An Eye-tracking Study of User Interactions with Query Auto Completion

Katja Hofmann, Bhaskar Mitra, Filip Radlinski, Milad Shokouhi
Microsoft, Cambridge, UK
{katja.hofmann, bhaskar.mitra, filiprad, milads}@microsoft.com

ABSTRACT

Query Auto Completion (QAC) suggests possible queries to web search users from the moment they start entering a query. This popular feature of web search engines is thought to reduce physical and cognitive effort when formulating a query.

Perhaps surprisingly, despite QAC being widely used, users' interactions with it are poorly understood. This paper begins to address this gap. We present the results of an in-depth user study of user interactions with QAC in web search. While study participants completed web search tasks, we recorded their interactions using eye-tracking and client-side logging. This allows us to provide a first look at how users interact with QAC. We specifically focus on the effects of QAC ranking, by controlling the quality of the ranking in a within-subject design.

We identify a strong position bias that is consistent across ranking conditions. Due to this strong position bias, ranking quality affects QAC usage. We also find an effect on task completion, in particular on the number of result pages visited. We show how these effects can be explained by a combination of searchers' behavior patterns, namely *monitoring* or *ignoring* QAC, and *searching* for spelling support or complete queries to express a search intent. We conclude the paper with a discussion of the important implications of our findings for QAC evaluation.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

Keywords

Query auto completion, eye tracking, evaluation

1. INTRODUCTION

Query Auto Completion (QAC)[1] is a popular feature of today's web search engines, and domain-specific retrieval systems. By suggesting a selection of possible completions based on the query prefix a user has typed so far (or even before the user has begun typing), QAC can help users formulate more effective queries in less time and with less effort. The resulting query formulation support is thought to reduce the effort of interacting with the search engine, leading to a more enjoyable and effective user experience (e.g. [32]). For instance, QAC may help users avoid spelling mistakes, discover relevant search terms, and avoid issuing overly ambiguous queries.

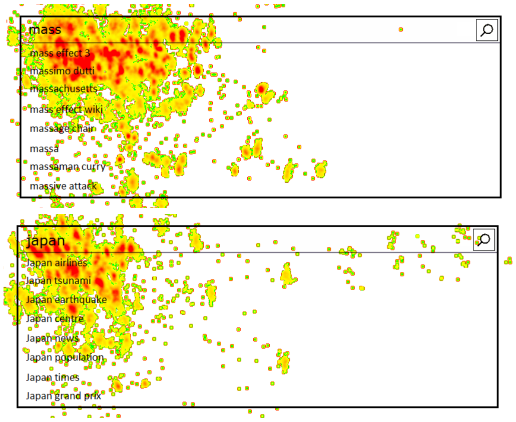

Figure 1: Example snapshots of query formulation with QAC for search tasks 1 and 5 (cf., Table 1), overlaid with heat maps of eye fixations for all participants on each task (across all query prefixes). Our results show that the focus on top suggestions is due to examination bias, not ranking quality (Section 5).

Despite the popularity of QAC, users' interactions with it are poorly understood. How many suggestions do users consider while formulating a query? Does position bias affect query selection, possibly resulting in suboptimal queries? How does the quality of QAC affect user interactions? And, inversely, can observed behavior be used to infer QAC quality? In this paper we present the results of an eye-tracking study designed to shed light on these questions.

We focus on the role that QAC ranking quality plays in this process. Information retrieval metrics, such as Mean Reciprocal Rank (MRR), have been proposed as evaluation metrics for QAC [4, 31], but without further insights in user interactions with QAC, it is not clear to what degree their underlying assumptions are warranted. This paper provides a first step towards addressing this gap.

As an example, the visualizations in Figure 1 show where searchers looked while formulating queries for two of the search tasks in our study. We see that in both tasks searchers examine primarily the top-ranked suggestions, in particular at ranks 1 to 3. This behavior could be due to the high quality of QAC rankings. If users aim to select a specific query they had in mind, and this query is shown towards the top of the suggestion list, then examining lower-ranked suggestions would be unnecessary. Alternatively, this behavior could be

[1] Also referred to as query completion [35], (dynamic) query suggestion [20, 30], and real-time query expansion [36].

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661922.

explained by *examination bias*, i.e., by users' expectations to find the best suggestions towards the top of the list. While position bias in interactions with QAC was identified in previous log studies, our study can distinguish between these alternative explanations.

The main contributions of this paper are:

- An eye-tracking study of user interactions with QAC; this includes the extension of the methodology for eye-tracking studies of user interactions with (web) search engines (Section 3), analysis (Section 4) and a discussion of potential limitations of the developed approach (Section 6.3).
- An experimental comparison of the effects of QAC ranking quality on users' search behavior (Section 5). We find a strong position bias that is consistent across ranking conditions. Ranking quality is found to affect suggestion use. Finally, we observe an effect on number of results visited to solve a search task, suggesting a link with query effectiveness.
- A descriptive analysis of user interactions with QAC. We identify common behavior patterns that generalize across participants, namely *monitoring*, *ignoring*, and *searching* (Section 6.1). We hypothesize that a combination of these patterns can explain the observed ranking effects.
- A detailed discussion of the implications of our findings for the evaluation of QAC systems, especially for the choice of evaluation metric (Section 6.2).

We now discuss related work (Section 2). Next, our experiment methodology is presented in Section 3 and our analysis method is described in Section 4. After presenting our results in Section 5, we discuss their implications, and potential limitations of our study, in Section 6. We conclude in Section 7.

2. RELATED WORK

To the best of our knowledge, this paper presents the first user study on the effects of QAC ranking quality on user interactions with QAC. However, eye-tracking studies are an established method for studying user interactions with search systems. They have led to important insights about how to interpret logs of user interactions with search results and how to evaluate search engine results [25].

Below, we first provide a short summary of the technical background of QAC (Section 2.1). We then overview user studies, and especially eye-tracking studies of search behavior (Section 2.2). Finally, we detail how QAC systems have been evaluated in the past, and how these evaluation setups and metrics relate to assumptions about user behavior that can be tested in user studies (Section 2.3).

2.1 Query Auto Completion

Early auto completion systems were developed for suggesting words and sentences in applications such as text editors and command shells. In *predictive auto completion* models, suggestion candidates are often filtered by exact-prefix matching. Ranking can happen on-the-fly [36] but in practice, for efficiency reasons, the order of suggestions is usually partially [14] or fully [4] precomputed.

In the context of QAC for web search, the most common approach is to rank suggestions according to their past frequency. This method effectively ranks suggestions according to their *maximum likelihood estimator* (MLE) upper-bound in the absence of any context [4]. However, Shokouhi and Radinsky [32] demonstrated that the MLE ordering of suggestions based on their predicted future popularity instead of their past leads to higher quality ranked lists.

