain the factors of bounded rationality and human biases presented in Fig. 3.3 and discuss the associated theories and empirical experiments (at both behavioral and neural levels) that support them.

It is worth noting that developing and testing user-oriented bias-aware user models and associated products (e.g., click model, session simulation model, offline

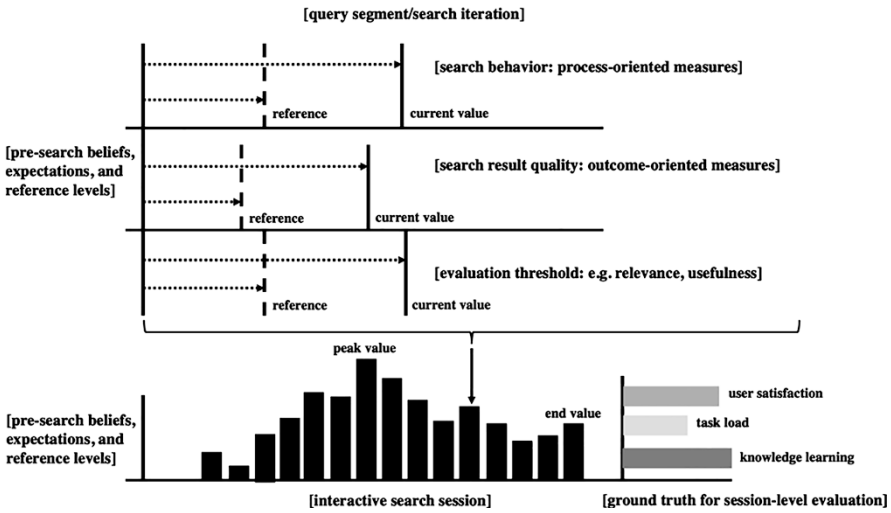

Fig. 3.3 Bias-aware IIR evaluation framework

evaluation metrics) would require extra efforts and more complicated representations than that of traditional Cranfield experiments and offline models (Liu, 2022). To further clarify the knowledge needed for at least partially addressing the gap and specifying bias-aware framework, researchers will need to develop deeper understanding on human biases and bounded rationality in decision-making (Chap. 4) and then leverage the knowledge in revising and improving existing assumptions and formal user models (Chap. 5).

Aligned with the evidence and arguments on metric-bias gaps offered in Sect. 3.2, in our framework, we argue that users' search behaviors and strategies, perceptions of information gains from search results, and evaluation thresholds are affected by both in situ changing references (e.g., reference dependence bias, loss aversion, decoy effect) and pre-search beliefs and expectations (e.g., confirmation bias, anchoring bias). In Fig. 3.3, users' judgments, perceptions, and search decisions are affected by the *delta values* both between current value and moving reference point and between current value and pre-search anchoring point. The perceptual changes or delta values over different dimensions (e.g., perceived search efforts, perceived informational gains) are not only associated with the mathematically calculatable differences in search actions (e.g., changes in number of clicks and dwell time on pages) and search outcomes (e.g., changes in precision and nDCG scores) but also related to the way in which information objects are framed and presented. With similar content and amount of useful information, different types of presentation (e.g., as organic search result or vertical blocks) may attract different levels of user attention and lead to different sizes of changes in perceived efforts.

According to the findings from relevant experiments (e.g., Liu & Han, 2020; Scholer et al., 2013; White, 2013), the weights of different reference points and dimensions in search evaluation vary over time and are associated with search task

type and individual characteristics. The hidden pre-search beliefs and moving reference points may contribute to the difficulty in characterizing and predicting user behavior and judgment in session search. This three-dimension part of the framework covers the core metric-bias gap *Problem 1* and *Problem 2*.

