Users Meet Clarifying Questions: Toward a Better Understanding of User Interactions for Search Clarification

JIE ZOU, MOHAMMAD ALIANNEJADI, and EVANGELOS KANOULAS,
University of Amsterdam, The Netherlands
MARIA SOLEDAD PERA, Boise State University, USA
YIQUN LIU, Tsinghua University, China

The use of clarifying questions (CQs) is a fairly new and useful technique to aid systems in recognizing the intent, context, and preferences behind user queries. Yet, understanding the extent of the effect of CQs on user behavior and the ability to identify relevant information remains relatively unexplored. In this work, we conduct a large user study to understand the interaction of users with CQs in various quality categories, and the effect of CQ quality on user search performance in terms of finding relevant information, search behavior, and user satisfaction. Analysis of implicit interaction data and explicit user feedback demonstrates that high-quality CQs improve user performance and satisfaction. By contrast, low- and mid-quality CQs are harmful, and thus allowing the users to complete their tasks without CQ support may be preferred in this case. We also observe that user engagement, and therefore the need for CQ support, is affected by several factors, such as search result quality or perceived task difficulty. The findings of this study can help researchers and system designers realize why, when, and how users interact with CQs, leading to a better understanding and design of search clarification systems.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *User studies*; • **Information systems** → *Users and interactive retrieval;*

Additional Key Words and Phrases: User study, information seeking systems, clarifying questions

ACM Reference format:
Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. 2023. Users Meet Clarifying Questions: Toward a Better Understanding of User Interactions for Search Clarification. *ACM Trans. Inf. Syst.* 41, 1, Article 16 (January 2023), 25 pages.
https://doi.org/10.1145/3524110

1 INTRODUCTION

In a typical search scenario, users formulate queries that describe their information needs and pose them to a search engine [6]. However, search queries are occasionally short, ambiguous, or

This research was supported by the NWO Smart Culture—Big Data/Digital Humanities (314-99-301), the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the H2020-EU.3.4.—SOCIETAL CHALLENGES—Smart, Green And Integrated Transport (814961). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Authors' addresses: J. Zou, M. Aliannejadi, and E. Kanoulas, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands; emails: {j.zou, m.aliannejadi, e.kanoulas}@uva.nl; M. S. Pera, Boise State University, 1910 University Dr., Boise, ID 83725, United States; email: solepera@boisestate.edu; Y. Liu, Tsinghua University, 30 Shuangqing Rd, Haidian District, Beijing City, China; email: yiqunliu@tsinghua.edu.cn.

1046-8188/2023/01-ART16 $15.00
https://doi.org/10.1145/3524110

incomplete, and thus they may be misinterpreted by search systems [6, 73]. Therefore, researchers have augmented search functionality by allowing search engines to ask ***clarifying questions*** (**CQs**), as a step toward a better understanding of users' information needs [6, 73], context, and preferences [50].

Asking CQs has attracted considerable attention within the information retrieval community because of the popularity of conversational information-seeking systems. Even though designing systems capable of having mixed-initiative interactions with users has been a long-standing goal [12, 24, 51], only recently have notable developments and achievements been observed in this area [4–6, 31, 66, 73, 79–81, 83]. Recent studies have demonstrated the significance and applicability of CQs for broad use cases, such as product search [76, 80], preference elicitation for recommendation [57, 79], and information-seeking conversations [6, 31, 44], and web search [74]. These studies highlight the effectiveness of CQs for system performance; however, the impact of asking CQs on users is to a big extent unknown. The findings of previous studies indicate that users enjoy voice query clarification even though it delays system responses [39]. Furthermore, CQ templates, candidate answer attributes, and query properties affect the user engagement rate [75]. However, the effect of CQ quality on user search performance, and the effect of user perception on search clarification remain unstudied. For example, displaying a clarification pane at the top of the **search engine results pages** (**SERPs**) or interrupting users in a conversation bears an unknown effect of cost and benefit on users. It can be beneficial to guide users through their search by asking one or multiple CQs, but low-quality CQs may come with a high risk of frustrating users.

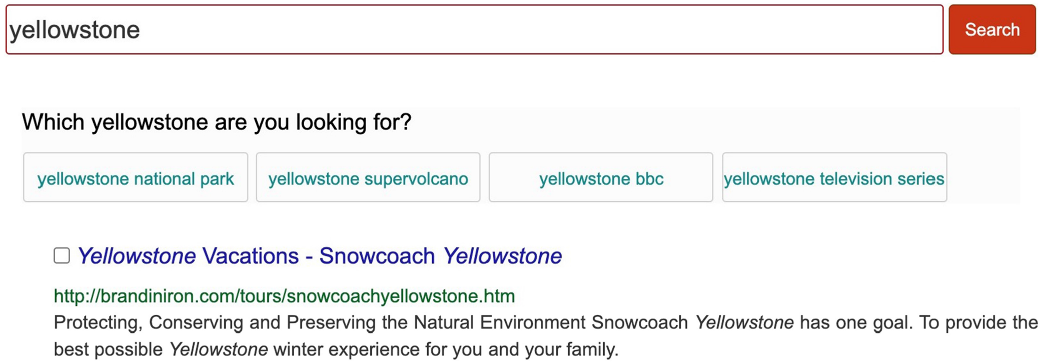
Here, we investigate (a) the effect of asking different-quality CQs on user search behavior, user ability to find relevant information, and user satisfaction, (b) the factors, pertaining to user background and perception intrinsic to the web search tasks, that prompt users to engage with CQs, and (c) the circumstances that lead to a high engagement level with CQs. To this end, we conducted a user study involving 106 participants who were asked to complete a set of web search tasks, following an existing laboratory user study setup [23, 30]. In particular, we simulated various conditions that a user and a system would encounter, and we studied the effect of system decisions on user behavior, as well as the effect of user decisions on system effectiveness. By design, the tasks spanned various topics, levels of difficulty, and CQ quality categories. We separated the participants into two groups: one group completed tasks using a plain search interface, and the other group using a search clarification interface designed to resemble Bing's[1] clarification pane [73]. As shown in the sample search interface in Figure 1, the clarification pane consists of a CQ in addition to the corresponding suggested answers that are displayed below the query input. Analysis of collected implicit and explicit user data allows us to examine user behavior and satisfaction within and across the two groups.

In this study, we answer three research questions:

> **RQ1:** To what extent does asking CQs affect user search behavior and satisfaction? Are users affected by being asked high-quality vs. low-quality CQs in a search session?

To address **RQ1**, we present users with CQs in different quality categories and compare behavioral measures capturing interaction and performance, such as querying, mouse movement, and bookmarking. We also investigate how much engaging with different quality categories of CQs affects user satisfaction. Moreover, we hypothesize that the effect of CQs on user performance spans the next SERP and roots in the entire search session. Accordingly, we analyze the effect on users not only immediately after the interaction with CQs (query-level) but also in the entire session

[1] http://www.bing.com.



Which yellowstone are you looking for?

□ *Yellowstone* Vacations - Snowcoach *Yellowstone*

http://brandiniron.com/tours/snowcoachyellowstone.htm

Protecting, Conserving and Preserving the Natural Environment Snowcoach *Yellowstone* has one goal. To provide the best possible *Yellowstone* winter experience for you and your family.

Fig. 1. Search engine UI of our user study system.

as they strive to complete a web search task (session-level). To deepen our understanding of the interaction with CQs, we also seek to answer:

> **RQ2:** How much do user background and task perception affect the interactions with CQs?

To address **RQ2**, we analyze responses to pre/post-task questionnaires and user demographic information. With **RQ1**, we examine how user behavior is affected after engaging with the CQs; here, we investigate the extent to which user task perception, such as expected difficulty and prior knowledge, affects the willingness of users to interact with CQs. Finally, we consider the user engagement with CQs under different circumstances:

> **RQ3:** How do users interact with CQs under various circumstances?

To address **RQ3**, we calculate user engagement metrics regarding the CQ pane (e.g., click-through rate and cursor hovering) and study the effect of factors such as CQ quality categories, task types, and SERP quality on the user engagement with CQs. As generally, users answer CQs (in the form of clicking on an answer) [74, 82], it is also critical to study how other engagement metrics, such as mouse movement, differ among CQs of different quality and types.

The results lead to the following conclusions: (a) when users engage with high-quality CQ panes, the interaction, performance, and satisfaction improve, compared with those on search engines that do not offer such an option. However, when the CQs are of low- or mid-quality, they actually negatively affect all measures, even if they are presented to the user, and the user does not engage with them; (b) the user expected and perceived difficulty of a web search task influences the degree of their engagement with CQs, while less experienced users incorrectly using the CQs and clicking more on irrelevant answers; and (c) the degree and quality of user engagement with CQ panes are affected by factors such as SERP quality, SERP diversity, and screen size, and they reduce as a search session evolves. As asking CQs is a necessary step toward developing mixed-initiative conversational search systems [51, 74], we believe that our findings can prove helpful in this direction.