While most current QAC systems ignore user context, Weber and Castillo [35] showed that query distributions change significantly across different demographics and suggested that QAC should reflect them. Inspired by similar observations, Bar-Yossef and Kraus [4] argued for using context in ranking QAC candidates. They leveraged users' recent searches as context and re-ranked the auto completion results accordingly. Their approach can rank `american presidents` before `american airlines` for prefix `am`, despite the overall higher popularity of the latter, in cases where the users have just searched for queries such as `richard nixon`. Shokouhi [31] showed that contextual (personalized) QAC rankers can be trained using supervised learning. For evaluating contextual QAC, reliably relating user behavior to QAC quality is crucial.

2.2 Eye-Tracking Studies in Search

Studies of search user behavior typically fall into one of two categories. First, large-scale log studies can test hypotheses based on large amounts of unobtrusively collected log data [12]. The large scale and realism of log data usually comes at the cost of control. In log-based studies, users' motivations and attitudes cannot be directly observed and the obtained results depend on the accuracy of the interpretations of the observable activity. Second, smaller-scale user studies are an important complement that allow researchers to conduct carefully controlled experiments with rich recordings of user activity [19]. For example, users' motivations for searches can be controlled by providing fixed search tasks. In this work, we investigate how users examine and interact with QAC in a detailed, controlled lab study. This setup allows us to collect data about, among other things, users' eye gaze while formulating queries.

Eye-tracking and studying variations in gaze patterns emerged as a field of research in the late 19th century, when researchers first took note of characteristic eye movements during reading [28]. Since then, eye-tracking has evolved, and thanks to advancement in technology has become a widely-used means for analyzing human-computer interaction. Goldberg et al. [13] were among the pioneers of investigating the gaze patterns of users when browsing different types of web-pages. Salojärvi et al. [29], Granka et al. [15] and Joachims et al. [18] were among the first to suggest that eye-tracking can be used to capture and understand user attention, and therefore infer relevance from behavioral observations.

Salojärvi et al. [29] proposed to infer document relevance directly from gaze data. Granka et al. [15] and Joachims et al. [18] instead asked what users look at prior to interacting with a search engine, manipulating the search results and monitoring changes to better understand how behavior is affected by relevance. Guan and Cutrell [16] showed that people tend to look at the top-ranked results only, and when they do not find their target they either click on the first result or reformulate their query. In addition to organic search results, other components of result pages have also been the focus of eye-tracking studies. For instance, Cutrell and Guan [10] discussed how changes in result snippets affect users' gaze patterns and search satisfaction. Buscher et al. [7] studied the correlation between visual attention on search results, and the quality of other components on the page such as advertisements and related searches. They showed that the quality and position of advertisements can substantially affect the gaze patterns of users on organic search results.

2.3 QAC Use and Evaluation

Despite the wide use of QAC, there have been few user studies in this area. White and Marchionini [36] were likely the first to propose QAC for web search, and found that it can improve query quality and users' search satisfaction. Shah et al. [30] compare exploratory search behavior on a baseline system to systems with QAC, and with QAC and dynamic search pages (Google Instant). They find that dynamic systems can decrease query formulation time and query length, and exposes users to a larger number of search result pages. Hawking and Griffiths [17] study QAC in enterprise search. They show that enhancing QAC with faceted-navigation can reduce the search time in known-item search. Compared to our work, none of the previous studies investigated the effect of QAC ranking quality.

Recent work on QAC has typically focused on log studies, both to examine hypotheses on user interactions with QAC, and to derive evaluation metrics. Mitra et al. [26] present a large-scale log study, and observe strong position bias toward top-ranked suggestions. They identify patterns in when users are likely to use suggestions, e.g., at word boundaries.

To evaluate QAC rankings, Bar-Yossef and Kraus [4] and Shokouhi [31] considered submitted queries as ground-truth. Shokouhi and Radinsky [32] and Strizhevskaya et al. [33] considered aggregated query frequency as an oracle QAC list for each prefix. The ground truth derived from log data was compared to generated suggestions in terms of established rank-based information retrieval metrics, such as MRR. This evaluation setup assumes that items placed towards the top of a ranked list receive more attention, and are therefore more useful to a searcher. In the context of spell correction for QAC, Duan and Hsu [11] additionally proposed the use of Minimum Keystroke Length (MKS), which measures the number of actions a user has to take to submit a target query.

Kharitonov et al. [20] propose a model of user interactions with QAC based on insights into user modeling and understanding click behavior in web search. They adapt Chapelle and Zhang's Dynamic Bayesian Network (DBN) click model [9] for QAC evaluation using query logs. Their model assumes that the user interacts with suggestions from top to bottom, seeking out specific suggestions they have in mind. Based on this model, they propose two metrics, *e-Saved* and *p-Saved*, for evaluating the quality of QAC ranked lists. The former measures the amount of effort saved in terms of key-strokes and the latter computes the expected QAC usage. Alternative evaluation procedures are proposed in [6, 24]. Bhatia et al. [6] use manual judgments for each suggestion, while Liu et al. [24] measure the quality of the search results retrieved by each suggestion.

We argue that our results are valuable in understanding user interactions with QAC, and can provide useful insights about the properties of an ideal evaluation metric. We will return to the assumptions underlying QAC evaluation metrics in Section 6.

Table 1: List of search tasks in our study. To balance the requirements for low variance and realism, we designed search tasks of different types, on a variety of topics.

ID	Search task
	Practice tasks
a	The Queen of England is the head of many commonwealth nations. Find out how many states she is sovereign of.
b	When was the car company with this logo founded? [logo presented]
	Navigational
1	Find the homepage of the Massachusetts General hospital in Boston, USA. What is their physical address?
2	Find the homepage of the Canadian tax office. What is the next payment date for the "Universal child care benefit"?
3	Go to the homepage of the Wall Street Journal. Which news story do you find the most interesting?
4	Find a web page where you can download Apple QuickTime. What version is currently available for download?
	Informational (easy)
5	Japan is the 10th most populated country in the world. Find out how many people live there.
6	The 2012 Olympic torch relay in the UK lasted for 70 days. Find the northern-most place in the route.
7	Arnold Schwarzenegger is an Austrian American actor, politician, businessman, investor, and former professional bodybuilder. Schwarzenegger served two terms as the 38th Governor of California. When was he sworn into office?
8	Queen Elizabeth II has had the 2nd longest reign of any king or queen in the various kingdoms of the British Isles. Can you find when she was crowned?
9	Machu Picchu is one of the most visited tourist destinations in South America. How far is it above sea level?
	Informational (complex)
10	You ran across a large pile of garbage while jogging along the river in Boston, USA. Find the homepage of an appropriate organization that you can report this to. What is their phone number?
11	Which two countries did the winner of the 2007 Nobel Prize in literature grow up in?
12	How many matches did Roger Federer win against Rafael Nadal in 2007?
13	How many children does the actress who plays the mother of the character played by Jim Parsons in the Big Bang Theory TV show have?
14	Who named the river that runs through Boston, and who is it named after?

3. USER STUDY

We describe our methodology in two parts. In this section, we describe the setup of our user study. Then, in Section 4, we describe how the collected data was analyzed.