Responding to *Problem 3*, we zoom out from individual query segment level and represent users' biased search action and evaluation at whole-session level. Specifically, in addition to the impacts of pre-search beliefs, general references, and loss aversion biases, users' peak- and end-moment experiences (e.g., maximum number of clicks and SERP dwell time in individual queries, SERP dwell time in last query segment) significantly affect their overall session evaluation (Liu & Han, 2020; Liu et al., 2019a). Also, the starting query may play a significant role in deciding the overall search strategies and facilitate early prediction of the characteristics of the overall search task (Mitsui et al., 2018). Knowledge learned from early prediction of task and session features would offer search systems the opportunity to provide and collect in situ feedback on not only *reactive* support but also *proactive* recommendations (e.g., query modification and document recommendation before predicted search failure, proactive search result re-ranking) and search interventions (Koskela et al., 2018; Liu & Shah, 2019b; Shah, 2018; Vuong et al., 2017).

In contrast to the implicit assumptions behind a variety of sum value and average value-based metrics, Liu and Han (2020) found that the total session dwell time do not have a significant impact on whole-session user experience. This result indicates that there are gaps between real-life users' perception of costs, gains, and efforts and the actual search interactions, which often lead to errors in search cost estimation and user satisfaction prediction. Thus, a session-level evaluation model should reflect the *nonlinear* relationship between in situ search experience and session experience and assign different weights to different reference points, rather than simply applying a monotonically decreasing weight function to all SERPs, search sessions, and search task types.

In addition to the three main problems presented in Fig. 3.2, the behavior-based and final-outcome-based measures also deviate from multiple aspects of search experience, which are often labelled and employed as ground-truth measures in meta-evaluation. Regarding this, it is worth noting that researchers can evaluate systems and meta-evaluate evaluation metrics over various dimensions or against different ground-truth measures, such as user satisfaction (Chen et al., 2017; Liu & Yu, 2021; Mao et al., 2016), task/cognitive load (Gwizdka, 2010; Hu & Kando, 2017), knowledge learning (Syed & Collins-Thompson, 2018; Yu et al., 2018), as well as other experience-related measures (e.g., user engagement; O'Brien & Toms, 2008). Some of these ground-truth measures (e.g., user satisfaction) could be deconstructed into separate facets to facilitate more accurate, reproducible user-oriented evaluations (Liu, 2021; J. Liu et al., 2020b). To achieve this, more detailed scales need to be designed and tested based on the knowledge about users' perceptions, their actual behaviors, as well as the human biases that separate them in search interactions. Many of the action-based and perception-based measures, constructs, and scales from management information systems research (e.g., Venkatesh & Davis, 2000; Venkatesh & Bala, 2008) could be adopted and applied to interactive

IR user modeling and system evaluation experiments across a variety of information access and human-computer interaction scenarios (e.g., desktop search, mobile search, chatbot, and spoken search).

3.5 Summary

To better evaluate IR systems and model users' search interactions, we revisit and reflect on the fundamentals of existing user models discussed in the previous chapter and focus on the implicit gaps between boundedly rational users (especially with respect to their cognitive and perceptual biases) and rational assumptions underpinning a variety of formal models and system evaluation metrics. Furthermore, based on the discussions on the limitations of current formal models, this work develops a general bias-aware evaluation framework to roughly characterize the connections between different components and human biases in search sessions. In contrast to the growing research attention on algorithmic biases (e.g., Ekstrand et al., 2019; Zehlike et al., 2017), users' systematic biases and their impacts have been scarcely studied in information seeking and retrieval (Azzopardi, 2021; Liu & Han, 2020). This is a timely opportunity to develop novel concepts, user models, and evaluation measures based on the insights from behavioral economics for this new branch of IR research and complement current IIR evaluations and user modeling. Also, leveraging the knowledge about human-bounded rationality in information seeking (e.g., Agosto, 2002; Chen, 2021) can strengthen the connection between the descriptive user models developed in information seeking community and computational evaluation metrics and techniques proposed in information retrieval community.

