$5-4 \neq 4-3$: On the Uneven Gaps between Different Levels of Graded User Satisfaction in Interactive Information Retrieval Evaluation

Jiqun Liu
The University of Oklahoma
Norman, OK, USA 73019
jiqunliu@ou.edu

Fangyuan Han
Xiamen University
Xiamen, Fujian, China 361005
hanxmu@outlook.com

Abstract

Similar to other ground truth measures, graded user satisfaction has been frequently employed as a continuous variable in information retrieval evaluation based on the assumption that intervals between adjacent grades are quantitatively equal. To examine the validity of equal-gap assumption and explore dynamic perceptual thresholds triggering grade changes in search evaluation, we investigate the extent to which users are sensitive to changes in search efforts and outcomes across different gaps of graded satisfaction. Experiments on four user study datasets (15,337 queries) indicate that 1) User satisfaction sensitivity, especially to offline evaluation metrics, changes significantly across gaps in satisfaction scale; 2) the size and direction of changes in sensitivity vary across study settings, search types, and intentions, especially within "$3-5$" scale subrange. This study speaks to the fundamentals of user-centered evaluation and advances the knowledge of heterogeneity in satisfaction sensitivity to search efforts and gains and implicit changes in evaluation thresholds.

Keywords: Information retrieval evaluation, graded user satisfaction, user satisfaction sensitivity

1. Introduction

Search evaluation has been a central topic to information retrieval (IR) research. Motivated by *user-oriented* interactive information retrieval (IIR) approach, a large body of evaluation studies have employed *user satisfaction* as the ground truth measure in evaluating search systems and meta-evaluating the effectiveness of *process-oriented* online metrics and *outcome-oriented* offline metrics (Y. Chen et al., 2017; J. Liu, 2021). According to Kelly (2009), satisfaction in system evaluation refers to the *"fulfillment of a specified desire or goal"*. Thus, level of satisfaction can be understood as the extent to which a system fulfills a user's goal. This definition echoes the goal of IR systems, which is to "support users in accomplishing the task/achieving the goal that led them to engage in information seeking" (Cole et al., 2009).

User satisfaction has been operationalized with five- or seven-point *grading scales* for facilitating statistical comparison and testing, prediction analysis, and regression modeling of varying forms (Zhang, Mao, Liu, Ma, et al., 2020). When IR researchers employ *graded user satisfaction* as a *continuous variable* for regression modeling or in further analysis, a critical assumption being implicitly made is that the gaps between adjacent grades are all *quantitatively equal*, and that users are able to *unbiasedly* divide their evaluation spectrum into several segments of equal length. This assumption relaxes many data distribution restrictions for satisfaction-related statistical modeling and serves as the basis for evaluating search interactions from user perspective (M. Liu et al., 2018; Zhang, Mao, Liu, Ma, et al., 2020). However, since graded satisfaction and many IR measures are not necessarily *interval-scaled*, using improper methods in evaluation experiments may produce questionable decisions on system evaluation (Ferrante et al., 2021). Although there are a small set of models specifically designed for analyzing ordinal variables (e.g. ordered logistic models), they have rarely been applied in evaluation studies and have certain restrictions on data distribution.

Given the key role of user satisfaction in evaluation, one question naturally arises: *to what extent are users' self-reported satisfaction grades sensitive to the changes in search interactions?* In other words, *to what extent does it make sense empirically to treat and analyze graded user satisfaction as a continuous variable?* This question becomes increasingly important as user-oriented evaluation approach has been adopted in multiple modalities of IR, but still remains unanswered.

Apart from statistical reasons, examining behavioral variations corresponding to different grading intervals will also advance our knowledge regarding two problems: (1) how do users actually label their search experiences with a grading scale? and (2) how do users'

URI: https://hdl.handle.net/10125/103141
978-0-9981331-6-4

grade-changing evaluation thresholds (e.g. the criteria triggering changes from score 5 to score 4 and from score 4 to score 3) and sensitivity to search results and efforts vary across different fixed grades? For instance, to what extent does the size and direction of variations in clicking, browsing, and the quality of ranked results that trigger the change from score 5 to score 4 in user satisfaction differ from that of the variations that motivate the transition from score 4 to score 3? Note that these questions are close to the central problems of IIR and HCI research and are also highly relevant to other similar self-reported measures applied in the evaluation of information and computing systems.

Given the open problems above, this study examines potentially uneven gaps in search interactions associated with mathematically equal intervals in graded satisfaction. We define *Satisfaction Sensitivity* as the sensitivity of a user's satisfaction evaluation to the changes in different dimensions of search interactions. Corresponding to the *equal-interval* assumption, our *Null Hypothesis* is: H_0: user satisfaction sensitivity remains *equal* across all intervals. In other words, there is no significant difference among the variations in search interactions associated with different intervals between adjacent grades. To test the null hypothesis and examine satisfaction sensitivity in varying contexts, this study addresses two research questions (RQs):

RQ1: To what extent does a user's satisfaction sensitivity vary across different between-grade gaps?
RQ2: To what extent does satisfaction sensitivity vary across search types and query intentions?

2. Background and Related Works

This section discusses the related works and open challenges that motivate our research on graded user satisfaction in IR evaluation.

2.1. User-Centered IR Evaluation

While Cranfield Paradigm remains to be the mainstream approach to evaluation in text retrieval, increasing research attention has been paid to user-centered evaluation problems and user-related contextual factors, such as motivating tasks, search states, and information seeking intentions (Hofmann et al., 2016; J. Liu, 2021). Under user-centered perspective, IR evaluation goes beyond examining the *topical relevance* of ranked documents to a given query and covers the informational support an IR system provide for facilitating users' task performances and improving interaction experiences (Cole et al., 2009). This extended scope of evaluation breaks the boundary between the user aspect of information seeking and the engineering aspect of IR, and diversifies IR evaluation methods, metrics, and theories, especially with the scales and tools adapted from cognitive psychology, behavioral sciences and economics (e.g. NASA Task Load Index, dual-task method, physiological signals) (Kelly, 2009; J. Liu and Shah, 2019). This user-centered perspective and the associated methods are critical, especially in whole-session IR evaluation where users issue multiple queries for addressing a task and often experience different local search intentions at different moments (J. Liu and Han, 2020). Due to the limitations of Cranfield Paradigm in representing user characteristics and task dimensions (Cole et al., 2009), IIR researchers sought to develop new metrics and procedures for evaluating search sessions (Koolen et al., 2017; J. Liu, 2021).