3.1 Overview

We designed a user study to examine the effect of ranking quality on searchers' examination of, and interaction with, QAC. The main challenge in designing the study was to create a setting where query formulation was as natural as possible, while controlling variance. In the resulting study, we asked participants to find short answers to each of 14 search tasks starting from the homepage of a commercial search engine. During this interaction, the users were aware that their gaze as well as all interactions with the search system were being recorded. However, they were unaware that our real interest was whether and how they interact with QAC.

3.2 Apparatus and Procedure

The user study took place in a standard office setting. We used a Tobii TX-300 eye tracker[2], which consists of a 23" monitor with built-in eye tracker. It supports gaze tracking at 300 Hz, with an accuracy of up to a 0.4° visual angle. All on-screen material was presented in full-screen mode, with a resolution of 1920 x 1080 px.

The study proceeded in the following steps. After being introduced to the study, the participants read and signed a consent form. Next, the eye tracker was calibrated to the participant, and the participant was asked to complete two practice tasks. Next, the participants completed our 14 search tasks (Section 3.3), followed by an exit questionnaire and a semi-structured interview. The experiment took between 45 minutes and 1 hour per participant. Afterwards, participants were debriefed and received a gift certificate.

Participants received instructions and search tasks on-screen. One task was displayed at a time, and the task was only displayed before the search for the given task could be started. Once the participant had read the task, they were directed to the search engine to start their search for the answer. This setup was chosen to create a realistic scenario for query formulation, and validated in a pilot study before data collection started. Displaying the task before each search ensured that participants could not copy query terms from the task description (using copy-paste, or by typing verbatim based on the task text). Once the answer was found, participants were instructed to note it on an answer sheet provided, and to press a key to move to the next task. If a participant was unable to find the answer to a task within three minutes, they were automatically asked to move on to the next task.

[2] http://bit.ly/1dHpTxq, last accessed 29 May 2014.

3.3 Search Tasks

All participants completed the same 14 search tasks (Table 1). We chose a fixed set of tasks with closed answers to achieve a good balance of realism and low variance. More open tasks are expected to be more realistic, but also to result in more varied search behavior.

The tasks were designed to address navigational and closed informational searches. The closed informational search tasks were further divided into easy tasks, which were expected to be addressed by a single query, and complex tasks, which were expected to require several queries. In a pilot test, we found that the navigational tasks resulted in some confusion, because participants did not know when the task was completed. Therefore, we extended the navigational tasks by asking for a piece of information that could be found easily on the target page. Further, we attempted to word the tasks to avoid suggesting a single succinct query to the participants, although it is likely that the particular query words used by participants were influenced by the choice of wording in the task [34].

Because our participants were expected to have a variety of (international) backgrounds, we chose topics from a range of interest areas and locations. Finally, we included named entities that we expected to be difficult to type (e.g., Schwarzenegger, Machu Picchu) and that could be abbreviated (e.g., WSJ for Wall Street Journal, MA for Massachusetts), to see how these would affect query formulation.

3.4 Experimental Conditions and Design

For each task, participants experienced one of two experimental conditions. In the control condition (termed *original*), default QAC suggestions were shown to users exactly as generated by the production algorithm of the commercial search engine used. The treatment condition (termed *random*) degraded the quality of the QAC rankings by permuting them (uniformly) at random. The treatment was applied on the server, and we verified that it did not affect response times or other aspects of the user experience. This treatment allows us to distinguish effects of position from suggestion quality. Note that *within* each task, condition was consistent, e.g., if the user typed submitted several searches to solve a given task, they would experience the same condition as for the first query for this task. The seed for randomization was fixed per query to avoid inconsistencies between subsequent QAC impressions.

The two conditions were counterbalanced in a within-subject design. Specifically, four sequences of *original* and *random* QAC were constructed in a way that ensured a maximum of two tasks in either *original* or two *random* condition would follow each other. These sequences were chosen to avoid a learning effect (e.g., a user may learn to avoid suggestions if they experience many tasks in the lower quality *random* condition).

In our experiments, we used a Graeco-Latin square design of task and condition, in such a way that each task would be completed equally often in each condition, and following each condition. The design for the first four participants is illustrated in Table 2. This design was chosen to control for effects of task and condition order. However, the experiment design could not be maintained due to technical problems, and because participants occasionally chose to skip tasks, or to go to a known URL without using our target search engine. This results in an incomplete Graeco-Latin square design, in which tasks and condition sequences are pseudo-randomized but not fully counterbalanced. Consequently, we selected an analysis method that is robust to missing data (see Section 4).

To verify that our treatment had not been detected, we asked participants in the follow-up interview whether they had experienced any variation in search performance between tasks. None of the participants had noticed any differences between the two conditions.

Table 2: Sketch of our experiment design. Task IDs (see Table 1) and condition (O = *original* and R = *random*) are shown for the first four participants. Four sequences of condition are used, indicated by shading.

Task	a	b	1	2	14	3	13	4	12	5	11	6	10	7	9	8
Condition	O	R	O	R	R	O	O	R	R	O	R	O	O	R	R	O
Task	a	b	2	3	1	4	14	5	13	6	12	7	11	8	10	9
Condition	R	O	O	R	O	R	R	O	O	R	O	R	R	O	R	O
Task	a	b	3	4	2	5	1	6	14	7	13	8	12	9	11	10
Condition	O	R	R	O	R	O	R	O	R	O	O	R	O	R	O	R
Task	a	b	4	5	3	6	2	7	1	8	14	9	13	10	12	11
Condition	R	O	R	O	O	R	O	R	O	R	R	O	R	O	O	R

3.5 Participants

We recruited participants using snowball sampling, making a particular effort to include participants with a variety of backgrounds. 25 participants were recruited, including researchers, administrative staff, medical practitioners, students in diverse fields, programmers and others. Due to the initial contacts used, there was a bias towards computer scientists, however non-computer science fields were also represented (anthropology, physics, political science and linguistics). The participants were aged between 21 and 63 years old (median 33) with education levels ranging from high-school to PhD. 11 of the participants were female, and 10 were native English speakers (although the study was conducted in an English-speaking locale, and all participants routinely used English at work / study). The native languages of the remaining participants were distributed over 14 languages (no dominant language, two bilingual participants).

Based on the exit questionnaire, all participants were familiar with web search engines, but usage ranged considerably from tens of searches per week (12 participants) to hundreds of searches per week (the remainder). The participants reported using a computer between 10 and 105 hours per week (median 50).

4. ANALYSIS

During our user study, we collected the following data as participants completed the search tasks:

- All eye fixations and saccades, visited URLs, participants' mouse clicks and keystrokes (using the eye tracker software).
- Screen capture videos showing exactly what each user saw on the screen at each point in time.
- Browser event data collected using a custom web browser plug-in developed for the study.

Next, we describe how this data was processed and analyzed.