Apart from investigating specific models, concepts, and evaluation measures, we are also interested in exploring and enhancing the potential broader impacts of boundedly rational user models. The ultimate goals for this line of research include (1) combining the knowledge learned from user biases and algorithmic biases studies in user modeling and system evaluation, (2) achieving a more comprehensive understanding on how users' biases interact with algorithmic biases and how these two types of biases jointly shape search interactions, and (3) developing unbiased system supports for critical decision-making, such as vaccination, housing, and financial investments. To achieve these goals and explore specific research problems that could be better solved with a bias-aware perspective, the following chapters will review and introduce the research progresses on bounded rationality in decision-making under uncertainty. Reviewing and synthesizing the theories and findings in this area will also provide a richer empirical basis for building formal models of boundedly rational users and developing bias-aware evaluation metrics. As intelligent interactive systems at large become more ubiquitous and complex, research into user biases and bounded rationality is going to be increasingly valuable and may prove to be computationally useful even beyond the field of interactive IR.

References

Agosto, D. E. (2002). Bounded rationality and satisficing in young people's Web-based decision making. *Journal of the American society for Information Science and Technology, 53*(1), 16–27. https://doi.org/10.1002/asi.10024

Asghar, H. M. (2015). Measuring information seeking through Facebook: Scale development and initial evidence of Information Seeking in Facebook Scale (ISFS). *Computers in Human Behavior, 52*, 259–270. https://doi.org/10.1016/j.chb.2015.06.005

Azzopardi, L. (2011). The economics in interactive information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 15–24). ACM. https://doi.org/10.1145/2009916.2009923

Azzopardi, L. (2014). Modelling interaction with economic models of search. In *Proceedings of the 37th ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 3–12). ACM. https://doi.org/10.1145/2600428.2609574

Azzopardi, L. (2021). Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 27–37). ACM. https://doi.org/10.1145/3406522.3446023

Azzopardi, L., Mackenzie, J., & Moffat, A. (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* (pp. 231–237). ACM. https://doi.org/10.1145/3471158.3472239

Azzopardi, L., Thomas, P., & Craswell, N. (2018). Measuring the utility of search engine result pages: An information foraging based measure. In *Proceedings of the 41st ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 605–614). ACM. https://doi.org/10.1145/3209978.3210027

Azzopardi, L., & Zuccon, G. (2016). An analysis of the cost and benefit of search interactions. In *Proceedings of the 2016 ACM SIGIR International Conference on the Theory of Information Retrieval* (pp. 59–68). ACM. https://doi.org/10.1145/2970398.2970412

Barnes, J. H., Jr. (1984). Cognitive biases and their impact on strategic planning. *Strategic Management Journal, 5*(2), 129–137. https://doi.org/10.1002/smj.4250050204

Barnfield, M. (2020). Think twice before jumping on the bandwagon: Clarifying concepts in research on the bandwagon effect. *Political Studies Review, 18*(4), 553–574. https://doi.org/10.1177/1478929919870691

Brown, T., & Liu, J. (2022). A reference dependence approach to enhancing early prediction of session behavior and satisfaction. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (pp. 1–5). ACM. https://doi.org/10.1145/3529372.3533294

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 621–630). ACM. https://doi.org/10.1145/1645953.1646033

Charness, G., & Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior, 104*, 1–23. https://doi.org/10.1016/j.geb.2017.02.015

Chen, T. (2021). A systematic integrative review of cognitive biases in consumer health information seeking: Emerging perspective of behavioral information research. *Journal of Documentation, 77*(3), 798–823. https://doi.org/10.1108/JD-01-2020-0004

Chen, N., Zhang, F., & Sakai, T. (2022). Constructing better evaluation metrics by incorporating the anchoring effect into the user model. In *Proceedings of the 45rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. https://doi.org/10.1145/3477495.3531953

Chen, Y., Zhou, K., Liu, Y., Zhang, M., & Ma, S. (2017). Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 15–24). ACM. https://doi.org/10.1145/3077136.3080804

Chuklin, A., Markov, I., & Rijke, M. D. (2015). Click models for web search. *Synthesis Lectures on Information concepts, Retrieval, and Services, 7*(3), 1–115. https://doi.org/10.2200/S00654ED1V01Y201507ICR043

Clarke, C. L., Vtyurina, A., & Smucker, M. D. (2020). Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (pp. 185–192). ACM. https://doi.org/10.1145/3409256.3409816

Cole, M., Liu, J., Belkin, N. J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. In *Proceedings of the Third Workshop on Human-Computer Interaction and Information Retrieval* (pp. 1–4). HCIR.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*(8), 775–780.

Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., & Carterette, B. (2020). Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 275–284). ACM. https://doi.org/10.1145/3340531.3411962

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 162–170). ACM. https://doi.org/10.1145/3159652.3159654

Ekstrand, M. D., Burke, R., & Diaz, F. (2019). Fairness and discrimination in retrieval and recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1403–1404). ACM. https://doi.org/10.1145/3331184.3331380

Gäde, M., Koolen, M., Hall, M., Bogers, T., & Petras, V. (2021). A manifesto on resource re-use in interactive information retrieval. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 141–149). ACM. https://doi.org/10.1145/3406522.3446056

Gao, R., & Shah, C. (2019). How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 229–236). ACM. https://doi.org/10.1145/3341981.3344215

Gatian, A. W. (1994). Is user satisfaction a valid measure of system effectiveness? *Information & Management, 26*(3), 119–131. https://doi.org/10.1016/0378-7206(94)90036-1

Gwizdka, J. (2010). Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology, 61*(11), 2167–2187. https://doi.org/10.1002/asi.21385

Harman, D. (2011). Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 3*(2), 1–119. https://doi.org/10.2200/S00368ED1V01Y201105ICR019

Hendahewa, C., & Shah, C. (2017). Evaluating user search trails in exploratory search tasks. *Information Processing & Management, 53*(4), 905–922. https://doi.org/10.1016/j.ipm.2017.04.001

Hofmann, K., Li, L., & Radlinski, F. (2016). Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval, 10*(1), 1–117. https://doi.org/10.1561/1500000051

Hu, X., & Kando, N. (2017). Task complexity and difficulty in music information retrieval. *Journal of the Association for Information Science and Technology, 68*(7), 1711–1723. https://doi.org/10.1002/asi.23803

Jiang, J., He, D., Kelly, D., & Allan, J. (2017). Understanding ephemeral state of relevance. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (pp. 137–146). ACM. https://doi.org/10.1145/3020165.3020176

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kairam, S., Morris, M., Teevan, J., Liebling, D., & Dumais, S. (2013). Towards supporting search over trending events with social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 7, No. 1, pp. 283–292).

Kim, K. S., Sin, S. C. J., & He, Y. (2013). Information seeking through social media: Impact of user characteristics on social media use. *Proceedings of the American Society for Information Science and Technology, 50*(1), 1–4. https://doi.org/10.1002/meet.14505001155

Koskela, M., Luukkonen, P., Ruotsalo, T., Sjöberg, M., & Floréen, P. (2018). Proactive information retrieval by capturing search intent from primary task context. *ACM Transactions on Interactive Intelligent Systems (TIIS), 8*(3), 1–25. https://doi.org/10.1145/3150975

Liu, J. (2021). Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management, 58*(3), 102522. https://doi.org/10.1016/j.ipm.2021.102522

Liu, J. (2022). Toward Cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing & Management, 59*(5), 103007. https://doi.org/10.1016/j.ipm.2022.103007

Liu, J., & Han, F. (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* (pp. 1141–1150). ACM. https://doi.org/10.1145/3397271.3401085

Liu, J., Liu, C., & Belkin, N. J. (2020b). Personalization in text information retrieval: A survey. *Journal of the Association for Information Science and Technology, 71*(3), 349–369. https://doi.org/10.1002/asi.24234

Liu, M., Mao, J., Liu, Y., Zhang, M., & Ma, S. (2019b). Investigating cognitive effects in session-level search user satisfaction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 923–931). ACM. https://doi.org/10.1145/3292500.3330981

Liu, J., Mitsui, M., Belkin, N. J., & Shah, C. (2019a). Task, information seeking intentions, and user behavior: Toward a multi-level understanding of Web search. In *Proceedings of the 2019 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 123–132). ACM. https://doi.org/10.1145/3295750.3298922