2 RELATED WORK

Asking CQs has shown great potential in enhancing the functionality of several applications, such as search [53, 76, 77, 80], recommender systems [57, 79], information-seeking conversations [6, 31, 45, 66, 73], and dialogue systems [22, 60, 72]. Four decades ago, Belkin et al. [13] explored early mixed-initiative systems by offering users choices in a search session and discussed the significance of mixed-initiative systems. Recently, Zamani et al. [73] proposed a neural approach for generating CQs. Hashemi et al. [31] used CQs to enrich representation learning in information-seeking conversations. Sekulic et al. [54] modeled search clarification prediction as a user engagement problem and proposed a transformer-based method to predict user engagement. Radlinski and Craswell [51] highlighted the importance of CQs for conversational search and recommender systems. Zhang et al. [76] presented a unified approach for conversational search and recommendation by asking questions over item "aspects" extracted from user reviews. Instead of item "aspects," Zou et al. [79, 80] constructed CQs based on extracted informative terms for recommendation and product search, respectively. Asking CQs about different item attributes is also applied to improve conversational recommender systems and dialog systems [4, 78]. Given that asking CQs is a prominent area of study, we have recently seen an influx of datasets and challenges facilitating research in this area. Notable examples include the Qulac dataset [6], the MIMICS dataset [74], and the Conversational AI challenge [4]. These datasets and challenges enable system training and evaluation on CQ-related tasks [55]. Existing studies on CQs primarily focus on model and representation learning as well as dataset construction. By contrast, we study the underlying mechanism of user interactions with search systems using CQs, offering insights into the design of these models.

Research discussing empirical studies examining CQs is broad, from the use of CQs on community question answering sites such as Stack Exchange, where answerers ask CQs to askers to better comprehend information requests [17], to the challenges of CQs for entity disambiguation [19]. Vtyurina et al. [65] compared three different conversational search systems: humans, assistants, and wizards; Kiesel et al. [39] studied the effect of query clarification over voice on user satisfaction, and they demonstrated that language proficiency affects user satisfaction. Trippas et al. [63] studied the effect of voice query clarification on user interaction, demonstrating that user queries and the average time on task become longer as task complexity increases.

More recently, Krasakis et al. [44] analyzed the effect of CQs on document ranking. Zou et al. [82] empirically quantified and validated user willingness and the extent of providing correct answers to CQs in existing question-based systems. Unlike that study, which primarily validates certain assumptions regarding existing CQ-based models, the present study is concerned with user behavior and engagement with CQs in various quality categories for search clarification.

Zamani et al. [73] conducted a user study showing that asking CQs is, in principle, beneficial. They constructed a taxonomy of clarifications for open-domain search queries to develop CQ templates. More notably, Zamani et al. [73] articulated the differences between search clarification and query reformulation, suggestion, or auto-completion. Although the candidate answers for CQs are similar to query suggestions, they found that CQs are substantially beneficial for user engagement in terms of the click-through rate (the reader is referred to Zamani et al. [73] for a more detailed comparison between search clarification and query reformulation, suggestion, or auto-completion). Based on their previous work, Zamani et al. [75] conducted a large-scale in-situ study, analyzed clarification panes for millions of queries, and developed representation learning methods to re-rank clarification panes. In particular, they analyzed the click rate received on CQ panes as a function of search query properties (e.g., query length), question template types (based on their template taxonomy [73]), and answer attributes. In addition, they examined the effect of clarification on proxies of user dissatisfaction. Our work is complementary to the work by Zamani et al. [75]. We conduct a smaller scale but controlled laboratory user study. This allows us to control certain variables (e.g., the quality of the CQ panes or the relevance of the results), collect explicit user information (e.g., via questionnaires), and user feedback (e.g., bookmarks). Furthermore, we focus our analysis on user search performance, behavior, and satisfaction at the query- and session-level, as well as the need for user engagement with CQs.

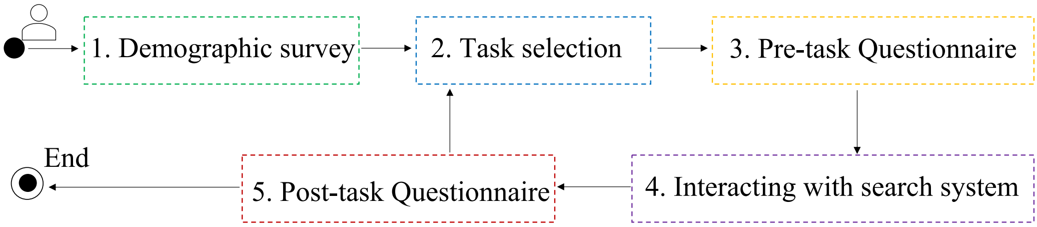

1. Demographic survey
2. Task selection
3. Pre-task Questionnaire
4. Interacting with search system
5. Post-task Questionnaire
End

Fig. 2. Study protocol.

3 METHOD

In this section, we present the study protocol, design, and participants.

3.1 Protocol

To answer the aforementioned research questions, we conducted a user study to capture implicit user behavior and explicit user feedback under various conditions. Following the protocol shown in Figure 2, we asked the study participants to perform the following:

(1) Complete a demographic survey.
(2) Select an experimental web search task from a list that was presented to them.
(3) Answer a pre-task questionnaire regarding their perception and opinions about the selected experimental task.
(4) Submit a search query, find relevant information, and bookmark it. During their session, they could choose to answer a CQ (if shown to them). However, it was not a requirement to answer or engage with CQs.
(5) Click on the "Finish" button once they are confident that they have found relevant information.
(6) Answer a post-task questionnaire regarding their perception and experience.
(7) Consider the selection of another task to complete. In this case, they return to step (2); otherwise, they finish the study participation.

We offered the participants detailed instructions regarding the study and clarified its goal. We demonstrated how our augmented search interface works in a short video and stressed that CQs are intended to support their experimental web search tasks but they are not always related and useful; we urged the participants to proceed with the experimental web search tasks as they normally would and take advantage of CQs only when deemed necessary.

We did not collect any data from the participants that can be used to breach their privacy. Furthermore, our study was approved by the ethics committee of the institute, and we specified to the participants that their data were securely encrypted and stored and that they could opt out at any point in the study.

3.2 Study Design

Herein, we provide insights into the design decisions that constitute the cornerstone of our study.

3.2.1 Search Interface. Our search interface is designed to mimic the commercial search system Bing [73], including the embedded CQs pane at the top of the search page—the search UI is shown in Figure 1.[2] In the text box on the search page, users enter a query related to the selected

[2] We allowed participants to report system interface issues; none was reported.

Table 1. Experimental Task Description

Task ID	Web Track ID	Task Categories
T1	133 (1)	fact-finding & faceted
T2	197 (1)	fact-finding & faceted
T3	52 (6)	information gathering & ambiguous
T4	60 (1)	information gathering & ambiguous
T5	200 (3)	information gathering & faceted

Web Track ID represents the original topic ID followed by the subtopic ID in parenthesis in TREC Web Track. Each experimental task corresponds to a task topic.

experimental task and wait for the results together with CQs (if any) to be presented. Users can (a) browse the results, (b) reformulate their query and repeat the search as many times as they wish, (c) answer the CQs (if any), and (d) click on the results that they find interesting and would like to know more about, and bookmark the ones they think are relevant for the experimental task.

3.2.2 Search Engine. The search engine used to produce SERPs is ChatNoir [15, 49], a web search engine widely used in previous research efforts focused on studying searchers' interactions and use of search engines [26, 27, 48, 59, 64], indexing the entire ClueWeb09 corpus.[3] ChatNoir search engine is fast and has been proven to generate comparable results to other search engines and BERT rankers [15, 18, 48]. It retrieves web pages from the ClueWeb09 Corpus by utilizing the field-oriented retrieval model BM25F and incorporates PageRank and SpamRank scores.

Note that in producing SERPs We removed duplicate results retrieved as well as spam web pages.[4]

3.2.3 Experimental Tasks. Different experimental web search tasks may influence information-seeking behavior [68]. Hence, we designed five such tasks derived from the **Text Retrieval Conference** (**TREC**) Web Track 2009–2012.[5] To clarify the experimental task the participants are expected to complete, we expanded the experimental task description using Simulated Work Task Situations [16], which creates a task scenario that offers participants a search context and a basis for relevance judgments. A sample task description is as follows:

> Imagine you are flying next week from the Ontario airport, located in California. Since this is your first time flying out from this airport, you are thinking of gathering some information about its facilities and services.

We categorized each experimental task according to its type: fact-finding or information gathering [25]. The former are simple tasks in which specific facts, files, or pieces of information are sought; the latter involves collecting information often from various sources to make a decision, write a report, or complete a project. The experimental tasks were also categorized based on the types defined by TREC: faceted and ambiguous (see Table 1).

Before the experiment, all experimental tasks were pilot-tested until no issues were reported. To motivate the search session,[6] of the participants were guided to read through all the task

[3] https://www.chatnoir.eu.

[4] Spam filtering was performed by applying Waterloo Spam Ranking for the ClueWeb09 Dataset, which was typically applied for TREC Web Track 2009 submissions.

[5] https://trec.nist.gov/data/webmain.html.

[6] A search session is an entire session for a user completing an experimental task.

Table 2. CQs Taxonomy and Examples

Taxonomy	Description	Examples of questions & answers
C1	Off-topic, unrelated CQs	Q: What do you want to know about the Idaho state flag? A: 1. Year adopted; 2. Pictures; 3. History; 4. Designer; 5. Colors.
C2	Related but not useful CQs	Q: What do you want to know about the Idaho State flower? A: 1. Growing seasons; 2. Growing conditions; 3. History; 4. Color; 5. Year adopted.
C3(i)	Related and useful CQs for specific/faceted details	Q: What do you want to know about the Idaho State flower? A: 1. Growing seasons; 2. Growing conditions; 3. history; 4. Color; 5. Scientific name.
C3(ii)	Related and useful CQs for disambiguation	Q: Which AVP are you interested in? A: 1. AVP program; 2. AVP company; 3. AVP association; 4. AVP airport; 5. AVP movie.

C1, C2, and C3(i) are from the experimental task "scientific name of Idaho State flower"; C3(ii) is from the experimental task "movie named AVP."

descriptions and select a task with which they felt most comfortable, thus avoiding any task assignment biases [33]. We encouraged the participants to complete as many experimental tasks as they could among the ones provided. After finishing a task, they could select another until they were no longer interested. Post-hoc analysis of the distribution of selected experimental tasks during the entire study indicates no obvious preference for any experimental task in either category,[7] and most participants (91.2%) stated that the experimental tasks were very clear.