In addition to evaluating systems with online *process-oriented* and offline *outcome-focused* metrics, researchers have also *meta-evaluated* IR evaluation metrics according to their associations with selected ground truth measures. For instance, Y. Chen et al. (2017) investigated the extent to which different evaluation metrics can reflect user satisfaction in varying search scenarios. Similarly, Xie et al. (2019) developed grid-based evaluation metrics for facilitating image search evaluation and found that the proposed metrics have strong correlations with user satisfaction. Z. Liu et al. (2021) examined three aspects (i.e., reliability, fidelity, and intuitiveness) of conversational search evaluation metrics and developed new metrics that achieve stronger correlation with ultimate user preference and satisfaction than existing metrics. Since a large pile of metrics and systems have been evaluated based upon *graded user satisfaction*, it is critical to look deeper into the moving thresholds behind grade changes and obtain a more comprehensive understanding of this underpinning for IIR evaluation.

2.2. User Satisfaction as Ground Truth

User satisfaction as a widely employed ground truth measure helps researchers define what "better" actually means in IR evaluation and meta-evaluation. At the operationalization level, apart from soliciting qualitative feedback (e.g. through interview transcripts, in-situ users' notes, records from think-aloud experiments) (J. Liu and Shah, 2019), researchers often ask participants to indicate their level of satisfaction using *grading scales* (Y. Chen et al., 2017; Z. Liu et al., 2021). In statistical analysis and prediction modeling, similar to many other self-reported measures, graded user satisfaction has often been assumed to be

a normally distributed continuous variable, which enables researchers to temporarily bypass a variety of distribution restrictions, especially for parametric testing. When comparing user satisfaction across different tasks and users, researchers often use average query-level scores as a representation of session-level satisfaction (Ayuningtyas and Janah, 2018). The implicit assumption is that the predefined intervals between grades correspond to empirically equal variations in satisfaction perception and search interaction. In other words, "$5-4=4-3$", and the score "3.5" indicates the exact half-way point between two adjacent satisfaction levels. Although the grading scales serve as a useful tool for users' in-situ and retrospective annotations, the assumptions above ignore the variations in users' sensitivity to the changes in search across different grades and may increase errors in real-time user behavior and search satisfaction estimation.

3. Methodology

This section introduces the characteristics of our diverse datasets and our analysis methods.

3.1. Datasets

This study selected *four user study datasets of varying types* with graded user satisfaction scores for 15,337 valid query segments in total. Aggregating four datasets allows our study to go beyond one or two specific study settings, compare results across varying conditions, and thereby test the generalizability and levels of context-dependence of the findings. A *query segment* refers to a single search iteration that start with a query, includes all user behaviors and interacted pages associated with the query, and ends before next query.

Each score in our datasets represents a user's level of satisfaction on *a list of search results within the same query segment*, rather than one single document or page. The question designed for collecting satisfaction grade in the four datasets was: *Were you satisfied with the search results in this query?* Participants were asked to record their answers with a 5-point scale, ranging from *unsatisfied / low* to *very satisfied / high*, without the meaning of grades between the two ends being specified.

The four diverse datasets employed here jointly cover both ad hoc retrieval (i.e., THU-2017) and whole-session interactive retrieval and contain category labels for varying search goals (informational, navigational, transactional) and cognitive levels (understand, remember) behind queries. The THU-17 dataset was gathered in a controlled setting where each participant completed a series of no more than 30 tasks. For each task, each participant was presented one predefined query and was asked to judge a list of ten fixed results. Once a participant completed a (single-query) session, they were required to label a satisfaction score for the result list or session.

Among the three session retrieval datasets, THU-KDD19 was collected through a lab study, whereas the TianGong-QRef and TianGong-SS-FSD datasets were gathered in field studies. The THU-KDD19 dataset consists of 450 unique search sessions from nine complex search tasks. Differing from the study settings in THU-17, participants in session retrieval studies can submit their own queries and interact with the retrieved results for completing the assigned tasks. Participants were asked to give a graded satisfaction feedback on each query segment. The two field study datasets were collected through a Chrome extension, which allowed researchers to collect data on search-related activities (e.g. actions, timestamps, URLs, cursor movements) remotely, and an annotation platform, which enabled participants to submit their explicit feedback on document usefulness and query-level satisfaction. The Tiangong-QRef study collected daily search logs and user feedbacks from 50 participants, and the TianGong-SS-FSD study gathered search behavior and evaluation data from 30 participants. Both studies lasted for one month.

All above session datasets have both behavioral data and users' explicit feedback on pages retrieved (e.g. graded relevance and usefulness labels), which allow us to compute online and offline metric values. Although all datasets used five-point satisfaction grade, they employed slightly different ranges for staisfaction labeling (i.e., $0-4$ and $1-5$). In analysis, we transformed all grades to the range of 1 to 5, in order to make our result presentation consistent without changing the nature of the data.

The user studies in which the datasets above were collected focus on their respective research problems (e.g. click modeling, understanding query reformulation behavior) that are completely different from ours. Besides, as separate user studies conducted in different settings, they investigated evaluation problems individually, without having an overarching goal that connects all of them. Combining the datasets together offers us a rich reusable empirical basis to explore the nature of graded user satisfaction and how the change of users' feedback on this is associated with variations in search efforts and outcomes under varying search contexts. In addition, the individual user studies above employed user satisfaction along with other ground truth measures in evaluating search system performance and user interaction experience. In contrast, our study combines them for including a

variety of study settings, search tasks, and intentions, and revisits the ground truth measure based on which a large body of IR evaluation studies were conducted.