Liu, J., Sarkar, S., & Shah, C. (2020a). Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 193–202). ACM. https://doi.org/10.1145/3343413.3377976

Liu, J., & Shah, C. (2019a). Interactive IR user study design, evaluation, and reporting. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 11*(2), i–93. https://doi.org/10.2200/S00923ED1V01Y201905ICR067

Liu, J., & Shah, C. (2019b). Proactive identification of query failure. *Proceedings of the Association for Information Science and Technology, 56*(1), 176–185. https://doi.org/10.1002/pra2.15

Liu, J., & Shah, C. (2022). Leveraging user interaction signals and task state information in adaptively optimizing usefulness-oriented search sessions. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (pp. 1–11). ACM. https://doi.org/10.1145/3529372.3530926

Liu, J., & Yu, R. (2021). State-aware meta-evaluation of evaluation metrics in interactive information retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 3258–3262). ACM. https://doi.org/10.1145/3459637.3482190

Luo, C., Liu, Y., Sakai, T., Zhou, K., Zhang, F., Li, X., & Ma, S. (2017). Does document relevance affect the searcher's perception of time? In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 141–150). ACM. https://doi.org/10.1145/3018661.3018694

Luo, J., Zhang, S., & Yang, H. (2014). Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th International ACM SIGIR conference on Research & Development in Information Retrieval* (pp. 587–596). ACM. https://doi.org/10.1145/2600428.2609629

Mao, J., Liu, Y., Zhou, K., Nie, J. Y., Song, J., Zhang, M., Ma, S., Sun, J., & Luo, H. (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 (pp. 463–472). ACM. https://doi.org/10.1145/2911451.2911507

Markle, A., Wu, G., White, R., & Sackett, A. (2018). Goals as reference points in marathon running: A novel test of reference dependence. *Journal of Risk and Uncertainty, 56*(1), 19–50. https://doi.org/10.1007/s11166-018-9271-9

Martin, V. (2017). When to quit: Narrow bracketing and reference dependence in taxi drivers. *Journal of Economic Behavior & Organization, 144*, 166–187. https://doi.org/10.1016/j.jebo.2017.09.024

Mitsui, M., Liu, J., Belkin, N. J., & Shah, C. (2017). Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1121–1124). ACM. https://doi.org/10.1145/3077136.3080737

Mitsui, M., Liu, J., & Shah, C. (2018). How much is too much? Whole session vs. first query behaviors in task type prediction. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1141–1144). ACM. https://doi.org/10.1145/3209978.3210105

Moffat, A., Bailey, P., Scholer, F., & Thomas, P. (2017). Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems (TOIS), 35*(3), 1–38. https://doi.org/10.1145/3052768

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS), 27*(1), 1–27. https://doi.org/10.1145/1416950.1416952

Nelson, T. E., Oxley, Z. M., & Clawson, R. A. (1997). Toward a psychology of framing effects. *Political Behavior, 19*(3), 221–246. https://doi.org/10.1023/A:1024834831093

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology, 59*(6), 938–955. https://doi.org/10.1002/asi.20801

Oeldorf-Hirsch, A., Hecht, B., Morris, M. R., Teevan, J., & Gergle, D. (2014). To search or to ask: The routing of information needs between traditional search engines and social networks. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 16–27). ACM. https://doi.org/10.1145/2531602.2531706

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

Pratt, J. W. (1978). Risk aversion in the small and in the large. In *Uncertainty in economics* (pp. 59–79). Academic Press. https://doi.org/10.1016/B978-0-12-214850-7.50010-3

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval, 4*(4), 247–375. https://doi.org/10.1561/1500000009

Schmitt-Beck, R. (2015). Bandwagon effect. *The International Encyclopedia of Political Communication, 1–5*. https://doi.org/10.1002/9781118541555.wbiepc015

Scholer, F., Kelly, D., Wu, W. C., Lee, H. S., & Webber, W. (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* (pp. 623–632). ACM. https://doi.org/10.1145/2484028.2484090