3.2.4 CQs and Candidate Answers. To study search clarifications, we developed a pool of CQs and candidate answers. We first constructed a CQ taxonomy capturing different quality categories based on the relatedness and usefulness of the CQs in the search process. Subsequently, two expert annotators generated and reviewed CQs as well as candidate answers for each experimental task following the proposed taxonomy. In case of disagreement, they would discuss and agree on a better formulation. Note that both expert annotators were trained so they could become familiar with the system, the CQ taxonomy, and the CQ generation pipeline, thus ensuring they understood their role. To inform the taxonomy design, we conducted a survey to ask users about factors that would lead them to interact with CQs. Based on 200 collected responses, most users indicated "related question asked" (33.5%) and "useful question asked" (21%); the latter aligns with the usefulness metric assessing the follow-up question suggestion in web search by Rosset et al. [53]. To refine the taxonomy, we also looked at existing literature on CQs [53, 73, 75] and public CQ datasets. The question taxonomy comprises three main categories: **(C1)** off-topic, unrelated CQs, **(C2)** related but not useful CQs, e.g., a duplicate question with a user query or a related question without useful answers, and **(C3)** related and useful CQs. To facilitate the development of CQs, we further defined two subcategories of C3 according to two fundamental purposes: **C3(i)** related and useful for specific/faceted details, for example, a question asking for a faceted attribute; **C3(ii)** related and useful for disambiguation, for example, disambiguating the query "apple" by asking whether

[7] Task T1 was selected 74 times, task T2 was selected 57 times, task T3 was selected 62 times, task T4 was selected 64 times, and task T5 was selected 73 times.

it is about the fruit or the brand. In general, category C3 refers to high-quality CQs that are substantially beneficial to the user. For example, they could be a means of providing new information, the next step to complete a task or exploratory options regarding a task.

The general principle in creating this taxonomy is to cover a variety of CQ quality categories and investigate the potential mechanism of search clarification under these quality categories. For example, CQs that are off-topic or useless may elicit user dissatisfaction and cause users to leave the session [66], whereas related and useful CQs can aid users [82]. The taxonomy and sample CQs are presented in Table 2.

In this study, for each experimental task, we generated three CQs with their respective candidate answers, each corresponding to a CQ category in our taxonomy. For CQs in C3, CQs in C3(i) were generated for experimental tasks in the faceted category, whereas CQs in C3(ii) were generated for experimental tasks in the ambiguous category. Following the Bing setting [75], each CQ had at most five answers, with each answer corresponding to a reformulated query for the next turn. A user can answer multiple CQs in one search session. However, given that the CQs are constructed before the search, the choice of CQs shown to the user does not depend on the current user query. This is why certain interactive and feedback effects may be missing in this step. We leave the investigation of user behavior and engagement in a multi-turn setting for future research. In this work, we generate a single CQ for each category in each experimental task, to ensure different users meet the same CQ under the same condition. However, it might be beneficial to extend the analysis by composing multiple CQs for each category for a given experimental task in the future.

3.2.5 Interactive Search Flow. Once a user agreed to participate in the study she was assigned to one of two groups: (a) the control group which completes the experimental task with a plain interface, or (b) the treatment group which is exposed to an interface with CQ panes, similar to the one in Figure 1. Once the user selected a task to complete, she was able to submit their query to our search engine. If the user was assigned to the treatment group, a CQ pane was shown along with the results. The CQs and corresponding answers for each experimental task were selected from a manually constructed CQ pool (Section 3.2.4). The CQ to be shown to a user was randomly selected from three CQs manually developed for each experimental task, following the related literature [21, 30, 36, 38] assigning conditions or subjects randomly. Each time a user chose to click on a CQ answer, the answer was concatenated to the user's query and resubmitted to the search engine, following the Bing setting [75]. Once a new query was issued, or a CQ was answered, a new CQ was selected to be displayed to the user. To mimic a real-world scenario in which a system would not ask the same question if it has already been answered by the user, the CQs clicked by a user were not shown again to the user in the same search session. Therefore a CQ from the other two CQ quality categories would be chosen. To avoid position bias, the answers to a CQ also appeared in random order on the CQ pane each time. To summarize, the presence or absence of CQs is a between-subjects variable, whereas the CQ category is a within-subjects variable. Specifically, approximately 25% of search sessions did not show CQs, and 25% * three (categories) of search sessions showed CQs.

3.2.6 Questionnaires. We provided participants with a set of questionnaires to obtain explicit feedback. We first presented participants with *demographic questions* eliciting information pertaining to their gender, age, career field, English language proficiency, and educational background. The purpose of these questions was to better understand users and determine whether their background would influence their interaction with CQs. Moreover, before and after completing each experimental task, we asked the participants to fill out short questionnaires. From the *pre-task questionnaire*, we collected the perception and opinions of the participants regarding the selected experimental task, including their prior knowledge, expected task difficulty, perceived task clarity, task interest, distraction level, and search expertise. The options for search expertise were "search daily," "search weekly," "search monthly," and "never search." For the remaining inquiries, users selected their answers on a scale of 1–5 . From the *post-task questionnaire*, we gathered information related to the experience of the participants with the system, including perceived helpfulness, attitudes toward future use of CQ-based systems, overall satisfaction rating, perceived task relevance, perceived task difficulty, and domain knowledge regarding the completed experimental task. The options for perceived helpfulness and attitudes toward future use of CQ-based systems were "positive," "negative," and "neutral." The overall satisfaction rating, perceived task relevance, perceived task difficulty, and domain knowledge were scored on a scale from 1 to 5. From responses to the aforementioned questionnaires, we collected user opinions about the experimental tasks and the system, allowing the investigation of the relationship between user interactions with CQs and user background, task perception, and user experience.

3.3 Participants

The participants in the user study were 106 volunteers recruited through email invitations (students and staff of two universities, one in Europe and one in the U.S.). Some of their personal data varied are as follows:

- Gender: 39 females, 65 males, two non-binary.
- Age: 69 participants were 18–24 years old, 26 were 25–34, seven were 35–44, and four were older than 44.
- Career field: 86 in science, computers and technology, three in education and social services, three in health care, three in law and law enforcement, two in management, business and finance, and one in architecture and civil engineering; eight did not specify.
- English language proficiency: 22 native and 51 proficient; the remaining were beginners.
- Highest education level completed: 58 high schools, 21 bachelor's, 11 master's, and four doctorate; 12 did not specify.

4 RESULTS

In this section, we present an analysis of the data collected through the user study. The data statistics are presented in Table 3. Note that the number of CQ showing times in C1, C2, or C3 in Table 3 are not balanced because of the CQ clicks. Given that each CQ category was randomly selected to be shown to users, initially, every category had an equal chance of being selected, and thus the number of CQ showing times in C1, C2, or C3 is similar. As we stopped showing the CQs already clicked by the user in a search session to avoid disturbing the user, more clicks resulted in relatively lower showing times. Unless otherwise reported, we performed *t*-tests [42] and one-way **analysis of variance** (**ANOVA**) for statistical analysis in this study, assuming the independence of different groups [58, 71]. In particular, for comparisons between two groups only, we used *t*-tests; for comparisons between more than two groups, we performed one-way ANOVA and **least significant difference** (**LSD**) post-hoc tests [69], thus controlling for Type I errors, as in [10].

To ensure the *data quality*, we performed two quality checks and filtered out low-quality participants: (a) we asked questions about the study instructions to ensure that participants had read it carefully and understood it, and (b) we measured the time participants spent reading the experimental task descriptions and filtered out participants who spent less than 10 seconds (a minimum expected threshold for a trustworthy worker [28]). We did not filter users based on their interactions with CQs.

Table 3. Statistics of Collected Data

# users	106
# experimental tasks	5
# search requests	1,334
# search sessions	330
# CQs	15
# CQ showing/hiding times	1,016/318
# CQ clicks	249
# user bookmarks	1,942
# user clicked results	705
# user cursor hovering records	17,780
# user page scrolling records	15,747
# CQ showing times in C1/C2/C3	417/368/231
# CQ clicks in C1/C2/C3	44/62/143
# CQs shown but ignored in C1/C2/C3	373/306/88
avg. # experimental tasks per user	3.11
avg. # search requests per user	12.58
avg. # search requests per experimental task	266.8
avg. # bookmarks per user	18.32
avg. # bookmarks per experimental task	388.4

4.1 RQ1: Effect of CQs on Search Behavior and Satisfaction

In **RQ1**, we investigate the effect of CQ quality categories on user search behaviors and satisfaction. Regarding search behavior, we consider three types of measures [38]:

- *Interaction*: number of queries issued, number of query terms, number of SERP scrolls, number of SERP hovers, number of SERP clicks, number of CQ clicks, and user engagement.
- *Performance*: number of results marked relevant (# bookmarks), number of correct bookmarks (# hit), and SERP quality measured by nDCG@10 (normalized discounted cumulative gain from rank 1 to 10).
- *Time spent*: dwell time on SERPs per query and the overall task time for an experimental task.

Regarding satisfaction, we use explicit feedback collected through the post-task questionnaires:

- Overall satisfaction rating.
- User-perceived helpfulness.
- User attitude toward future use of CQ-based search systems.

4.1.1 Query-level Behaviors by CQ Quality Category. We begin the analysis of search behavior at the query-level. Table 4 presents behavioral measures under different conditions, that is, when a user clicked on a CQ pane of a certain quality (C1, C2, and C3), and when no CQs were shown to the user ("No CQs"). Throughout Table 4 we observe similar behavior for all the metrics reported. When users engage with low-quality CQs (i.e., C1), all metrics are low, typically, considerably lower than the metrics for a search interface that does not offer CQ panes. The metrics increase when users engage with mid-quality CQs (i.e., C2); in this case, all metrics are on par with those in the case of searching without CQs. All metrics significantly increase when users engage with high-quality CQs (i.e., C3).