This sub-section aims to provide enough details regarding the data collection procedures to facilitate result interpretation. The sample sizes and distribution are provided in Table 1. More details about other aspects of the user studies are offered in the cited references.

3.2. Online Process-oriented and Offline Outcome-oriented Measures

To examine users' satisfaction sensitivity to the changes in different dimensions of search interactions (e.g. longer dwell time on pages, lower nDCG scores, more clicks), this study employed a variety of online and offline evaluation metrics extracted from previous studies (Y. Chen et al., 2017). With these metrics or dimensions, we investigated how satisfaction sensitivity varies across different between-grade intervals, and how these differences in sensitivity differ across varying dimensions. Findings from these analyses can illustrate the heterogeneity in satisfaction sensitivity: The significant divergences or uneven gaps in satisfaction sensitivity may be more evident in some metrics, but less frequent in other metrics. Further analysis will show that these between-gap divergences are also conditional on specific search scenarios and query intents. Note that for each offline evaluation measure in the KDD19 dataset, we had two different versions: *relevance*-based and *usefulness*-based (see Figure 1). Specifically, for each page, participants were asked to annotate both *topical relevance* to the *query* and the *practical usefulness* of the page for accomplishing the *task*. We used different prefixes to differentiate relevance-based (e.g. `QueryPrecision@5`) and usefulness-based (e.g. `TaskPrecision@5`) measures. More details about the measures are provided in Table 2.

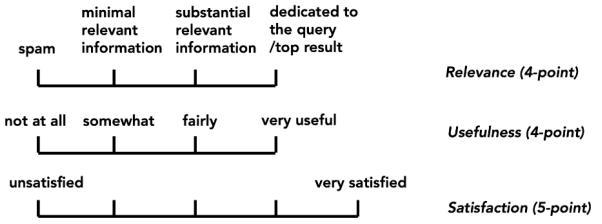


Figure 1. Grading scales of evaluation metrics

3.3. RQ1: Between-Gap Differences

To address RQ1 and fully test the Null Hypothesis proposed in *Section 3*, we first examined between-gap differences in different dimensions of search interaction. To get the sample distribution of gap between grade 4 and grade 5, we randomly sampled 1,000 sample pairs and each sample pair consists of one 5-score satisfaction query segment and one 4-score query segment. Then for both online and offline metrics such as query dwell time and nDCG, we computed differences in metric values between grades. The same sampling process was applied to all other pairs (i.e., 4-3, 3-2, 2-1). After sampling, we get a sample distribution of between-grade intervals for every metric listed in Table 2.

We did *Kruskal–Wallis H test* on all four gaps (5-4, 4-3, 3-2, and 2-1) for each metric individually since the search data were not normally distributed. Then, we did Dunn's post hoc pairwise tests with Benjamini-Hochberg multiple comparisons correction.

3.4. RQ2: By-Group Analysis

To address RQ2, we dug deeper into the variations in satisfaction sensitivity under different conditions. In addition to the cross-gap and cross-metric changes analyzed above, this step examines the differences in satisfaction sensitivity across varying *search types* (i.e., single-query ad hoc retrieval or whole session retrieval) and *query intents* characterized by *search goal* and *cognitive level*. Results from the by-group analysis can further enhance our understanding of the uneven gaps between grades in satisfaction sensitivity.

Following the taxonomies developed by Anderson and Krathwohl (2001) and the IR experiments conducted by Y. Chen et al. (2017), the THU-17 dataset offered two categorical variables (search goal and cognitive level) to characterize query intents. The dataset offers three label values or categories under search goal, including informational goal, navigational goal, and transactional goal, and includes two types of cognitive levels (i.e., understand and remember) defined based on the cognitive complexity associated with the intent behind each query, with the *Understand* level being more complex and intellectually challenging than *Remember* level which only requires memorizing retrieved factual information. The three-category search goal taxonomy was also adopted in classifying queries within sessions in the TianGong-SS-FSD dataset. Note that under this session search dataset, we combined transactional and navigational categories together due to the unbalanced data distribution across search goals.

4. Results

To clarify the contribution of our study, we organize the results from data analyses according to the RQs. In each result table, we ranked metrics according to the corresponding level of *statistical significance of among-gap differences* measured by Kruskal–Wallis

Table 1. Characteristics of Datasets

Dataset	# Sessions	# Queries	User Satisfaction Grade (Number of Queries)				
THU-2017 (Y. Chen et al., 2017)	-	2435	1(95)	2(139)	3(336)	4(638)	5(1227)
THU-KDD19 (M. Liu et al., 2019)	450	1548	1(286)	2(175)	3(265)	4(396)	5(426)
TianGong-QRef (J. Chen et al., 2021)	2353	7479	1(259)	2(743)	3(1032)	4(2080)	5(3365)
TianGong-SS-FSD (Zhang, Mao, Liu, Xie, et al., 2020)	1169	3875	1(221)	2(282)	3(595)	4(1345)	5(1432)