4.1 Processing

All results presented in this paper are obtained by analyzing only those search episodes, where the participant submitted the first query using our target web search engine. In 19 episodes (5.4%), participants searched directly on a more specific search engine (e.g., Wikipedia, IMDB, image search), or entered a URL directly in the browser address bar. We exclude these episodes because QAC cannot affect search behavior there. The resulting dataset consists of 331 search episodes (94.6% of the maximum possible 350).

For all 331 search episodes, we detected *query formulation*, *search result page examination*, and *result visits* (see Figure 2) using the URLs and timestamps obtained from our client-side logs.

The presence of QAC suggestions was detected using video analysis, by matching the location of the search box and QAC block on the screen. The complete suggestion block was then divided into individual suggestions based on the known suggestion size (30 px high). For all query formulations, QAC was displayed at least for initial (short) prefixes. Since we did not manipulate the number of suggestions shown, whether or not QAC was displayed depended only on the typed query prefixes, and the QAC algorithm of our target web search engine.

Participants' gaze data was analyzed using the Tobii Studio 3.2 software. Eye tracking data tracks the motion of the eyes and yields two types of events. *Fixations* occur when a person focuses their vision on a particular point, while changes in eye position are referred to as *saccades* [28]. Following standard practice, we focus our analysis on where and when fixations occur. Fixations are extracted using velocity threshold identification (I-VT, [21]) as implemented in Tobii Studio. This approach identifies fixations based on the eyes' angular velocity. We use the default settings, as detailed in [27]. The resulting output provides gaze timestamps, gaze type label (Fixation, Saccade), and the XY-coordinates obtained after noise reduction.

We then measure fixation time on QAC suggestion as the time during which fixations were recorded with XY-coordinates that fall in the region of the QAC suggestion areas. On the 23" monitor used in our experiments (Section 3.2), suggestions had a height of 8mm, or 0.67° visual angle (for a user seated 65cm away from the screen). While the area in which information can be perceived is wider than this single focal point (typically, 2° [28]), this approach ensures that participants saw at least the identified suggestion with high probability. The observed fixation times on QAC should be regarded as conservative estimates of attention to QAC.

4.2 Statistical Analysis

We analyzed the collected data using generalized linear mixed-effects models (GLMMs) [3, 23]. Linear mixed-effects distinguish between *fixed effects*, which are due to the experimental condition, and *random effects*, that are due to the variation within a random sample of e.g., participants, but do not generalize to the whole population. *Generalized* linear mixed-effects models generalize beyond linear response variables to, e.g., binary responses. In contrast to a traditional ANOVA analysis, GLMMs do not require a balanced experiment design, can handle missing data, and can simultaneously model crossed random effects due to both participant and task (traditional (repeated measures) ANOVA would only handle nested effects). GLMMs have become a popular analysis tool, due to recent advances in fitting these models [3]. In the context of information retrieval, they have been proposed as a tool to model topic effects in retrieval system evaluation [8].

We train a GLMM with crossed random effects for each response variable (detailed in Section 4.3). There is one fixed effect, QAC condition (2 levels), and two random effects, participant and task. Modeling participant and task as random effects is justified, because the random effects assumptions holds (they are independent of condition due to our experiment design). This results in the following set of models:

$$g(y_{ij}) = \beta_0 + \beta_1 x_{ij} + p_i u_i + t_j v_j + \epsilon_{ij}. \qquad (1)$$

Here, y_{ij} is the response observed for participant i and task j, and $g(\dot{)}$ is a link function, depending on the type of the variable (e.g., *logit* for binary variables). Further, β_0 and β_1 model the fixed effect due to condition, where β_0 is the intercept associated with the base level (here: condition = *original*) and β_1 defines the slope and is associated with our treatment condition = *random*; x_{ij} is a binary indicator of whether the observation for participant i and task j was in the treatment condition; p_i and t_j are the coefficients for participant and task effects, and u_i and v_j are indicators that identify participant i and task j. Finally, ϵ_{ij} is the residual noise associated with each observation.

Table 3: Overview of the response variables in our analysis.

QAC examination	
CFT	*Cumulative fixation time on QAC* – the total amount time during which fixations were recorded on any QAC during query formulation.
TFF	*Time to first fixation* – the time between the first keystroke and the first recorded fixation on any QAC suggestion, measured on the first query for a task.
Query formulation	
QFT	*Query formulation time* – the time between the first keystroke and query submission, measured for the first query formulated for a task.
QU	*QAC suggestion used* – binary, indicates whether the participant used QAC for their first submitted query (by mouse click, keyboard, within or at the end of a query).
QR	*QAC rank* – the rank of the QAC suggestion used (zero when QAC is not used).
QL	*Query length* – the length of the first submitted query, in characters.
CS	*Characters saved* – the difference between the length of the submitted query, and the number of characters that the user typed (zero when QAC is not used).
Task completion	
UQ	*Unique queries submitted* – the number of unique queries submitted while completing a task (matched by query text).
UR	*Unique result pages* – the number of unique result pages visited while completing a task (matched by URL).
TFC	*Time to first result click* – the time between the first query submission and the first click on a search result.
TCT	*Task completion time* – the time between being directed to the search page to indicating the end of the task using the keyboard (cut off at threshold ts).

In addition to the model used here, we tested more complex models, where participant and task could also affect the slope, in addition to the intercept. However, the more complex models did not improve model fit (measured by the Akaike information criterion, AIC), and coefficients were highly correlated, indicating an over-specified model. We chose to train separate models for each response variable, to keep the number of fitted parameters per model manageable, and to keep the models interpretable.

The type of each response variable was determined as follows. First, we selected the appropriate type based on the quantity that each response measures (logit for binary responses, Poisson for counts, and log for times). We compared the resulting models to a baseline linear response model, and found that the chosen response types improved model fit. Finally, we visually interpreted the residual plots to verify that the residuals were approximately normally distributed.

We fitted models using the `R` package `lme4` [5]. Binary and Poisson responses were fitted using the method `glmer`. For time-based data, we applied a log transform to the observed response [2] before fitting a linear response model using `lmer`.

To test for significant effects of condition, we use the `R` package `lmerTest` [22], which implements ANOVA for mixed-effects models using the Satterthwaite approximation to estimate degrees of freedom. We chose an alpha of 0.05 to determine whether to reject a null hypothesis.