Table 4. Objective Behavior Measures by Condition, i.e., Clicking on an Answer Related to Each CQ Quality Category

	No CQs	C1	C2	C3
# bookmarks/page	1.56(2.14)††	0.41(0.86)***† † †	1.06(1.54)† † †	2.22(2.30)**
# hits/page	0.73(1.21)† † †	0.07(0.25)***† † †	0.79(1.12)†	1.17(1.39)***
nDCG@10	0.27(0.30)† † †	0.12(0.19)**† † †	0.25(0.27)† † †	0.42(0.39)***
SERP scrolls	12.73(14.70)† † †	3.64(5.99)***† † †	7.63(9.85)*† † †	19.20(17.63)***
SERP hovers	12.93(13.39)† † †	6.55(6.78)*† † †	9.85(10.79)† † †	19.55(21.25)***
SERP clicks	0.55(1.15)†	0.09(0.47)*† † †	0.27(0.65)††	0.83(1.43)*
dwell time(s)	56.32(68.83)†	36.61(83.18)† † †	34.38(33.82)*† † †	73.87(65.41)*

* and † denote significant difference with No CQs and C3, respectively (*/† p-value < 0.05; **/†† p-value < 0.01; ***/† † † p-value < 0.001).

Bookmark quality. Table 4 indicates that compared with searching without CQs ("No CQs"), engaging with high-quality CQs (i.e., C3) leads to a significant increase in the number of (correct) bookmarks, that is, "# bookmarks/page" and "# hits/page"; the opposite occurs when the user engages with CQs that belong to C1. The number of (correct) bookmarks also significantly increases from either C1 or C2 to C3. Accordingly, high-quality CQs aid users in finding relevant information, whereas mid- and low-quality CQs negatively affect the finding of relevant information.

SERP quality. SERP quality, measured by nDCG@10, significantly increases after users click on high-quality CQs (i.e., C3), but it significantly decreases after low-quality CQs (i.e., C1) are clicked on, compared with the SERP quality corresponding to "No CQs." Compared with engaging with C3, engaging with C1 or C2 CQs significantly lowers SERP quality.

SERP scrolls and hovers. Cursor movements, such as scrolling and hovering, are valuable signals for inferring user behavior and preferences [34]. Thus, we investigate the effect of different quality categories of CQs on SERP scrolls and hovers. We observe that the number of scrolls significantly increases after users engage with C3 compared with the number corresponding to C1/C2/"No CQs"; it significantly decreases after C2 or C1 CQs are clicked on, compared with the corresponding number for "No CQs." Furthermore, we note that the number of hovers also significantly increases after C3 CQs are clicked on compared with the number corresponding to C1/C2/"No CQs"; it significantly decreases after C1 CQs are clicked on compared with the corresponding number for "No CQs." These observations for scrolls and hovers indicate that users scan a SERP more extensively when they engage with a high-quality CQ and less in the case of mid- and low-quality CQ.

SERP clicks. The number of clicks on SERP results in significantly increases after users engage with high-quality CQs compared with the corresponding number for mid- and low-quality CQs or "No CQs"; it significantly decreases after engagement with low-quality CQs compared with the corresponding number for "No CQs." We attribute these outcomes to improved SERP quality, which can cause users to see more interesting results on the SERP after they engage with high-quality CQs, but less interesting results after they click on C1/C2 CQs.

Dwell time. Users spend significantly more time on SERPs after engaging with high-quality CQs than with C1/C2 CQs. This may be because users realize the absence of useful information and quickly move on after clicking on C1/C2 CQs. By contrast, they pay more attention and attempt to find more relevant results (# bookmarks) after clicking on C3 CQs. Compared with the dwell time corresponding to "No CQs," the dwell time after C3 CQs are clicked on increases significantly, but it decreases after users click on C1 CQs (not significantly) or C2 CQs (significantly, $p < 0.05$). This

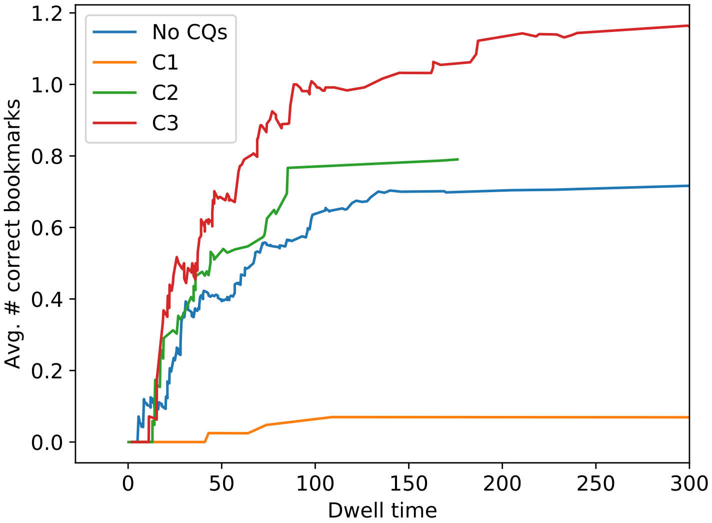


Fig. 3. Dwell time vs. # correct bookmarks (# hits).

again suggests that users scan a SERP more extensively when they engage with a high-quality CQ and less in the case of mid- and low-quality CQs.

Evolutionary trend of # hits with dwell time. To better understand the effect of different CQ quality categories, we explore the evolutionary trend of the average number of hits with dwell time on the secondary search result page (landing page) after users engage with a CQ pane (and hence a new query is submitted to the system), to consider the tradeoff between CQ engagement and dwell time. From Figure 3, it is seen that the average number of hits increases for C1, C2, C3, and “No CQs,” that is, spending more time leads to finding more relevant information. However, the growth rate greatly declines at later stages, indicating the diminishing return of gain [2, 8, 9]. Furthermore, clicking on high-quality CQs (i.e., C3) always leads to more relevant information found and is beneficial with respect to the dwell time cost, whereas clicking on low-quality CQs (i.e., C1) is less beneficial than “No CQs.”

In summary, findings emerging from our query-level exploration demonstrate that user behavior is greatly affected by CQ quality. The analysis indicated that interactions and user performance (relevant information found) are significantly improved when the user is offered and clicks on a high-quality CQ, whereas if the CQs are of low- or mid-quality, it is better for the user not to engage with them. This echoes the claims that users are willing to answer CQs if these are relevant and well-selected [82], suggesting that posing CQs to users causes higher interaction costs (in terms of time spent), yet, it does not always ensure better returns. Therefore, CQs introduce a high risk of user dissatisfaction and frustration [23]. Accordingly, future search systems should model the risk of asking CQs and optimize their performance based on that [66].

4.1.2 Session-level Behaviors by CQ Quality Category. In addition to examining the immediate effect of interacting with CQs on user behavior, we explore the effect of CQs across the entire session. The sum values for the selected measures in the entire session are presented in Table 5. After each user query, the CQs were randomly selected (for the group of participants that viewed CQs). Therefore, sessions with a single CQ category are rare. Hence, we split sessions into four categories, along two axes: (a) sessions without any C3 CQs (w/o C3) vs. sessions with C3 CQs (w C3), and (b) sessions in which users engaged (i.e., clicked on an answer) with a CQ vs. sessions

Table 5. Objective Behavior Measures by Condition, i.e., Viewing and Clicking/Not Clicking on CQs in a Session

		Click on Answer		No Click on Answer	
	No CQs	w/o C3	w/ C3	w/o $C3^-$	w/ $C3^-$
# bookmarks/session	5.36(2.47)	5.64(3.52)	6.05(3.18)	6.08(3.39)	6.10(2.93)
# hits/session	2.40(1.86)†	2.53(1.99)	3.17(2.45)*	2.66(2.21)	2.57(2.20)
nDCG@10	0.87(1.23)† † †	1.11(1.00)	1.42(1.29)***§§§	0.70(0.74)† † †	0.90(1.05)†
SERP scrolls	46.03(32.48)	38.03(29.69)†	55.67(43.13)§§	36.34(34.04)††	37.3(31.63)†
SERP hovers	42.90(30.36)† † †	43.14(29.20)††	65.15(48.80)*** §§§	41.06(32.70)† † †	39.37(23.48)† † †
SERP clicks	2.22(3.02)	0.96(1.74)†	2.34(3.19)	1.76(3.91)	2.7(3.41)
# queries/session	3.18(2.30)† † † §	4.50(2.13)*§§§	5.02(2.88)***§§§	2.28(1.28)*† † †	3.57(2.17)†† §
# query terms	11.43(12.19)† † † §	17.96(10.01)**§§§	19.24(12.70)***§§§	7.18(5.74)*† † †	11.3(7.77)† † †
task time(s)	204.75(173.20)§	188.32(187.84)	244.70(192.72)§§§	136.22(129.34)*† † †	170.85(128.32)†

, † and § denote significant difference with No CQs, w/ C3 and w/o $C3^-$, respectively. (/†/§ p-value < 0.05; **/††/§§ p-value < 0.01; ***/† † †/§§§ p-value < 0.001).

in which users did not engage with (i.e., skipped) the CQ panes. In principle, the results in Table 5 when the users click on high-quality CQ panes exhibit similar trends to those in Table 4. When users do not click on CQ panes (the last two columns), we still observe an interesting gap between high-quality CQ panes and lower-quality CQ panes, indicating that the quality of CQ panes has indirect effects on user behavior in this case.