Table 2. Online and Offline Evaluation Metrics

Feature	Description
	Online Metrics - Mouse and keyboard based
QueryLength	Number of terms used in an issued query.
ActionCount	Number of actions (page click, scroll, query formulation).
ClickCount	Number of clicks.
AvgClickRank	Average rank of clicked results.
Clicks@3	Number of clicks between ranks 1-3.
Clicks@5	Number of clicks between ranks 1-5.
Clicks@5+	Number of clicks below rank 5.
ClickDepth	The deepest or lowest rank of clicked result.
	Online Metrics - Dwell time based
SERPtime	Total dwell time on search engine result page (SERP).
AvgContent	Average dwell time on content pages.
TotalContent	Total dwell time on content pages.
QueryDwellTime	Total dwell time within a query segment.
TimeFirstClick	Time delta between the start of session and the first click.
TimeLastClick	Time delta between the start of session and the last click.
	Offline Metrics - Search outcome based
RR	Reciprocal rank.
DCG	Discounted cumulative gain on the first SERP page.
nDCG@3	Normalized discounted cumulative gain (rank 1 to 3).
nDCG@5	Normalized discounted cumulative gain (rank 1 to 5).
nDCG@10	Normalized discounted cumulative gain (rank 1 to 10).
Precision@3	The proportion of relevant pages (rank 1 to 3).
Precision@5	The proportion of relevant pages (rank 1 to 5).
Precision@10	The proportion of relevant pages (rank 1 to 10).
RelDocCount	Number of relevant documents (relevance score > 0).
KeyDocCount	Number of key documents (relevance score > 1).
RBP	Rank-biased precision on first SERP page.
ERR	Expected reciprocal rank on first SERP page.
nERR	Normalized expected reciprocal rank on first SERP page.
INST3	A weighted precision metric (rank 1 to 3).
INST5	A weighted precision metric (rank 1 to 5).
INST10	A weighted precision metric (rank 1 to 10).
Qmeasure	A discounted gain measure with a *persistence* parameter.
	Ground Truth Measure
Q-SAT	satisfaction on a query segment (five-point scale).

(K-W) multi-group tests. Therefore, the same metric might be ranked at different positions under different datasets and conditions. Due to the space limit, in each table, we only presented top-ranked metrics that best reflect the divergence of sensitivity across gaps.

In addition, since the grade scores 3, 4 and 5 were among the most frequent grades assigned to query segments by users and the between-interval differences were significant across multiple metrics (see the frequency distribution in Table 1), our result figures present the number of each type of metrics in the entire metric pool (i.e., where $5-4$ significantly greater than $4-3$; $5-4$ significantly less than $4-3$; no significant between-gap difference) for each dataset. These results jointly highlight the changes in sensitivity across subranges and complement the limited top-ranked metrics by offering an overall picture of metric distributions within the $5-3$ subrange.

4.1. RQ1: Differences in Sensitivity

As the response to RQ1, this section presents the results of K-W multi-group tests and post hoc pairwise tests with multiple comparisons corrections for each of the four datasets in Tables 3, 4, 5, and 6 respectively. In each table, we present the descriptive statistics under each gap, K-W test results for each metric, as well as the pairwise comparison results. Since our analysis includes a long list of metrics, we use different background colors to better show the satisfaction sensitivity changes.

4.1.1. Between-group Differences Overall, the results reject the Null Hypothesis and demonstrate that the between-grade gaps are empirically uneven on diverse metrics, which cast doubts on the equal-interval assumption. In particular, we observed significant between-gap divergences in satisfaction sensitivity on a series of offline evaluation metrics (e.g. `Precision`, `DCG@K`, and `INST`). In the controlled lab whole-session retrieval context (i.e., THU-KDD19), we found that there were significant changes in satisfaction sensitivity associated with both relevance-based and usefulness-based metrics, with several usefulness-based measures being ranked on the top position. This result demonstrates the impacts of perceived usefulness on user satisfaction sensitivity and the importance of understanding usefulness evaluation for developing a more accurate, unbiased session-level evaluation. Moreover, we found that under most metrics associated with significant between-gap differences, the changes in metric scores corresponding to the "5-4" gap are significantly smaller than that of the "4-3" gap, with a few exceptions mainly from ad hoc retrieval dataset. This result suggests that when users were deciding or moving between very satisfied (5) and satisfied (4) in search evaluation, they were very sensitive to the changes in search interactions, especially in the rank order, relevance and usefulness of retrieved results. When it comes to the changes between satisfied (4) and less satisfied (3), satisfaction sensitivity decreased and it would require larger variations in result quality to trigger any score change between 4 and 3 for a user.

Regarding the direction of sensitivity changes, in contrast to most results from "5-4" and "4-3" gaps comparisons, the test results from "4-3" and "3-2" pairwise comparisons are less consistent among

Table 3. Between-Gap Differences: THU-KDD19

Measures	5-4	4-3	3-2	2-1	Kruskal–Wallis posthoc test
task_INST10***	0.12(0.7)	0.17(0.71)	0.16(0.77)	0.59(0.89)	54<43*, 43>32, 32<21***
QuerynERR***	0.0(0.2)	0.0(0.3)	0.1(0.3)	0.23(0.52)	54>43*, 43>32, 32<21***
TasknERR***	0.0(0.3)	0.01(0.29)	0.0(0.3)	0.23(0.55)	54<43*, 43>32, 32<21***
TaskPrecision@10***	0.03(0.2)	0.04(0.21)	0.04(0.25)	0.24(0.4)	54<43, 43>32, 32<21***
TaskPrecision@5***	0.03(0.22)	0.04(0.25)	0.04(0.29)	0.23(0.43)	54<43*, 43>32, 32<21***
QueryPrecision@10***	0.02(0.15)	0.04(0.18)	0.01(0.19)	0.2(0.37)	54<43*, 43>32*, 32<21***
task_INST5***	0.11(0.71)	0.18(0.73)	0.16(0.78)	0.59(0.91)	54<43**, 43>32, 32<21***
query_INST3***	0.06(0.65)	0.22(0.75)	0.08(0.77)	0.58(1.01)	54<43*, 43>32, 32<21***
task_INST3***	0.11(0.74)	0.19(0.76)	0.16(0.81)	0.6(0.93)	54<43*, 43>32, 32<21***
QueryRBP***	0.06(0.47)	0.18(0.54)	0.03(0.57)	0.42(0.69)	54<43**, 43>32**, 32<21***

Note: *:p<.05, **:p<.01, ***:p<.001. Mean value and standard deviation are listed. dark grey : offline metrics. light grey : online metrics.

different datasets. Specifically, in THU-KDD19 and THU-17, most of the between-gap differences in offline metrics are not statistically significant. In the two field study datasets (see Tables 5 and 6), however, we noticed that the differences associated with "4-3" gap are significantly smaller than that of the "3-2" gap, indicating that user satisfaction sensitivity kept decreasing on a variety of metrics. Thus, it would need significantly larger changes in search efforts or search result features to trigger a grade change between 3 and 2. In particular, according to the results in Table 6, we observed a higher sensitivity corresponding to the "4-3" gap compared to the "3-2" gap in terms of click counts and click ranks. This may be because users are sensitive to the perceived costs associated with clicking behavior in natural settings, and this satisfaction sensitivity tends to be higher in the "5-3" grade range. We also observed diverse directions of significant sensitivity changes (especially in offline metrics) between the "3-2" and "2-1" gaps under different datasets.