4.3 Response Variables

The primary goal of our study is to understand the effect of QAC ranking quality on users' search behavior. We capture user behavior

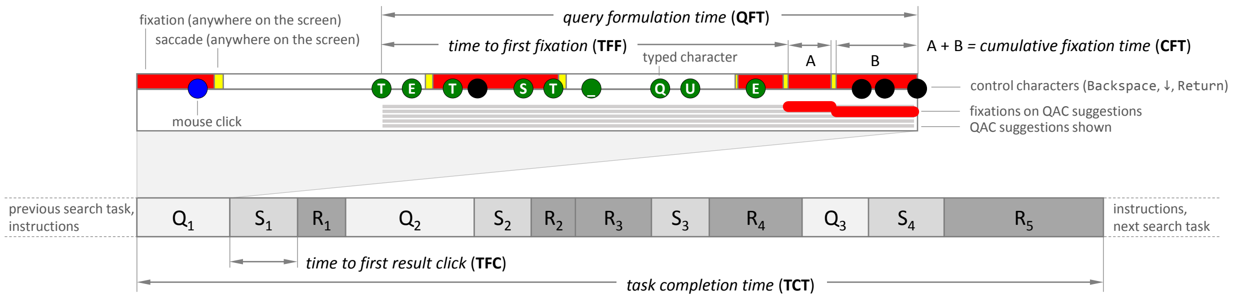
fixation (anywhere on the screen)
saccade (anywhere on the screen)
query formulation time (QFT)
time to first fixation (TFF)
typed character
A
B
A + B = cumulative fixation time (CFT)
T
E
T
S
T
_
Q
U
E
control characters (Backspace, ↓, Return)
mouse click
fixations on QAC suggestions
QAC suggestions shown
previous search task, instructions
Q1
S1
R1
Q2
S2
R2
R3
S3
R4
Q3
S4
R5
instructions, next search task
time to first result click (TFC)
task completion time (TCT)

Figure 2: Overview of a search episode, highlighting the time-based response variables used in this study. Shown is a search episode with 3 query formulation episodes $Q_{1\ldots3}$, examination of 4 search result pages $S_{1\ldots4}$, and 5 result page visits $R_{1\ldots5}$. For the first submitted query, the users' eye gaze and interactions are sketched. Here: the user submits a query using QAC.

in the *QAC examination*, *query formulation* and *task completion* response variables detailed below. An overview of the key definitions is given in Table 3. Time-based metrics are additionally illustrated in Figure 2 (all are measured in milliseconds).

To reduce variance, and in particular to avoid any potential interaction between search result quality and QAC engagement, we limit our analysis to the first query formulated per search episode. This also ensures that all our search tasks have equal influence on the results, rather than tasks requiring more queries dominating.

QAC examination. These measures reflect how users examine QAC suggestions. In particular, TFF is not expected to be affected by QAC ranking condition, and serves to validate our experimental setup (users can only be affected by QAC ranking condition after examining suggestions). CFT captures how much time users spend examining suggestions. When QAC ranking quality degrades, they may have to spend more time identifying the best suggestion, or may give up examination sooner, if the top-ranked suggestions do not appear promising.

Query formulation. This group of measures is designed to capture the behavior of the searcher when formulating a query. If QAC ranking primarily affects how searchers formulate queries, then we would expect effects on these measures. For example, if searchers seek the best suggestion to submit, QFT would be expected to increase in the random condition. QU may be affected if searchers ignore lower-ranked suggestions, and instead type the query they have in mind, while CS may be affected if users select top-ranked suggestions, regardless of ranking quality.

Our query formulation response variables capture user behavior as follows: QFT captures the intuition that higher quality QAC rankings may speed up query submission; QU reflects the hypothesis that higher-quality QAC rankings may lead to higher suggestion usage, following [20]; QR reflects the use of suggestion rank for QAC ranking evaluation [4, 31]; QL and CS test whether submitted queries differ qualitatively between ranking conditions.

Task completion. The last group of measures is designed to capture the behavior of the searcher on the overall task. Because condition is consistent within a task, it may affect how efficiently search tasks are completed. In particular, we would expect effects on these variables if searchers follow a "satisficing strategy", where they select query suggestions they encounter as long as they reasonably approximate their search intent.

This set of response variables reflects task completion as follows: UQ captures the intuition that users may submit relatively fewer queries when QAC ranking quality is degraded and query submission becomes more difficult [1]. UR is a proxy for query quality, as lower-quality queries may lead to the user needing to visit more result pages. In both cases, we count *unique* items to avoid effects that are due to e.g., back button use. TFC similarly measures the time it takes from submitting the first query to selecting the first result page. Again, this measure may be affected by the quality of the submitted query. Finally, TCT reflects whether higher quality QAC suggestions can lead to faster task completion [17].

5. RESULTS

We now present our main results on the effect of QAC ranking quality condition on user interactions (Table 4). We examine the effects of ranking condition on our three groups of response variables: QAC examination (Section 5.1), query formulation (Section 5.2), and task completion (Section 5.3).

331 query formulation episodes were collected (see Section 4.1), of which 169 (51.1%) were completed in the *original* QAC ranking condition and 162 (48.9%) in the *random* condition. On average, we observed 23.6 participants perform each task (minimum 21, median 24, and maximum 25). An average of 13.2 search episodes were obtained per participant (min 5, median 14, max 14). Unless otherwise noted, all analysis of statistical significance was performed using GLMMs (Section 4.2) trained on this data.

5.1 Effect of ranking on QAC examination

QAC suggestions were shown during part of all query formulation sessions, at least for short prefixes. On average, QAC was shown throughout 86% of the query formulation episodes (median: 95.4%). Most users examine suggestions, with the probability of fixating on suggestions at some point during query formulation estimated as $logit^{-1}(\beta_0) = logit^{-1}(3.468) = 0.97$ under the base model (**CTF** > 0). Ranking condition does not have a statistically significant effect here ($\beta_1 = -0.22$, $p = 0.604$). When suggestions are examined, the amount of time spent doing so is substantial. It is estimated as $exp(\beta_0) = exp(7.124) = 1,241$ ms under the base model (**CFT | CFT** > 0, $\beta_0 = 7.124$). Again, there is no evidence that QAC condition affects examination behavior ($\beta_1 = -0.043$, $p = 0.685$). To verify that our treatment had no unintended side-effects, we also examine **TFF**, which should not depend on ranking order. Participants are estimated to engage with suggestions after 667 ms under the base model ($\beta_0 = 6.503$). No evidence of an effect of condition is found, which validates our treatment.

Interestingly, while participants spent a substantial amount of time examining QAC suggestions, we find no evidence of adapting QAC examination to ranking condition. If searchers were to seek out the best possible suggestion, they would be expected to consider lower ranks in more detail under randomization (assuming that the best suggestions are shown towards the top in the original ranking). To further investigate this phenomenon, we split the time searchers

Table 4: Results: coefficients for the (generalized) linear mixed-effects models, with standard error (SE) and p-value, for each response. Vertical bars, as in "a | b" indicate that a was analyzed for a subset of data, given that b is true. For example, "CS | QU" means that CS (characters saved) was analyzed for the data where QU (QAC Use) is true (n indicates the resulting number of samples the analysis was run on). Statistically significant effects are marked with "*".