Bookmark quality. As is the case with query-level trends, clicking on an answer from a C3 CQ pane (w/ C3) significantly increases the number of correct bookmarks in the entire session compared with the corresponding number for "No CQs." Sessions without CQs ("No CQs") or sessions in which users click on an answer to C1 or C2 CQs (w/o C3) result in a comparable number of correct bookmarks. Furthermore, viewing C1 or C2 CQs with no clicks (i.e., w/o $C3^-$) results in a slightly higher number of correct bookmarks than clicking on C1/C2 CQs (w/o C3), indicating that clicking on a low-quality or mid-quality CQ is harmful.

SERP quality. Clicking on C3 CQs in a session (w/ C3) significantly increases the SERP quality measured by nDCG@10 compared with the values corresponding to "No CQs" and to skipping C3 (w/ $C3^-$).

SERP scrolls and hovers. Engaging with C3 CQs in a session (w/ C3) significantly increases the SERP scrolls and hovers compared with the numbers corresponding to w/o C3 and w/ $C3^-$ sessions (skip C3).

SERP clicks. The number of clicks for SERP results significantly increases when users click on C3 CQs compared with the number corresponding to C1 or C2 CQs (w/ C3 vs. w/o C3).

User queries. Clicking on CQs of either w/o C3 or w/ C3 sessions significantly increases the number of query terms and session length. This indicates that users tend to formulate significantly shorter queries by themselves ("No CQs") than automated reformulated queries by clicking CQs, demonstrating an advantage of CQs when long queries are required.

Overall task time. Clicking CQs of w/ C3 sessions slightly increases the mean dwell time per experimental task, whereas clicking CQs of w/o C3 sessions slightly decreases it (not significantly different). This may be because users pay more attention and can locate more relevant results to bookmarks (average number of bookmarks per search session: 6.05 vs. 5.36). Skipping C3 (w/ $C3^-$) significantly decreases dwells time compared with clicking on C3 (w/ C3). In addition, we also find that the mean dwell time per experimental task corresponding to clicking on C3 CQs for information gathering tasks is higher than for fact-finding tasks (260.95 s vs. 223.47 s). This may

Table 6. User Satisfaction Measures by Condition, i.e., Viewing and Clicking/Not Clicking CQs in a Session

		Click on Answer		**No Click on Answer**	
	No CQs	w/o C3	w/ C3	w/o $C3^-$	w/ $C3^-$
satisfaction	2.88(0.99)†	3.14(1.06)	3.26(1.14)*§§	2.72(0.98)††	2.70(1.00)†
future use(%)	48.61/16.67	53.57/32.14	70.63/8.39	44.00/22.00	50.00/23.33
helpfulness(%)	57.14/28.57	34.62/53.85	59.86/29.58	16.67/77.78	13.33/73.33

Satisfaction shows means followed by standard deviations in parenthesis. *, † and § denote significant difference with No CQs, w/ C3 and w/o $C3^-$, respectively (*/†/§ p-value < 0.05; **/††/§§ p-value < 0.01; ***/† † †/§§§ p-value < 0.001). Future use and helpfulness are represented by ratio of positive/negative ratings.

be because information-gathering tasks are more complex than fact-finding tasks, and thus more time is needed to locate the relevant information.

Overall, from session-level analysis, we observed that interacting with CQs influences user behavior and satisfaction in the entire session. Sessions with high-quality CQs lead to higher session-based search performance than sessions without CQs. This is true even when users do not actively engage with high-quality CQs by clicking on their answers. Moreover, we noticed that while engaging with low- or mid-quality CQs decreases performance measures at the query-level, it improves them at the session-level. This suggests that even though these CQs lead to worse immediate performance, they may improve the search performance for the session. A plausible explanation is that low- or mid-quality CQs implicitly aid users by providing hints about the domain and the topic, so that users may effectively reformulate their queries. In fact, one of our participants mentioned that "I forgot the name of a state, so the question helped me clarify my search."

4.1.3 Satisfaction by CQ Quality Category. Regarding satisfaction, we collected explicit feedback from users through post-task questionnaires. From Table 6, it is seen that user satisfaction significantly improves when users interact with C3 compared with that corresponding to "No CQs" (w/ C3 vs. "No CQs"); when they skip C3, user satisfaction decreases significantly compared with that corresponding to clicking C3 CQs (w/ $C3^-$ vs. w/ C3).

To gauge user attitudes toward future usage of CQs-based systems and perceived helpfulness, we inquired about users' positive, neutral, or negative attitudes in the post-task questionnaires. Based on the percentage of positive and negative ratings across groups, we note that adding C3 CQs in the session improves user attitude toward future use of CQ-based systems (w/ C3 vs. w/o C3: more positive ratio and less negative ratio for w/ C3). Users who engage with high-quality CQs are more positive than those who do not (w/ C3 vs. w/ $C3^-$). Regarding user-perceived helpfulness, most users in w/ C3 (59.86%) are positive, whereas users in w/o C3 are in principle negative (53.85% negative). Adding C3 CQs in the session also improves user-perceived helpfulness (w/ C3 vs. w/o C3). Viewing CQs but not clicking them yields a considerably lower percentage of positive users and a considerably higher percentage of negative users (w/o $C3^-$ and w/ $C3^-$ vs. "No CQs"), indicating that showing improper CQs can lead to user dissatisfaction.

Previous studies [35, 46] suggest that the last impression (query) in a session may have a stronger correlation with user search satisfaction. Accordingly, we study whether there is an effect depending on when the interaction with CQs took place. By dividing each session into three segments (first query, in-between queries, and last query) [35], we indeed note the last impression effect: users are significantly more satisfied when user interaction with CQs occurs in the last query than in the first query ($p < 0.05$) or in-between query ($p < 0.01$) (average satisfaction score: 3.27, 3.04, and 3.62 for first, in-between, and last query, respectively). This indicates that the last impression contributes more to user satisfaction, suggesting that user satisfaction in a session may be better measured by individually modeling user interaction in each query for future studies.

In summary, clicking on high-quality CQs improves user satisfaction, whereas users feel less satisfied when they skip CQs. This suggests that user interactions with CQs have a positive effect on user-perceived satisfaction and overall attitude [61]; however, CQs may also introduce a risk of user dissatisfaction and frustration [23].

4.2 RQ2: Effect of User Background on CQ Interactions

In **RQ2**, we explore the effect of user background, as indicated by demographics, in addition to user perception of the extent of interactions with CQs, as measured by the CQ answer click-through rate (answer CTR, that is, total clicks divided by total showing times).

4.2.1 User Demographics. As stated by Weber and Jaimes [67], user demographics, such as age, education, and gender, are among the most important predictors of online information search behavior. This motivates our study of intrinsic characteristics that may induce users to interact with CQs in their quest for information, so that we may gain valuable insights regarding whether and when CQs should be shown to different users.

Gender. Our analysis indicates a significantly higher answer CTR for female users than for male users (25.2% vs. 23.6%, $p < 0.05$), as well as a lower correct answer CTR, i.e., the number of correct answers clicked compared with the total answer clicks (51.4% vs. 55.6%), pointing to information processing differences between females and males [40].

Language. Language proficiency affects interactions, as a decrease in the proficiency level (native speaker → proficient → beginner), leads to a decrease in answer CTR for C3 CQs (67.5% → 63.6% → 56.1%). Beginners engage significantly less with C3 CQs than native speakers ($p < 0.05$) and proficient speakers ($p < 0.01$). In addition, the overall answer CTR and correct answer CTR drop (26.3% → 25.7% → 21.8%, and 56.8% → 55.6% → 50.6%, respectively) with a decrease in the proficiency level. These observations on language proficiency are in agreement with those by Kiesel et al. [39] regarding voice query clarification.

Education. Answer CTR for C3 CQs grows among users with a higher education background (57.1%, 65.6%, 78.3%, and 100% for high school, bachelor's, master's, and doctorate degrees, respectively). Users with bachelor's degrees or above engage significantly more with C3 CQs than users without bachelor's degrees ($p < 0.05$); this is also the case for users with doctorate degrees ($p < 0.001$).

Career field. As anticipated, computer science students, likely more well-versed in search literacy instruction, exhibited higher overall answer CTR with CQs than participants of other occupations (25.2% vs. 22.7%). While the difference is not significant in the overall answer CRT, we notice a higher answer CTR in C3 (64.1% vs. 56.3%) and a higher correct rate for correct answer clicks (55.7% vs. 48.4%).

Age. Age did not emerge as a factor influencing CTR, that is, there were no obvious trends or significant differences across age groups.

The outcomes of the demographic analysis demonstrated that user demographic traits impact the way in which users interact with CQs. In general, user interactions with CQs are affected by gender, language proficiency, and educational background, but they are not affected by age.

4.2.2 User Perception. In addition to user demographics, we examine the effect of various perceived factors based on explicit feedback collected from pre- and post-task questionnaires.

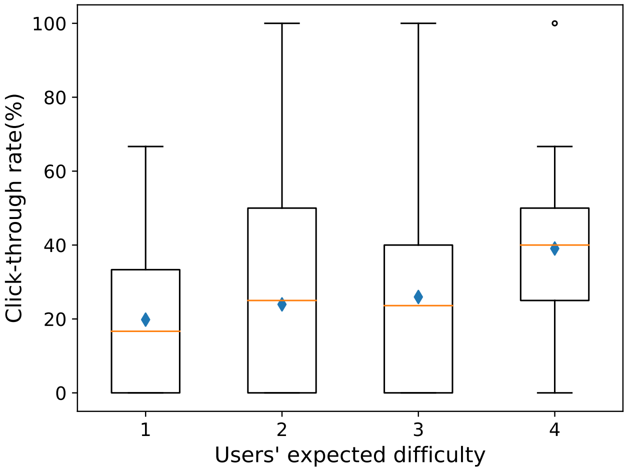


Fig. 4. Expected task difficulty level on answer CTR.