In addition to the between-gap differences in a diverse set of offline metrics, we also observed significant variations along a series of online behavioral metrics, especially in THU-17 and the TianGong-SS-FSD datasets. For instance, in the TianGong-SS-FSD dataset, user satisfaction sensitivity shows a significant decrease from the "4-3" gap to "3-2" gap on `AvgClickRank`, `ClickDepth`, `ClickCount` and `Clicks@3`, suggesting that users were more sensitive to the changes in experienced efforts (e.g. number of clicks made, especially at lower ranked results) in the "$4-2$" subrange. We did not see much difference between "$5-4$" and "$4-3$" gaps, which is largely different from the results on offline metrics from the other field dataset TianGong-QRef. This result suggest that there might be other hidden contextual factors that shape satisfaction sensitivity to the changes in search. In the ad hoc retrieval context, we observed significant changes from the "5-4" gap to the "4-3" gap. Users were more sensitive to the changes in `TimeFirstClick`, `TimeLastClick`, and `ClickDepth` when moving between "very satisfied" (5) level and "satisfied" (4) level. However, we observed the opposite direction of sensitivity changes in a few offline metrics (e.g. `DCG@3`, `INST3`), with users being more sensitive in the "$4-3$" gap.

In summary, our results from between-gap difference tests illustrate the uneven gaps in graded user satisfaction and also demonstrate the cross-metric divergence in satisfaction sensitivity. For most of the metrics under all datasets, we found that users were more sensitive to the "5-4" gap than to the "4-3" gap. However, there is less cross-dataset consistency in "4-3" and "3-2" sensitivity comparisons: in the two field study datasets, users generally showed high sensitivity to the "4-3" gap. However, for the session retrieval study conducted in controlled lab setting, there were not much significant differences between the two gap groups. In addition, we observed almost opposite directions of sensitivity changes from "3-2" to "2-1" between KDD19 and TianGong-QRef datasets, despite the fact that most of these sensitivity variations were associated with offline metrics in both datasets. This between-dataset discrepancy in the changes of satisfaction sensitivity indicates that the environment of search (i.e., controlled lab versus natural settings) and task source or motivation (i.e., assigned task versus authentic task) could significantly affect users' satisfaction thresholds and sensitivity to varying sizes of changes in search outcomes.

In the ad hoc retrieval study (Table 4), we did not observe significant variations in satisfaction sensitivity across "4-3", "3-2" and "2-1" groups, except for `TimeFirstClick`. This may be because in ad hoc retrieval contexts, users were less sensitive to the gains and efforts on less satisfactory result lists as they were separated from satisfactory query segments and thus generate less direct contrasts or threshold priming effects. In contrast, users' in-situ experiences in continuous sessions are constantly affected by their ongoing interactions and past experiences, which contribute to the cross-gap variations in sensitivity.

In addition to study settings, there are other possible factors that might lead to differences in the size and

Table 4. Between-Gap Differences: THU-2017

Measures	5-4	4-3	3-2	2-1	Kruskal–Wallis posthoc test
TimeFirstClick***	-2.6(13.9)	12.6(11.1)	-1.4(10.7)	1.8(10.1)	54<43***, 43>32**, 32<21*
DCG3**	4.1(12.1)	1.3(12.7)	2.2(12.3)	2.2(10.1)	54>43**, 43<32, 32<21
DCG5**	4.9(15.1)	1.8(15.3)	3.1(14.8)	2.7(11.7)	54>43***, 43<32, 32>21
ActionCount**	-112.4(219.4)	-28.1(251.5)	-28.7(261.1)	6.2(261.4)	54<43**, 43>32, 32<21
DCG10**	5.2(18.3)	2.3(17.4)	4.4(16.7)	3.7(13.3)	54>43***, 43<32, 32>21
ERR**	0.0(0.1)	0.0(0.1)	0.0(0.1)	0.0(0.1)	54<43**, 43>32, 32<21
QueryDwellTime**	-22.4(56.4)	-2.1(62.4)	-11.9(79.8)	5.7(93.7)	54<43**, 43>32, 32<21
TimeLastClick*	-18.6(40.9)	-2.2(47.7)	-12.1(61.0)	3.6(74.1)	54<43*, 43>32, 32<21
AveClickRank*	-1.2(2.9)	-0.4(3.1)	-0.1(3.1)	-0.2(3.0)	54<43**, 43<32, 32>21
ClickDepth*	-2.0(4.4)	-0.6(4.6)	-0.3(4.5)	-0.3(4.6)	54<43**, 43<32, 32<21
INST3*	0.25(0.83)	0.1(0.84)	0.19(0.82)	0.15(0.7)	54>43**, 43<32, 32>21
KeyDocCount*	0.85(3.99)	0.37(3.77)	0.69(3.42)	0.73(2.62)	54>43**, 43<32, 32<21

Note: *:p<.05, **:p<.01, ***:p<.001.