				Intercept			Condition = *random*		
	Response	Type	n	β_0	*SE*	*p*	β_1	*SE*	*p*
QAC examination	**CFT** > 0	binary	331	3.468	0.775	$\ll 0.001$*	-0.220	0.424	0.604
	CFT | **CFT** > 0	log	284	7.124	0.189	$\ll 0.001$*	-0.043	0.106	0.685
	TFF | **CFT** > 0	log	284	6.503	0.217	$\ll 0.001$*	-0.094	0.204	0.643
Query formulation	**QFT**	log	331	8.680	0.104	$\ll 0.001$*	0.058	0.053	0.272
	QL	Poisson	331	3.224	0.064	$\ll 0.001$*	-0.007	0.022	0.737
	QU	binary	331	-0.915	0.401	0.026*	-0.508	0.285	0.075
	CS | **QU**	Poisson	99	2.192	0.110	$\ll 0.001$*	0.223	0.070	0.001*
	QR | **QU**	Poisson	99	0.344	2.731	$\ll 0.006$*	0.044	0.177	0.802
Task completion	**UQ**	Poisson	331	0.357	0.076	$\ll 0.001$*	0.044	0.091	0.631
	UR $= 0$	binary	331	-3.654	1.082	$\ll 0.001$*	-0.022	0.402	0.956
	UR | **UR** > 0	Poisson	282	0.703	0.105	$\ll 0.001$*	0.161	0.078	0.040*
	TFC | **UR** > 0	log	282	8.625	0.109	$\ll 0.001$*	-0.036	0.083	0.670
	TCT $\geq ts$	binary	331	-3.217	0.531	$\ll 0.001$*	0.764	0.416	0.066
	TCT | **TCT** $< ts$	log	297	11.096	0.109	$\ll 0.001$*	-0.021	0.046	0.650

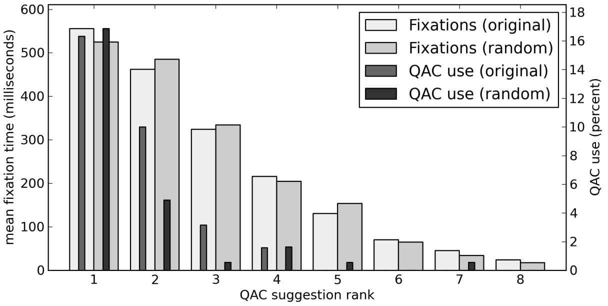

Figure 3: Mean fixation time and QAC use by suggestion rank.

spent examining suggestions by condition and suggestion rank. Figure 3 shows a strong bias towards examining top-ranked suggestion, which is consistent across conditions. In both conditions, we see a clear decrease in attention with rank (as measured by fixation time). Under the original ranking, users spend on average 555 ms examining the first suggestion, 14% more than for the second suggestion. This decreases to only 24 ms for the lowest-ranked suggestion. Under the random condition, users spend on average 524 ms examining the top suggestion, just 5.6% less than in the original condition. The difference in fixation time on the first suggestion is distributed relatively evenly over the remaining suggestions, but much less so than would be expected given the aggressive uniform randomization.

Our findings suggest that searchers follow consistent patterns when examining QAC suggestions, rather than seeking out the best suggestion, or a query they had in mind. We now investigate whether and how the identified position bias affects query formulation.

5.2 Effect of ranking on query formulation

We find no evidence of an effect of QAC ranking condition on **QFT**. QFT under the base model is estimated as $5,884$ ms ($\beta_0 = 8.680$), which is estimated to increase to $6,235$ ms under the random condition (not significant with $\beta_1 = 0.058$, $p = 0.272$). Comparing the base model estimates to our observation on query examination, we see that searchers tend to examine suggestions early (after roughly 11% of QFT has passed), and that a substantial amount of the QFT is spent examining suggestions (more than 20%). We find no evidence that **QL** is affected by condition, with an estimated length of 25 characters under the base model ($\beta_0 = 3.224$).

Turning to QAC usage, we find that the probability of engaging with suggestions is estimated as 0.29 (**QU**, $\beta_0 = -0.915$). This decreases to 0.19 under the random condition, i.e., participants where 32% less likely to engage with suggestions. While this effect is not statistically significant ($p = 0.075$), it does provide some evidence that QAC ranking quality may affect interaction with suggestions. We examine this phenomenon in more detail, by focusing on characters saved and on the selected suggestion rank when QAC is used. We find a strong and statistically significant effect on **CS**. Under the base model, participants are estimated to save 8.95 characters per query ($\beta_0 = 2.192$). This increases to 11.19 characters under the random condition ($\beta_1 = 0.223$, $p = 0.001$). We verified that this increase in CS is not due to a change in query length. Rather, it is a result of a marked drop in engagement for shorter queries (up to 20 characters long), that tend to be placed higher in the *original* condition. This suggests that QAC ranking condition affects for *which* queries suggestions are used.

Under the base model, **QR** is estimated as 1.41 ($\beta_0 = 0.344$). We find no evidence of an effect of ranking condition on the ranks of selected suggestions ($\beta_1 = 0.044$, $p = 0.802$). This is further confirmed when considering QAC use per rank under the two conditions in Figure 3. Similar to the position bias observed previously for QAC examination, we see a strong bias in QAC use towards the top-ranked suggestions. Under the original condition, participants selected the top-ranked suggestion in 16.3% of all query formulation episodes. Under the random condition, we see a slight increase, to 16.9%. While this small small shift of usage towards the top-ranked results may be due to noise, it is consistent with findings on user interactions with search result pages, where users increasingly trusted the top-ranked results when the best result was made harder to identify [16].

QAC examination and the ranks of suggestions that users engage with do not appear to change with QAC ranking. As a result, we see a difference in the suggestions used. Here, we see a marked increase in characters saved under the random condition, indicating a strong

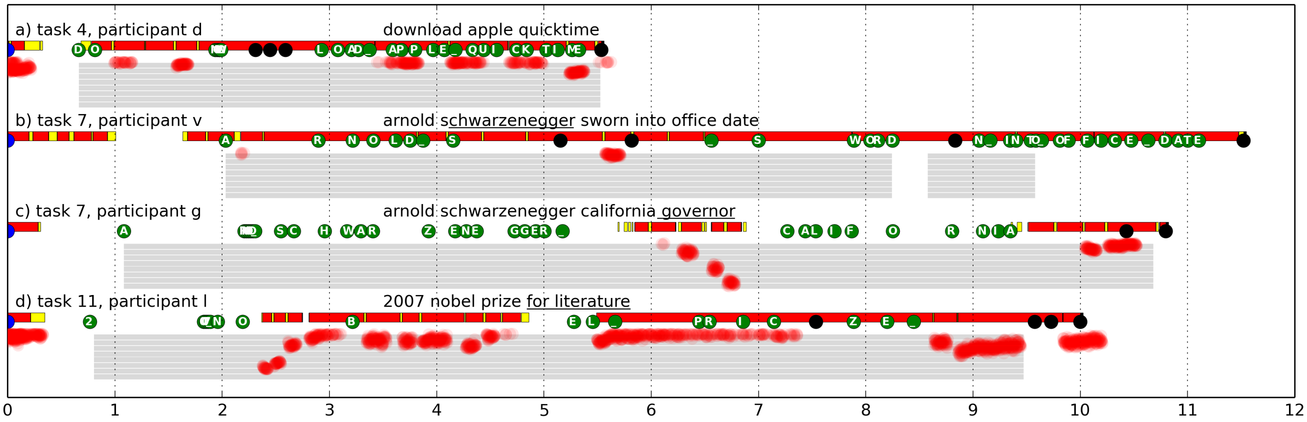

Figure 4: Visualization of example query formulation sessions for different study participants on tasks 4 (a), 7 (b-c), and 11 (d). The x-axis shows time in seconds (see Figure 2 for details of the visualization). We identify *monitoring* behavior in (a), *ignoring* in (b) and (c), and *searching* in (c) and (d). The final query is shown for each example, with characters submitted using QAC underlined.

effect on the specific suggestions that study participants engaged with. The question that remains is whether this affects the quality of submitted queries. We investigate this question by comparing user behavior as captured by task completion metrics.