User expected difficulty. As shown in Figure 4, users engage more with CQs when completing an experimental task that is expected to be more difficult. The one-way ANOVA test shows significant differences among different groups ($p < 0.05$), and the post-hoc LSD test indicates that users completing experimental tasks of high difficulty level "4" are significantly more engaged with CQs than those engaging with experimental tasks of lower levels of difficulty, that is, "1", "2", and "3" ($p < 0.05$). Moreover, when users expect an experimental task to be more difficult, the answer CTR on C3 panes increases (46.3%, 64.2%, 67.2%, and 78.6% for difficulty level "1", "2", "3", and "4", respectively).

User-perceived difficulty. Users engage more with C3 as the user-perceived difficulty level increases after task completion (50.9%, 64.3%, 64.7%, 65.6% for difficulty level "1", "2", "3", and "4", respectively), as is the case with user expected difficulty.

User distraction. We expect users to be more eager to interact with CQs when they are distracted, as we assume that CQs could provide some assistance to ease the process; indeed, the overall answer CTR increases (not significantly) with the distraction level: 22.9%, 24.7%, and 30.6% for "not distracted" (distraction level "1"), "moderately distracted" (distraction level "2"–"4"), and "highly distracted" (distraction level "5"), respectively.

Search expertise. We posit that less-experienced users would be more willing to interact with CQs to successfully complete experimental web search tasks. Less experienced users (those using search engines weekly) indeed achieve higher mean and median values of the overall answer CTR (24.2% vs. 31.3%) and answer CTR in C3 (61.7% vs. 66.7%) than those using search engines daily (not significantly). However, they also obtain a lower correct answer rate (55.1% vs. 33.3%). This indicates that adding a new feature to the search can be confusing to less experienced users.

Other traits. We also considered users' prior knowledge of the task, users' task interest, perceived task clarity, perceived task relevance, and domain knowledge. No obvious trends or significant differences were observed.

Overall, the analysis of user perceptions indicate that user interactions with CQs are severely affected by user expected difficulty for the experimental web search task. Users engage significantly more with CQs when completing an experimental task with higher expected difficulty.

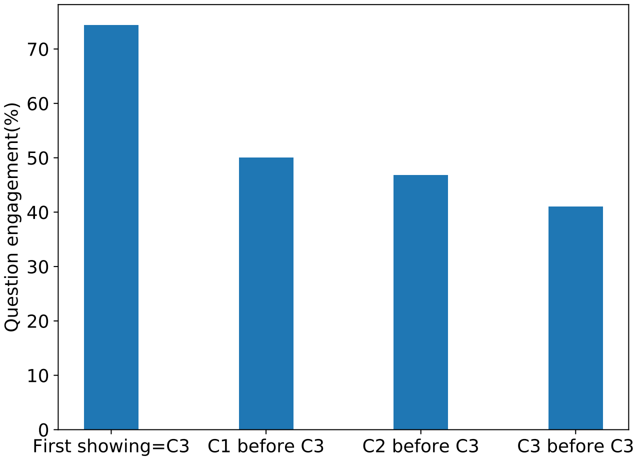


Fig. 5. User engagement on C3 questions, for different preceding CQ categories. For example, "C1 before C3" reports user engagement for cases when users saw C1 first and then C3.

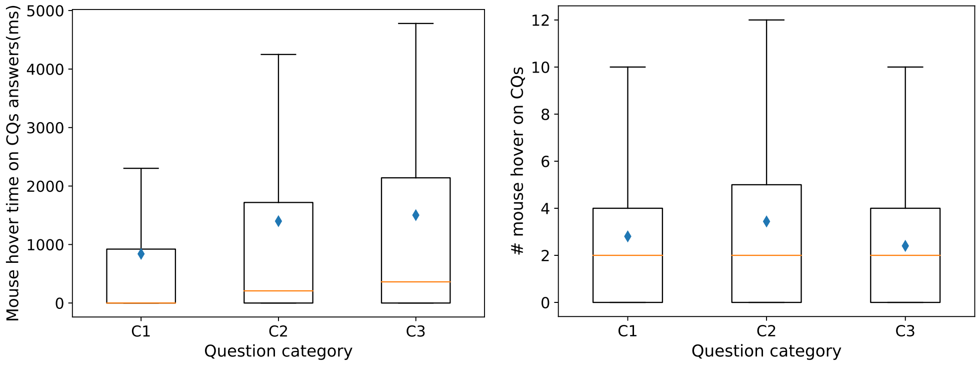
User-perceived difficulty, distraction level, and search expertise also impact user interactions with CQs, but not as much as the user expected difficulty. Users' prior knowledge of the task, users' task interest, perceived task clarity, perceived task relevance, and domain knowledge seldom influence user interactions with CQs.

As indicated by Kim [41], there is a positive relationship between pre-task difficulty and web search interactions such as page viewing. Similarly, findings from our study suggest that user-expected difficulty is an effective indicator of CQ interactions. Our observation for search expertise is in line with that by Kiesel et al. [39] observed when studying voice assistants: expertise has a weak effect. Recent studies [3, 30] have shown that the current user context can lead to different levels of distraction, thus affecting user performance and behavior. Our findings suggest a similar effect concerning CQs. Moreover, we observed that users do not always click the correct answers. They also click wrong answers irrespective of quality categories, which is in line with the findings of Zou et al. [82], who report that users provide noisy answers to CQs, and future research should drop the assumption that all questions are answered correctly.

4.3 RQ3: User Engagement with CQs under Various Circumstances

Given the absence of explicit feedback in a real-world setting, it is important to understand the dynamics of user engagement with CQs. For instance, how much do users engage with high-quality CQs as opposed to low-quality ones? How much does the quality of the search results affect users' tendency toward engaging with CQs? Even though Zamani et al. [75] conducted a large-scale industrial study to examine various queries, CQ templates, and answer attributes on engagement, they could not control variables, such as task type, CQ quality, and SERP quality. Our study setup enables us to provide considerably deeper insights into *why* and *how* users engage with CQs under various circumstances. Similar to Piccardi et al. [47] and Baird et al. [11], we use two engagement metrics to measure user interest in CQs: (a) CQ answer CTR and (b) cursor hovering over the CQ pane.

CQ quality categories. First, we examine how different CQ quality categories affect the answer CTR. As expected, the CTR increases as C1 → C2 → C3. C3 achieves the highest CTR (61.9%) compared with C1 (10.6%), C2 (16.8%), and overall CTR (24.5%). Significant differences are observed



(a) (b)

Fig. 6. Cursor hovering on CQ quality categories.

between C1 and C2 ($p < 0.05$), C1 and C3 ($p < 0.001$), and C2 and C3 ($p < 0.001$). We attribute this to C3 consisting of related and useful CQs, that is, it is the highest-quality category.

We also consider the effect of showing a low-quality CQ before a high-quality CQ, as shown in Figure 5. Showing C1 or C2 CQs before C3 CQs results in lower CTR in C3 (50% and 46.9%, respectively) than only showing C3 CQs (74.4%). This suggests that showing low-quality CQs before high-quality CQs reduces the level of user engagement with CQs. We also observe that users may click on C3 CQs even though they encountered the same CQs but did not click on them, however with a relatively low CTR (41.0%).

We investigate variations, if any, on cursor hover time and the number of hovers across the different qualities of CQs. As shown in Figure 6(a), cursor hover time on answers to CQs per search request significantly increases from C1 to C2 ($p < 0.001$) to C3 ($p < 0.001$). In Figure 6(b), it is seen that C2 obtains a significantly higher number of hovers on the CQ pane (including questions and answers) than C3 ($p < 0.01$) and C1 ($p < 0.05$). This may be due to the confusion that C2 can cause users, as this category offers useful questions but useless answers.

Task types. As in previous studies indicating that search tasks may influence information-seeking behavior [25, 68], we also observe that the overall answer CTR varies among experimental web search tasks (Figure 7). The fact-finding tasks T1 and T2 have a higher overall CTR than the information gathering tasks T3, T4, and T5. This may be because fact-finding tasks are simpler than information-gathering tasks [25], and users can foresee the benefit of answering the CQs. Moreover, ambiguous tasks (T3 and T4) received a lower overall CTR than faceted tasks (T1, T2, and T5), but with a higher correct answer click rate (0.66, 0.63, 0.45, 0.55, and 0.51 for T3, T4, T1, T2, and T5, respectively). The ANOVA test indicates significant differences in the overall CTR among the experimental tasks ($p < 0.01$). The post-hoc pairwise comparison indicates significant differences between T1 and T3 ($p < 0.01$), and T1 and T4 ($p < 0.01$), which are category-orthogonal tasks, that is, they have no category overlap.

Query index. Zamani et al. [75] demonstrated that the CTR increases for longer queries. Instead, we explore whether the CTR increases with the query index (i.e., ith query). We first compare user engagement between the first CQ pane shown to the user and the subsequent panes, with the CTR being 39.9% and 19.3%, respectively. A similar observation can be made regarding the number of cursor hovers (mean: 3.57 vs. 2.73, $p < 0.01$) and cursor hover time (mean: 1,885 ms vs. 958 ms, $p < 0.001$). Moreover, among all CQ clicks, 41.4% occur the first time a CQ is shown to the user. This suggests that, for each search session, the first instance of showing a CQ is highly important, and

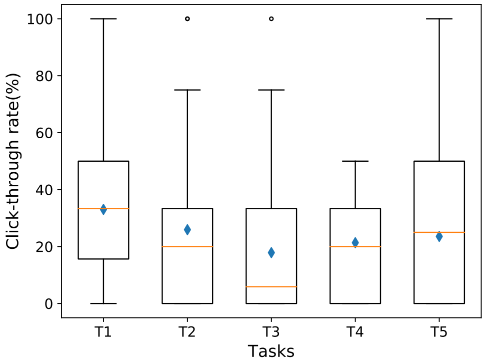


Fig. 7. CTR by experimental tasks.