Table 5. Between-Gap Differences: TianGong-QRef

Measures	5-4	4-3	3-2	2-1	Kruskal–Wallis posthoc test
Precision@10***	0.0(0.1)	0.01(0.14)	0.08(0.13)	0.03(0.11)	54<43***, 43<32**, 32>21**
INST10***	0.0(0.32)	0.09(0.37)	0.17(0.32)	0.07(0.23)	54<43***, 43<32, 32>21***
INST5***	0.02(0.38)	0.11(0.43)	0.2(0.37)	0.08(0.26)	54<43***, 43<32, 32>21***
INST3***	0.05(0.46)	0.14(0.51)	0.25(0.44)	0.1(0.31)	54<43***, 43<32, 32>21***
ERR***	0.01(0.14)	0.04(0.16)	0.11(0.16)	0.05(0.12)	54<43***, 43<32, 32>21***
RBP***	0.03(0.3)	0.09(0.33)	0.16(0.29)	0.07(0.2)	54<43***, 43<32, 32>21***
DCG10***	0.91(4.22)	1.1(4.4)	1.7(3.7)	0.67(2.28)	54<43***, 43<32, 32>21***
DCG5***	0.97(4.14)	1.1(4.3)	1.6(3.6)	0.65(2.15)	54<43**, 43<32, 32>21***
DCG3***	0.98(4.02)	1.0(4.1)	1.5(3.3)	0.59(1.97)	54<43**, 43<32, 32>21***
nDCG10***	0.04(0.44)	0.06(0.48)	0.35(0.55)	0.19(0.55)	54<43, 43<32***, 32>21*
nDCG5***	0.05(0.45)	0.06(0.49)	0.34(0.56)	0.18(0.55)	54<43, 43<32***, 32>21
nERR***	0.01(0.41)	0.05(0.49)	0.43(0.63)	0.23(0.64)	54<43***, 43<32***, 32>21*

Note: *:p<.05, **:p<.01, ***:p<.001.

direction of satisfaction sensitivity variations across varying metrics, such as task topic, task complexity, and user intents (J. Liu, 2021). Although examining all possible contextual variables is beyond the scope of our study, we further explored some of the goal-related query-level factors in by-group analyses under RQ2.

In addition to highlighting the top-ranked metrics which best demonstrate large divergences between different gaps, we also introduce the overall distribution of metrics (i.e., number of metrics involving statistically significant difference and type of difference) for the "$5-4$" and "$4-3$" comparison as they represent the most frequent satisfaction grades and intervals. As it is shown in Figure 2, in terms of online process-oriented evaluation metrics, we observed more significant differences in the "$5-4 < 4-3$" group within controlled lab study settings (i.e., THU-17 and THU-KDD19) compared to the results in naturalistic settings (i.e., TianGong-QRef and TianGong-SS-FSD), indicating that users are more sensitive to relatively smaller changes in search efforts when moving between score 5 and score 4. This result also suggests that study setting may play an important role in affecting users' search satisfaction criteria and perceived search efforts. In controlled lab settings where search sessions tend to be short, participants could be more sensitive to the changes in search efforts reflected in online metrics.

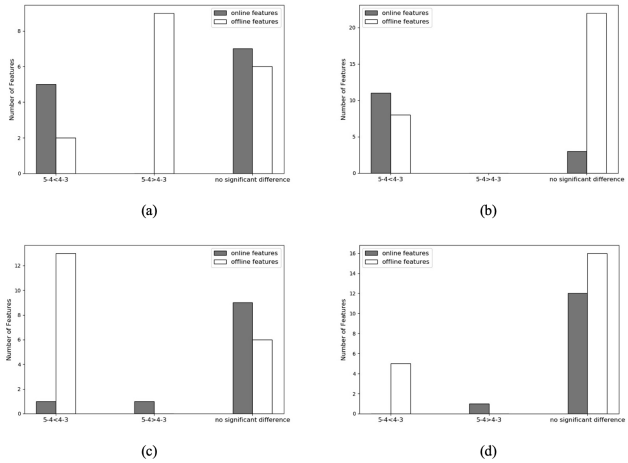


Figure 2. Differences between "$5-4$" and "$4-3$" groups: Metric distribution
(a): THU-17. (b): THU-KDD19. (c): TianGong-QRef. (d): TianGong-SS-FSD. Light bar: offline evaluation metrics; Dark bar: online evaluation metrics.

4.1.2. "$5-4$" and "$4-3$" Comparison: Metrics Distribution With respect to offline metrics, in ad hoc retrieval context (i.e., THU-17), a majority of measures fall into the $5-4 > 4-3$ category, indicating that users are more sensitive to the changes in multiple aspects of search outcomes when moving between score 4 and score 3. This pattern is different from that of online evaluation metrics (where users' search satisfaction sensitivity is higher in the "5-4" interval), showing that satisfaction sensitivity varies in different directions across different types of evaluation metrics in certain search settings. In contrast, under naturalistic settings, most offline metrics are included under the "$5-4 <$

Table 6. Between-Gap Differences: TianGong-SS-FSD

Measures	5-4	4-3	3-2	2-1	Kruskal–Wallis posthoc test
AvgClickRank***	-0.1(0.6)	-0.1(0.7)	0.12(0.71)	0.01(0.64)	54<43*, 43<32***, 32>21
ClickDepth***	-0.9(5.7)	-0.8(7.1)	1.3(7.2)	0.1(6.55)	54<43, 43<32***, 32>21
ClickCount***	-0.1(1.3)	-0.1(1.5)	0.47(1.46)	0.05(1.38)	54<43, 43<32***, 32>21
nERR**	0.0(0.32)	0.02(0.4)	0.2(0.56)	0.05(0.68)	54<43*, 43<32, 32>21
Clicks@3**	0.08(1.66)	-0.1(2.4)	-0.2(3.3)	0.34(3.57)	54>43, 43>32**, 32<21
ERR**	0.0(0.1)	0.01(0.18)	0.08(0.22)	0.02(0.27)	54<43*, 43<32*, 32>21
Precision@10**	0.0(0.4)	0.07(0.46)	0.19(0.5)	0.08(0.5)	54<43*, 43<32*, 32>21
RelDocCount*	0.2(4.2)	0.68(4.57)	1.8(5.0)	0.86(4.99)	54<43*, 43<32, 32>21
ActionCount*	-9.3(64.0)	-11.6(76.5)	5.8(72.7)	-3.3(72.7)	54>43, 43<32**, 32>21
Precision@3*	0.0(0.1)	0.02(0.14)	0.06(0.16)	0.01(0.17)	54<43*, 43<32, 32>21

Note:*: p<.05, **:p<.01, ***:p<.001.