5.3 Effect of ranking on task completion

When query formulation becomes more difficult, participants may compensate by issuing fewer queries, or by examining more search result pages [1]. Here, we find no evidence that the number of unique queries is affected by ranking condition (**UQ**, $\beta_1 = 0.044$, $p = 0.631$).

To analyze the number of result pages visited, we first consider the probability of not visiting any result pages (**UR** $= 0$). This occurs when the solution for a search task is found directly on the search result page. We find that this happens with low probability (less than 0.03) and is not affected by condition ($\beta_1 = -0.022$, $p = 0.956$). However, we do find a substantial and statistically significant effect of QAC ranking condition on the number of visited result pages when any are visited. Under the base model, **UR | UR** > 0 is estimated at 2.02 ($\beta_0 = 0.703$). This increases to 2.37 under the random condition ($\beta_1 = 0.161$, $p = 0.04$), a relative increase of 17%. The time to make the first click decision (**TFC**) is not affected by condition ($\beta_1 = -0.036$, $p = 0.670$).

Regarding the overall task completion time, we find some evidence that tasks were more likely to time out under the random condition. In our study, we set the timeout threshold $ts = (3 * 60 * 1000) - 2000$ milliseconds to account for small inaccuracies in system timing (this threshold resulted in the best fit, but results were qualitatively similar for other choices). The probability of a timeout was 0.039 under the base model ($\beta_0 = -3.217$). This increased to 0.079 under the random condition ($\beta_1 = 0.764$, $p = 0.066$). When tasks were completed within the alloted time, no effect of condition on **TCT** was observed ($\beta_1 = -0.021$, $p = 0.650$).

Our results regarding task completion indicate that QAC ranking quality can affect task completion. The significant effect on UR, and the additional evidence that timeouts may be more likely under the random condition, suggest that lower-quality QAC rankings can impede task completion.

6. DISCUSSION

In Section 5, we identified a strong position bias in QAC examination. Participants focused on examining top-ranked suggestions, independently of QAC ranking condition. Ranking condition was found to affect which queries were submitted with QAC support. It also affected task completion. Under the *random* condition, the number of unique result pages users visited to solve search tasks increased, suggesting an effect on query effectiveness.

In this section we discuss our results and their implications in more detail. First, we detail behavior patterns that characterize query formulation behavior in Section 6.1. Then we discuss the implications of our findings on QAC evaluation in Section 6.2, and our approach to minimize potential limitations of our study setup in Section 6.3.

6.1 Common behavior patterns

We identified behavior patterns by visually analyzing all collected query formulation episodes. Together, these provide insights into the kinds of patterns that can generate the observed effects of QAC ranking on search behavior. We exemplify the identified behavior patterns in the four examples shown in Figure 4.

A behavior that strongly affects query formulation is whether the searcher touch-types. In our data set we identified touch-typing by the presence of gaze fixations being detected on the screen during typing (as indicated by the red bar in Figures 4 (a) and (b), but absent in Figure 4 (c)). We found that 12 study participants (48%) primarily touch-typed. The remaining participants primarily looked at the keyboard while typing. Occasionally, participants typed the beginning of the query while looking at the keyboard (often the first 2-3 characters of the query), and completed typing while looking at the screen (eg., Figure 4 (d)).

Typing behavior was strongly associated with the times at which participants noticed QAC. Touch-typists tended to *monitor* suggestions, either continuously, or in intervals (Figure 4 (a)). Monitoring is characterized by frequent fixations on QAC, and on the top-ranked suggestions in particular. We hypothesize that in the monitoring case, QAC is playing a role in confirming to the user that they are typing the query correctly.

While most monitoring involved the user considering just the top suggestion, we noticed many examples (eg., Figure 4 (c)), where the participant scanned the QAC list from top to bottom. We term this *searching* behavior. In this behavior pattern, participants actively scan, and engage with, the QAC list. The examples in Figure 4 (b) and (c) illustrate two distinct types of searching. In example (b), the searcher is seeking spelling support by selecting the top suggestion for the second query word (schwarzenegger) while entering a query they have in mind. In example (c), the searcher is not seeking spelling support, but rather looks for a complete query that appropriately expresses their information need. Interestingly, while the direction of scanning suggestions was typically from top to bottom, we found a substantial number of examples where this pattern was reversed. This pattern appears to result from the user looking up from the keyboard (eg., to verify their typed query), and scanning query suggestions along the way. This was observed at least once for 12 of the study participants (48%).

Four participants (16%), who primarily looked at the keyboard while typing, largely *ignored* QAC suggestions. They typically only looked up from the keyboard when they had finished typing.

With the broad patterns identified above, we can characterize searchers as follows. Touch-typists tend to monitor suggestions as they type. They primarily focus on the top suggestions and use them when convenient. They search for spelling support or query completions occasionally, but not as much as searchers who do not touch type. This second group has more frequent and clearly defined intervals at which they examine suggestions, while few completely ignore suggestions.

6.2 Implications for evaluation

The strong position bias observed in QAC use can make it difficult to devise a general evaluation strategy for QAC rankings. In this section, we revisit the evaluation strategies that have been proposed in previous work, and observe how they relate to our findings.

QAC evaluation is commonly based on log data, where an observed query [4, 20, 31], or set of queries [32, 33] is taken as the ground truth. Several authors further propose the use of rank-based evaluation metrics, such as MRR to assess the quality of suggestion rankings [4, 31–33]. In these metrics, it is assumed that lower-ranked results are less likely to be examined by searchers, resulting in lower utility for suggestions placed at lower ranks. This effect is confirmed by our results on the distribution of gaze across ranks. Treating suggestions at all ranks identically, as proposed in [24], is not supported.

While our findings support the *training* of QAC rankers to move useful suggestions to top ranks, selected ranks may not be an adequate basis for evaluating QAC rankings under live user behavior. While higher-ranked suggestions receive more attention, we find no evidence of searchers adapting their examination strategy when rankings degrade. As a result, the ranks of the suggestions that users engage with may not be affected as expected. The use of differences in QAC usage, as proposed by Kharitonov et al. [20] is a promising direction for QAC evaluation.