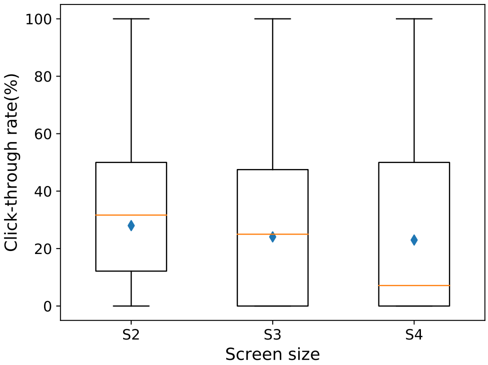


Fig. 8. CTR by screen sizes.

users pay particular attention to it. We also consider whether the overall CTR increases with the query index. We observe that users gradually lose their enthusiasm for CQs as the query index increases (39.9%, 32.6%, 21.1%, 11.4%, 18.0%, 4.1%, 6.5%, 0%, and 0% for CTR from the 1st query to the 9th query, respectively).

SERP diversity and quality. To increase user satisfaction, search engines often show diverse results [70]. We explore whether more diverse SERPs prompt more CQ interactions. As in Web TREC, we use nERR-IA@10 as a diversity metric. We divided the nERR-IA@10 values into four equal bins: [0, 0.2), [0.2, 0.4), [0.4, 0.6), and [0.6, 0.8). We observe that users engage more with CQs as SERP diversity increases (overall CTR: 22%, 22%, 29%, and 36% for [0, 0.2), [0.2, 0.4), [0.4, 0.6), and [0.6, 0.8), respectively).

As determined in the study of the effect of SERP quality using NDCG@10 on CTR, we also observe that users engage more with CQs as SERP quality increases (overall CTR: 22%, 22%, 24%, and 36% for [0, 0.2), [0.2, 0.4), [0.4, 0.6), and [0.6, 0.8), respectively).

Screen size. According to the layout change of the CQ pane, we divided screen size into five categories based on screen width: S0 (width 0–690 px), S1(691–1,063 px), S2 (1,064–1,437 px), S3 (1,438–1,811 px), S4 (1,812 px–∼). From Figure 8, we see that a larger screen obtains a relatively lower overall CTR (28.0%, 24.1%, and 23.0% for S2, S3, and S4, respectively), but with a higher correct answer click rate (47.6%, 57.3%, and 57.4% for S2, S3, and S4, respectively). Users with a smaller screen size of S2 engage significantly more with CQs than those with a larger screen size of S4 ($p <$ 0.05). This indicates that, for users, there is a greater need for CQs on smaller screens (e.g., smartphones). This may be because it is less practical to scan SERPs on small screens [73]. Our finding provides a better understanding of how users interact with CQs under different screen sizes. In turn, it can help researchers and system designers improve search clarification systems by realizing when to expose users to more CQs depending upon users' screen sizes, since always showing CQs to the users may introduce a risk of user dissatisfaction and frustration [23]. Moreover, previous studies have demonstrated that small-screen devices, such as smartphones, are usually used on the go leading to fragmented user attention [3, 30]. Therefore, our findings motivate further investigation of the effect of CQs under various interaction modalities (e.g., smartphone screen vs. voice [39]) and contexts (e.g., walking vs. driving [62]).

In summary, as a result of exploring various circumstances, we note that user engagement with CQs varies with the CQ quality category, task type, query index, degree of SERP diversity and quality, as well as screen size. Specifically, we observed a significantly higher CTR for high-quality CQs. More importantly, we observed that engagement decreases toward the end of a session. We found that users pay more attention to the first CQ shown to them in a session, whereas showing low- and mid-quality CQs before high-quality CQs reduces user engagement in terms of the answer CTR. In addition, we observed users need more support from CQs on smaller screens.

We also demonstrated that there are significant differences in cursor hovering signals among the CQ quality categories. This suggests that implicit user feedback, such as cursor hover time on answers to CQs, and the number of cursor hovers on the CQ pane, could be effective indicators of CQ quality. As we observed significant differences in the manner that users interact with CQs of different quality (e.g., different cursor hover behavior), a combination of such signals can potentially be used to predict the quality of a CQ [54, 56]. These findings related to inferring CQ quality have practical applications. Predicted CQ quality can be leveraged for the design of methods that learn from user interactions [1, 14] and online evaluation methods [32, 43]. Moreover, they can be used as weak labels to train models to generate high-quality CQ panes, especially in commercial systems with a large number of interactions between multiple users and CQ panes.

5 CONCLUSIONS AND DISCUSSION

In this article, we presented the results of a user study we conducted to investigate user interactions with search clarification panes. Our goal was to understand the effect of adding such an element to user behavior and experience. We first explored the effect of asking different quality categories of CQs on user behavior and satisfaction at the query- and session-level. We then investigated types of user background and perception that could lead to high engagement with CQs, and how users interact with CQs in various circumstances. From our analysis, we learned that systems should ask CQs only when it is certain that the questions are of high quality and the answers provided are appropriate, as asking useful CQs improves user performance and satisfaction, whereas asking unrelated or useless CQs can be harmful. We also identified features that affect user engagement with CQs. This can be used to predict user engagement and provide more personalized services for CQ-based systems.

Our results can facilitate the design of effective search clarification systems based on the query-level and session-level user interactions with different types of web search tasks and CQs. Even though this study focused on search systems, some of the findings can also be extended to conversational systems. For instance, in a conversational search setup, query difficulty prediction [7] could inform whether the system asks a CQ or shows the answer to the user. However, the results also demonstrated the complexity of user interaction with CQs at the session-level, which requires further investigation using other forms of user studies, such as eye-tracking or thinking aloud.

This study depends on an online CQ-based system developed on the ClueWeb09 corpus on the basis of ChatNoir. Even though ChatNoir is a proven and widely used web search engine [26, 27, 48, 59, 64] and the user interface mimics that of Bing, its retrieval effectiveness may not be on par with that of commercial search engines like Bing. This constitutes a limitation of our study, given that search quality affects user engagement with CQ panes. We leave the extended analysis based on other commercial search engines as future work.

A second limitation is that the number of experimental web search tasks considered—only 5—is limited. Many studies in our community (e.g., [20, 21, 23, 29, 30, 37, 38]) used a limited number of experimental tasks to control multiple factors while maintaining the cost of the study at a reasonable level. Nevertheless, it would be beneficial to extend the analysis to more experimental tasks in the future.

As a laboratory user study, it is common in our community to use a limited number of participants recruited from universities (e.g., [20, 21, 23, 29, 30, 38, 39, 65]). Despite the fact that we follow previous laboratory user study setups, participants represent university students and may not generalize to all user groups. It is, therefore, worth extending the user study with more participants from different user groups in the future. Also, as we involve participants with various backgrounds to explore the effect of user background on the interactions with CQs, the distribution of users with different backgrounds may be unbalanced (e.g., age). Hence, the findings are as good as our simulation of user cases, and it is worth studying the effect of user background involving balanced and large-scale users in the future.

There are no appropriate models to automatically generate various qualities of CQs and candidate answers. Thus, we manually create CQs and candidate answers, same as many studies in our community (e.g., [4, 6, 52]). Nevertheless, this may introduce some biases. We aimed at minimizing these biases by ensuring the expert annotators understand the CQ taxonomy and CQ generation pipeline, and then generate CQs and candidate answers following the CQ taxonomy and generation pipeline together with a review check. A possible future direction to mitigate this concern would be to provide a limited number of pre-generated CQs (e.g., 10 human-generated CQs and model-generated CQs [73]) to the experts to select from. One can also generate a pool of CQs either relying on humans or an automatic CQ generation model [73] and then label the CQ quality. As we found implicit user feedback, such as cursor hover time on answers to CQs, and the number of cursor hovers on the CQ pane, could indicate CQ quality, a plausible fully automatic approach could be employed, which is firstly using an automatic CQ generation model to generate CQ and then using implicit user feedback to detect the CQ quality.

Another clear future direction to extend this study is to consider different search settings, including different interfaces for interaction. For instance, it would be interesting to examine whether similar conclusions hold for voice-only conversational systems [3, 30].

REFERENCES

[1] Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2018. Never-ending learning for open-domain question answering over knowledge bases. In *Proceedings of the 2018 World Wide Web Conference.* 1053–1062.

[2] Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing mixed initiatives and search strategies during conversational search. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management.* 16–26.

[3] Mohammad Aliannejadi, Morgan Harvey, Luca Costa, Matthew Pointon, and Fabio Crestani. 2019. Understanding mobile search task relevance and user behaviour in context. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval.* 143–151.

[4] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. ConvAI3: Generating clarifying questions for open-domain dialogue systems (ClariQ). arXiv preprint arXiv:2009.11352.

[5] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail S. Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the EMNLP (1).* 4473–4484.

[6] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 475–484.

[7] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, and Ebrahim Bagheri. 2020. Neural embedding-based metrics for pre-retrieval query performance prediction. In *Proceedings of the European Conference on Information Retrieval.* 78–85.

[8] Leif Azzopardi. 2014. Modelling interaction with economic models of search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval.* 3–12.

[9] Leif Azzopardi, Mohammad Aliannejadi, and Evangelos Kanoulas. 2022. Towards building economic models of conversational search. In *Proceedings of the ECIR.*

[10] Leif Azzopardi, Mark Girolami, and Keith Van Risjbergen. 2003. Investigating the relationship between language model perplexity and IR precision-recall measures. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval.* 369–370.