Sels, L., Ceulemans, E., & Kuppens, P. (2019). All's well that ends well? A test of the peak-end rule in couples' conflict discussions. *European Journal of Social Psychology, 49*(4), 794–806. https://doi.org/10.1002/ejsp.2547

Shah, C. (2018). Information fostering-being proactive with information seeking and retrieval: Perspective paper. In *Proceedings of the 2018 International ACM SIGIR Conference on Human Information Interaction & Retrieval* (pp. 62–71). ACM. https://doi.org/10.1145/3176349.3176389

Shah, C., & González-Ibáñez, R. (2011). Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 913–922). ACM. https://doi.org/10.1145/2009916.2010038

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118. https://doi.org/10.2307/1884852

Syed, R., & Collins-Thompson, K. (2018). Exploring document retrieval features associated with improved short- and long-term vocabulary learning outcomes. In *Proceedings of the 2018 ACM SIGIR Conference on Human Information Interaction & Retrieval* (pp. 191–200). ACM. https://doi.org/10.1145/3176349.3176397

Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., & Staake, T. (2018). Overcoming salience bias: How real-time feedback fosters resource conservation. *Management Science, 64*(3), 1458–1476. https://doi.org/10.1287/mnsc.2016.2646

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *The Quarterly Journal of Experimental Psychology, 37*(4), 571–590. https://doi.org/10.1080/14640748508400920

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics, 106*(4), 1039–1061. https://doi.org/10.2307/2937956

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences, 39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Voorhees, E. M. (2008). On test collections for adaptive information retrieval. *Information Processing & Management, 44*(6), 1879–1885. https://doi.org/10.1016/j.ipm.2007.12.011

Vuong, T., Jacucci, G., & Ruotsalo, T. (2017). Proactive information retrieval via screen surveillance. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1313–1316). ACM. https://doi.org/10.1145/3077136.3084151

White, R. (2013). Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3–12). ACM. https://doi.org/10.1145/2484028.2484053

White, R. W. (2016). *Interactions with search systems*. Cambridge University Press.

White, R. W., & Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in web logs. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 587–594). ACM. https://doi.org/10.1145/1835449.1835548

Wixom, B. H., & Todd, P. A. (2005). A theoretical integration of user satisfaction and technology acceptance. *Information Systems Research, 16*(1), 85–102. https://doi.org/10.1287/isre.1050.0042

Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., & Dietze, S. (2018). Predicting user knowledge gain in informational search sessions. In *Proceedings of the 41st ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 75–84). ACM. https://doi.org/10.1145/3209978.3210064

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). FA* IR: A fair Top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1569–1578). ACM. https://doi.org/10.1145/3132847.3132938

Zhang, J., Liu, Y., Mao, J., Xie, X., Zhang, M., Ma, S., & Tian, Q. (2022). Global or local: Constructing personalized click models for Web search. In *Proceedings of the ACM Web Conference* (pp. 213–223). ACM. https://doi.org/10.1145/3485447.3511950

Zhang, Y., Liu, X., & Zhai, C. (2017). Information retrieval evaluation as search simulation: A general formal framework for IR evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 193–200). ACM. https://doi.org/10.1145/3121050.3121070

Zhang, F., Mao, J., Liu, Y., Ma, W., Zhang, M., & Ma, S. (2020b). 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* (pp. 389–398). ACM. https://doi.org/10.1145/3397271.3401163

Zhang, T., & Zhang, D. (2007). Agent-based simulation of consumer purchase decision-making and the decoy effect. *Journal of Business Research, 60*(8), 912–922. https://doi.org/10.1016/j.jbusres.2007.02.006

Zhang, W., Zhao, X., Zhao, L., Yin, D., Yang, G. H., & Beutel, A. (2020a). Deep reinforcement learning for information retrieval: Fundamentals and advances. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2468–2471). ACM. https://doi.org/10.1145/3397271.3401467

Zviran, M., & Erlich, Z. (2003). Measuring IS user satisfaction: Review and implications. *Communications of the Association for Information Systems, 12*(1), 5. 10.17705/1CAIS.01205.