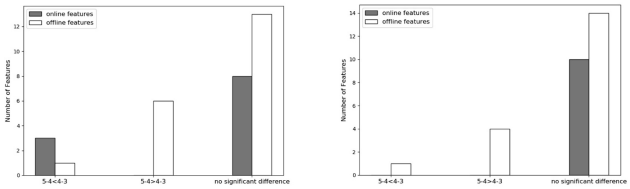
$4-3$" category. This result to some extent confirms the potential impacts of search/study settings.

It is worth noting that the significantly uneven sensitivity gaps only appear in some of the evaluation metrics. This result also illustrates the heterogeneity in user evaluation: Users are not equally sensitive to all evaluation metrics; Instead, there are non-significant differences, differences between online and offline metrics, as well as differences in the direction of changes in satisfaction sensitivity between different intervals across varying study settings. Examining this heterogeneity and further explore the reasons behind the systematic differences discussed above is essential for better understanding users' evaluation decisions.

4.2. RQ2: By-Group Analysis

As the first part of the answer to RQ2, we discusses overall effects of search type from between-gap difference tests. Comparing the results from THU-17 and that of other three datasets allowed us to further understand the differences in satisfaction sensitivity between ad hoc retrieval and whole-session retrieval contexts. In summary, the results on between-gap differences from THU-17 are similar to that of the session retrieval datasets in the following two aspects: 1) significant differences between "5-4" and "4-3" gaps were identified, indicating that sizable variations in satisfaction sensitivity may happen on multiple metrics in the "5-3" grade range; 2) user satisfaction grades were sensitive to a variety of offline evaluation metrics (the specific metrics differ across different datasets). However, we also identified differences across search types: Unlike the results from experiments on THU-KDD19 and TianGong-QRef, in the ad hoc retrieval context, user satisfaction sensitivity was also sensitive to a series online behavioral metrics. Compared to the online metrics listed in Table 6 (TianGong-SS-FSD), online measures identified in the ad hoc retrieval context were more diverse and covered both click-based (e.g. `AveClickRank`, `ClickDepth`) and dwell-time-based metrics (e.g. `TimeFirstClick`, `QueryDwellTime`). This result indicates that users may be more sensitive to search efforts when conducting single-query retrieval.

Dividing search dataset by cognitive level allowed us to examine how the between-gap differences in satisfaction sensitivity differ across query intents with varying levels of cognitive complexity. Due to the space limit, this section summarizes the metric distribution results in "$5-4$" and "$4-3$" comparison as this subrange frequently occur and best demonstrates the by-group differences. Full results including sensitivity variations on all metrics will be made available upon publication.



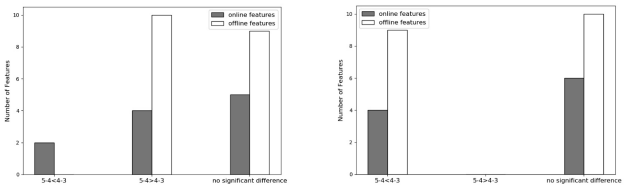
(a): Cognitive Level: Remember (b): Cognitive Level: Understand

Light bar: offline evaluation metrics; Dark bar: online evaluation metrics.

Figure 3. Results by Cognitive Level: THU17

With respect to the cognitive level of search in ad hoc retrieval, Figure 3 indicates that there were significant changes in user satisfaction sensitivity associated with search effort variations in searches focusing on retrieving and remembering facts. In contrast, we did not observe significant changes in online features for searches that require understanding information. In both groups, we found significant sensitivity changes associated with offline features, especially under the "$5-4 < 4-3$" condition. This result shows a pattern that differs from that of the between-gap analysis in other subranges, indicating that the change of sensitivity happens within different *subrange* for different metrics, especially in the *Understand* searches. For offline features, changes in satisfaction sensitivity mostly happened between "$5-4$" and "$4-3$" intervals.

Regarding search goals, The results presented in Figure 4 and Figure 5 illustrate the major divergence in satisfaction sensitivity patterns between controlled lab ad hoc retrieval study and naturalistic whole-session retrieval study, confirming the impacts of study setting

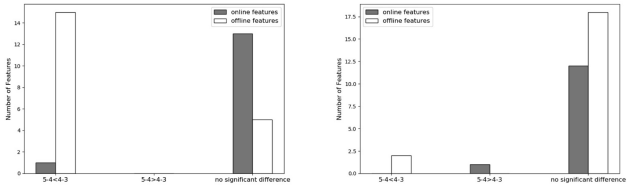


(a): Informational (b): Navigational + Transactional

Light bar: offline evaluation metrics; Dark bar: online evaluation metrics.

Figure 4. Results by Search Goal: THU17

on user evaluation. Specifically, in THU17 study, the results demonstrate that the changes in satisfaction sensitivity were associated with the variations in both online and offline metrics, whereas in naturalistic session searches, offline metrics play a major role in affecting satisfaction sensitivity. Overall, we observed more significant changes in informational searches compared to navigational and transactional searches. In particular, for navigational and transactional searches conducted in naturalistic settings, user satisfaction sensitivity remained unchanged over most of the evaluation metrics. When searches are restricted to predefined search tasks and limited search time (i.e., in controlled lab settings), users became more sensitive to the variations in experienced search efforts and outcomes, resulting in more uneven search-related gaps between different grades in the satisfaction scale.



(a): Informational (b): Navigational + Transactional

Light bar: offline evaluation metrics; Dark bar: online evaluation metrics.