[11] James Baird, Nick Redmond, Gregory Harrison, Brian Gebala, and John Kawamoto. 2016. Systems and methods for capturing and reporting metrics regarding user engagement including a canvas model. US Patent 9,390,438. [Access 12 Jul. 2016].

[12] N. Belkin, Colleen Cool, A. Stein, and U. Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems With Applications* 9, 3 (1995), 379–395.

[13] Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications* 9, 3 (1995), 379–395.

[14] Michael Bendersky, Xuanhui Wang, Donald Metzler, and Marc Najork. 2017. Learning from user interactions in personal search via attribute parameterization. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining.* 791–799.

[15] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic chatnoir: Search engine for the clueweb and the common crawl. In *Proceedings of the European Conference on Information Retrieval.* 820–824.

[16] Pia Borlund. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research* 8, 3 (2003), 8–3.

[17] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What do you mean exactly? Analyzing clarification questions in CQA. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval.* 345–348.

[18] Viktoriia Chekalina and Alexander Panchenko. 2021. Retrieving comparative arguments using ensemble methods and neural information retrieval. In *Proceedings of the Working Notes of CLEF 2021*, Vol. 2936. 2354–2365.

[19] Anni Coden, Daniel Gruhl, Neal Lewis, and Pablo N. Mendes. 2015. Did you mean A or B? Supporting clarification dialog for entity disambiguation. In *Proceedings of the SumPre-HSWI@ ESWC.*

[20] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2015. User activity patterns during information search. *ACM Transactions on Information Systems* 33, 1 (2015), 1–39.

[21] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval.* 163–172.

[22] Marco De Boni and Suresh Manandhar. 2003. An analysis of clarification dialogue for question answering. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.* 48–55.

[23] Ashlee Edwards and Diane Kelly. 2017. Engaged or frustrated?: Disambiguating emotional state in search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 125–134.

[24] George Ferguson, James F. Allen, and Bradford W. Miller. 1996. TRAINS-95: Towards a mixed-initiative planning assistant. In *Proceedings of the AIPS.* 70–77.

[25] Jacek Gwizdka. 2008. Revisiting search task difficulty: Behavioral and individual difference measures. *Proceedings of the American Society for Information Science and Technology* 45, 1 (2008), 1–12.

[26] Matthias Hagen, Martin Potthast, Payam Adineh, Ehsan Fatehifar, and Benno Stein. 2017. Source retrieval for web-scale text reuse detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 2091–2094.

[27] Matthias Hagen, Martin Potthast, Michael Völske, Jakob Gomoll, and Benno Stein. 2016. How writers search: Analyzing the search and writing logs of non-fictional essays. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval.* 193–202.

[28] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. Crowd worker strategies in relevance judgment tasks. In *Proceedings of the 13th International Conference on Web Search and Data Mining.* 241–249.

[29] Shuguang Han, Zhen Yue, and Daqing He. 2015. Understanding and supporting cross-device web search for exploratory tasks with mobile touch interactions. *ACM Transactions on Information Systems* 33, 4 (2015), 1–34.

[30] Morgan Harvey and Matthew Pointon. 2017. Searching on the Go: The effects of fragmented attention on mobile web search tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 155–164.

[31] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2020. Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1131–1140.

[32] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Measuring user satisfaction on smart speaker intelligent assistants using intent sensitive query embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 1183–1192.

[33] Chien-Ju Ho and Jennifer Vaughan. 2012. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[34] Jeff Huang, Ryen W. White, Georg Buscher, and Kuansan Wang. 2012. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 195–204.

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

[36] Diane Kelly. 2009. *Methods for Evaluating Interactive Information Retrieval Systems with Users.* Now Publishers Inc.

[37] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-Ching Wu. 2015. Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval.* 101–110.

[38] Diane Kelly and Leif Azzopardi. 2015. How many results per page? A study of SERP size, search behavior and user experience. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 183–192.

[39] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward voice query clarification. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1257–1260.

[40] Dae-Young Kim, Xinran Y. Lehto, and Alastair M. Morrison. 2007. Gender differences in online travel information search: Implications for marketing communications on the internet. *Tourism Management* 28, 2 (2007), 423–433.

[41] Jeonghyun Kim. 2006. Task difficulty as a predictor and indicator of web searching interaction. In *Proceedings of the CHI'06 Extended Abstracts on Human Factors in Computing Systems.* 959–964.

[42] Tae Kyun Kim. 2015. T test as a parametric statistic. *Korean Journal of Anesthesiology* 68, 6 (2015), 540.

[43] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval.* 121–130.

[44] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the effect of clarifying questions on document ranking in conversational search. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval.* 129–132.

[45] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How am i doing?: Evaluating conversational search systems offline. *ACM Transactions on Information Systems* 39, 4 (2021), 1–22.

[46] Mengyang Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating cognitive effects in session-level search user satisfaction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 923–931.

[47] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2020. Quantifying engagement with citations on wikipedia. In *Proceedings of the Web Conference 2020.* 2365–2376.

[48] Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, and Benno Stein. 2021. The information retrieval anthology. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2550–2555.

[49] Martin Potthast, Matthias Hagen, Benno Stein, Jan Graßegger, Maximilian Michel, Martin Tippmann, and Clement Welsch. 2012. ChatNoir: A search engine for the ClueWeb09 corpus. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1004–1004.

[50] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the SIGdial.* 353–360.

[51] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval.* 117–126.

[52] Pengjie Ren, Zhongkun Liu, Xiaomeng Song, Hongtao Tian, Zhumin Chen, Zhaochun Ren, and Maarten de Rijke. 2021. Wizard of search engine: Access to information through conversations with search engines. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 533–543.

[53] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020.* 1160–1170.

[54] Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. 2021. User engagement prediction for clarification in search. In *Proceedings of the European Conference on Informatin Retrieval.*

[55] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining.*

[56] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Exploiting document-based features for clarification in conversational search. In *Proceedings of the European Conference on Information Retrieval.*

[57] Anna Sepliarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. 2018. Preference elicitation as an optimization problem. In *Proceedings of the 12th ACM Conference on Recommender Systems.* 172–180.

[58] Yunqiu Shao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2019. On annotation methodologies for image search evaluation. *ACM Transactions on Information Systems* 37, 3 (2019), 1–32.

[59] Benno Stein, Tim Gollub, and Dennis Hoppe. 2012. Search result presentation based on faceted clustering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management.* 1940–1944.

[60] Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2014. Towards natural clarification questions in dialogue systems. In *Proceedings of the AISB Symposium on Questions, Discourse and Dialogue*, Vol. 20.

[61] Hock-Hai Teo, Lih-Bin Oh, Chunhui Liu, and Kwok-Kee Wei. 2003. An empirical study of the effects of interactivity on web user attitude. *International Journal of Human-computer Studies* 58, 3 (2003), 281–305.

[62] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White. 2020. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval.* 43–52.

[63] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 325–328.

[64] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Modeling the usefulness of search results as measured by information use. *Information Processing & Management* 56, 3 (2019), 879–894.

[65] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* 2187–2193.

[66] Zhenduo Wang and Qingyao Ai. 2021. Controlling the risk of conversational search via reinforcement learning. arXiv preprint arXiv:2101.06327.

[67] Ingmar Weber and Alejandro Jaimes. 2011. Who uses web search for what: and how. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining.* 15–24.

[68] Ryen W. White, Mikhail Bilenko, and Silviu Cucerzan. 2007. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 159–166.

[69] Lynne J. Williams and Herve Abdi. 2010. Fisher's least significant difference (LSD) test. *Encyclopedia of Research Design* 218 (2010), 840–853.

[70] Zhijing Wu, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Does diversity affect user satisfaction in image search. *ACM Transactions on Information Systems* 37, 3 (2019), 1–30.

[71] Xiaohui Xie, Yiqun Liu, Maarten de Rijke, Jiyin He, Min Zhang, and Shaoping Ma. 2018. Why people search for images using web search engines. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining.* 655–663.

[72] Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and SUN Xu. 2019. Asking clarification questions in knowledge-based question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.* 1618–1629.

[73] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the Web Conference 2020.* 418–428.

[74] Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020. Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management.* 3189–3196.

[75] Hamed Zamani, Bhaskar Mitra, Everest Chen, Gord Lueck, Fernando Diaz, Paul N. Bennett, Nick Craswell, and Susan T. Dumais. 2020. Analyzing and learning from user interactions for search clarification. arXiv preprint arXiv:2006.00166.

[76] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 177–186.

[77] Zhiling Zhang and Kenny Q Zhu. 2021. Diverse and specific clarification question generation with keywords. arXiv preprint arXiv:2104.10317.

[78] Kun Zhou, Wayne Xin Zhao, Hui Wang, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. Leveraging historical interaction data for improving conversational recommender system. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management.* 2349–2352.

[79] Jie Zou, Yifan Chen, and Evangelos Kanoulas. 2020. Towards question-based recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 881–890.

[80] Jie Zou and Evangelos Kanoulas. 2019. Learning to ask: Question-based sequential bayesian product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 369–378.

[81] Jie Zou and Evangelos Kanoulas. 2020. Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems* 38, 3 (2020), 35 pages.

[82] Jie Zou, Evangelos Kanoulas, and Yiqun Liu. 2020. An empirical study on clarifying question-based systems. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. 2361–2364.

[83] Jie Zou, Dan Li, and Evangelos Kanoulas. 2018. Technology assisted reviews: Finding the last few relevant documents by asking yes/no questions to reviewers. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* 949–952.

Received 21 May 2021; revised 13 January 2022; accepted 4 March 2022