Figure 5. Results by Search Goal: TianGong-SS-FSD

This section presents the result of a *meta-analysis* on the statistical significance of satisfaction sensitivity changes under different conditions in "$3-5$" subrange. Overall, the results demonstrate that the changes of satisfaction sensitivity and evaluation criteria were not only affected by *global* environment, such as study settings and type of search (i.e., ad hoc retrieval and whole-session retrieval), but also shaped by users' *local* query-level intents and goals. Moreover, these contextual effects are multidimensional and include the direction of sensitivity changes (increase or decrease from "$5-4$" interval to "$4-3$" interval), type of evaluation metrics involved, and number of metrics where significant changes happened. Understanding this heterogeneity can enhance our understanding of the variations in *in-situ* satisfaction feedback.

5. Discussion and Conclusion

Our study examined the extent to which user satisfaction sensitivity varies across different gaps under varying *usefulness-based* and *topical-relevance-based* metrics, search interaction features, and search goals and contexts. Our experiments on four diverse user study datasets reject the Null Hypothesis from the Introduction section and demonstrate that: 1) there are statistically significant *between-gap divergences* in user satisfaction sensitivity, especially to offline evaluation metrics, in both ad hoc and session-based retrieval scenarios; 2) the size and direction of between-gap variations in evaluation measures vary across different search goals, cognitive levels, and study settings, suggesting that satisfaction sensitivity is dynamic and context-dependent. Since there are limited overlaps in terms of the metrics that reflect cross-gap sensitivity variations across datasets and groups, to better understand the cross-metric and cross-context differences, we need to explore a broader range of judgment dimensions (credibility, readability, etc.).

Note that we do not object to using graded user satisfaction or other similar self-reported measures in evaluating IR and other computing systems. In fact, we believe that these grading-scale-based measures are useful in collecting information about users' *in-situ* and retrospective satisfaction evaluations under certain settings. However, we argue that it is important to fully understand the satisfaction measure we use as the ground truth and be aware of the empirically confirmed uneven gaps between grades and the potential risks of ignoring them when making certain statistical assumptions. Due to the scope limit of our analysis, we could only cover one ground truth measure, *query-level user satisfaction*. In fact, the proposed concerns and analytical approach are also highly relevant to other self-reported measures, such as perceived task difficulty, topic familiarity, cognitive loads, and user engagements.

Our study echoes the argument in Ferrante et al. (2017) for calibrating IR evaluation measures at the query-resultset level to narrow the uneven gap regarding user perception. Findings from this study can 1) enhance our understanding of the uneven gaps and heterogeneity in satisfaction evaluation across different local contexts (e.g. search intents and goals) and global settings (e.g. tasks); 2) support the modeling of user preferences and the development of unbiased metrics that are on an interval scale and consistent with users' perceptions; and 3) contribute to future studies on calibrating IR evaluation measures as well as developing

robust evaluation infrastructures for IIR.

6. Acknowledgement

This work is partially supported by the Faculty Investment Program (FIP) from the Research Council of the University of Oklahoma Norman Campus.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. London, UK: Longman.

Ayuningtyas, K., & Janah, N. Z. (2018). Development and UI/UX usability analysis of pinjemobil web-based application using user satisfaction model. *2018 International Conference on Applied Engineering (ICAE)*, 1–6.

Chen, J., Mao, J., Liu, Y., Zhang, F., Zhang, M., & Ma, S. (2021). Towards a better understanding of query reformulation behavior in web search. *Proceedings of The Web Conference 2021*.

Chen, Y., Zhou, K., Liu, Y., Zhang, M., & Ma, S. (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*, 15–24.

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. *Proceedings of the Third Workshop on Human-Computer Interaction and Information Retrieval Cambridge*, 1–4.

Ferrante, M., Ferro, N., & Fuhr, N. (2021). Towards meaningful statements in IR evaluation: Mapping evaluation measures to interval scales. *IEEE Access*, *9*, 136182–136216.

Ferrante, M., Ferro, N., & Pontarollo, S. (2017). Are IR evaluation measures on an interval scale? *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 67–74.

Hofmann, K., Li, L., & Radlinski, F. (2016). Online evaluation for information retrieval. *Foundations and trends in information retrieval*, *10*(1), 1–117.

Kelly, D. (2009). *Methods for evaluating interactive information retrieval systems with users*. Now Publishers Inc.

Koolen, M., Kamps, J., Bogers, T., Belkin, N., Kelly, D., & Yilmaz, E. (2017). Report on the second workshop on supporting complex search tasks. *ACM SIGIR Forum*, *51*(1), 58–66.

Liu, J. (2021). Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management*, *58*(3), 102522.

Liu, J., & Han, F. (2020). Investigating reference dependence effects on user search interaction and satisfaction: A behavioral economics perspective. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1141–1150.

Liu, J., & Shah, C. (2019). Interactive IR user study design, evaluation, and reporting. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *11*(2), i–93.

Liu, M., Liu, Y., Mao, J., Luo, C., Zhang, M., & Ma, S. (2018). Satisfaction with failure or unsatisfied success: Investigating the relationship between search success and user satisfaction. *Proceedings of the 2018 World Wide Web Conference*, 1533–1542.

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

Liu, Z., Zhou, K., & Wilson, M. L. (2021). Meta-evaluation of conversational search evaluation metrics. *ACM Transactions on Information Systems (TOIS)*, *39*(4), 1–42.

Xie, X., Mao, J., Liu, Y., de Rijke, M., Shao, Y., Ye, Z., Zhang, M., & Ma, S. (2019). Grid-based evaluation metrics for web image search. *The World Wide Web Conference*, 2103–2114.

Zhang, F., Mao, J., Liu, Y., Ma, W., Zhang, M., & Ma, S. (2020). Cascade or recency: Constructing better evaluation metrics for session search. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 389–398.

Zhang, F., Mao, J., Liu, Y., Xie, X., Ma, W., Zhang, M., & Ma, S. (2020). Models versus satisfaction: Towards a better understanding of evaluation metrics. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 379–388.