A drawback of log studies is that logs that were collected while an existing QAC system is running may suffer from position bias, as valuable suggestions that are not currently highly ranked may receive little attention. As an alternative to using log-based ground truth data, Bhatia et al. [6] relied on manual judgments for each suggestion. Such an approach can more objectively measure suggestion quality, but may fail to capture context-dependent aspects of suggestion utility. A possible extension may be to combine manual judgments with log information that captures search intents as expressed by the landing pages users visited after issuing a query. Liu et al. [24] followed a more indirect log-based approach, where the quality of a result page retrieved by a query suggestion was measured in terms of NDCG. Our results indicated a significant effect of QAC ranking quality on the number of results visited, indicating that approaches measuring result page quality are promising. Finally, the use of targeted exploration of alternative rankings may be a promising avenue to collecting unbiased data for evaluation.

Effort-based metrics for QAC evaluation were proposed for spelling support [11]. In our study, we did not find evidence of searchers explicitly trading off effort in our more general setting.

Probably due to the reliance on log data for evaluation, experiments so far have mainly focused on the "input support" scenario, where users are assumed to benefit most from saving time and keystrokes when entering a query. Our study identified a variety of motivations. Spelling and query formulation support are two usage scenarios that can be addressed directly. For spelling support, it may be easier to predict an individual, difficult to type, query term instead of a complete query. For query formulation support, it may be possible to guide users towards expressive queries. Comprehensive QAC evaluation setups should take these scenarios into account.

6.3 Potential limitations

It is important to note a number of potential limitations in our study, and how we designed our study specifically to minimize their effect. Below, we address the influence of study participant and task selection, the study environment, and eye-tracking setup.

The participants and search tasks in a user study are typically not randomly sampled from an overall population. Nevertheless, eye tracking user studies in web search have generally been successful in identifying patterns of user behavior that proved meaningful for characterizing more general search behavior. Here, we made a specific effort to recruit participants from a wide variety of ages and backgrounds resulting in a variety of behaviors. A fixed set of closed search tasks was used, to afford the necessary control needed to detect relationships between our variables of interest. To alleviate the effect of using predefined search tasks, we selected a variety of search tasks to represent different typical types of interaction. In addition, the analysis method we used (mixed-effects models with crossed random effects for participants and tasks) is specifically designed for modeling variation across participants and tasks as random effects, while finding a generalizable model of the fixed effect (condition) [3].

The fact of being part of a user study may affect behavior, as users who are being watched may behave differently when searching than they do in a personal context. To limit this effect, our user study was performed in an office environment, configured to be maximally representative of an actual office. We chose an eye tracker that is minimally invasive. In the follow-up questions, our participants indicated that they believed that they behaved naturally, and none were aware that the focus of our study was the use of QAC.

Finally, eye tracking in a natural environment is often prone to error when detecting the user's gaze locations. In laying out our study location, we attempted to minimize potential sources of error (such as direct sunlight), performed per-participant calibration, and minimized typical further sources of error (for example, the participant's chair did not have wheels to minimize motion). Moreover, both experimental conditions studied would be subject to the same sources of error.

7. CONCLUSION

In this paper, we presented the first eye-tracking study of query auto completion (QAC) ranking quality on searchers' examination of, and interactions with QAC. During our study, participants saw suggestions in the original order generated by a commercial search engine for half the search tasks they completed (assigned at random). For the other half of the tasks, searchers saw a random permutation of suggestions.

Across ranking conditions, we found a strong bias towards examining and using top-ranked suggestions, a finding that is surprisingly consistent with similar observations on user interactions with web search results. While earlier log studies reported position bias, our study is the first to demonstrate that this effect is due to examination bias, not ranking quality.

Due to position bias, ranking condition affected QAC usage both in terms of engagement and QAC effectiveness. We found effects on task completion, in particular the number of result pages participants visited to complete a task. This second effect suggests that queries issued with lower-quality QAC rankings may be less effective. The observed effects can be seen as a result of several behavioral patterns, in particular *monitoring*, *ignoring*, and *searching* for QAC suggestions. Finally, we discussed the implications of these findings on QAC evaluation.

In this study, we focused on closed information seeking tasks, and on short-term effects of QAC ranking quality. Interesting directions for future research include extending this work to more open-ended search tasks, and to long-term effects of changes in QAC ranking quality. In addition, we plan to revisit evaluation metrics based on our findings. Finally, identifying and predicting types of user behavior, such as seeking spelling support or query formulation support during query formulation would allow QAC systems to better support searchers.

Acknowledgements. We would like to thank the Bing AutoSuggest team for their support throughout this project. In particular, Panagiotis (Panos) Tigkas and Istvan Simon provided technical support and generously shared their ideas in many lively discussions. We are grateful to Chris O'Prey, Ken Woodberry, and Mitch Goldberg, for sharing their eye tracking equipment and expertise, and Natasa Milic-Frayling, Siân Lindley, and Gabriella Kazai for their help in designing the user study. We thank Fernando Diaz, Susan Dumais, and Thore Graepel for insightful feedback, questions, and suggestions.

References

[1] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *SIGIR '13*, pages 23–32, 2013.

[2] R. H. Baayen and P. Milin. Analyzing reaction times. *Intl. Journal of Psychological Research*, 3(2):12–28, 2010.

[3] R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.

[4] Z. Bar-Yossef and N. Kraus. Context-sensitive query auto-completion. In *WWW '11*, pages 107–116, 2011.

[5] D. Bates. Fitting linear mixed models in R. *R news*, 5(1):27–30, 2005.

[6] S. Bhatia, D. Majumdar, and P. Mitra. Query suggestions in the absence of query logs. In *SIGIR '11*, pages 795–804, 2011.

[7] G. Buscher, S. T. Dumais, and E. Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proc. SIGIR*, pages 42–49, 2010.

[8] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *CIKM '11*, pages 611–620, 2011.

[9] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *WWW '09*, pages 1–10, 2009.

[10] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *CHI '07*, pages 407–416, 2007.

[11] H. Duan and B.-J. P. Hsu. Online spelling correction for query completion. In *WWW '11*, pages 117–126, 2011.

[12] S. Dumais, R. Jeffries, D. M. Russell, D. Tang, and J. Teevan. Understanding user behavior through log data and analysis. In J. S. Olson and W. A. Kellogg, editors, *Ways of Knowing in HCI*, pages 349–372. Springer New York, 2014.

[13] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky. Eye tracking in web search tasks: Design implications. In *ETRA '02*, pages 51–58, 2002.

[14] K. Grabski and T. Scheffer. Sentence completion. In *SIGIR '04*, pages 433–439, 2004.

[15] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *SIGIR '04*, pages 478–479, 2004.

[16] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *CHI '07*, pages 417–420, 2007.

[17] D. Hawking and K. Griffiths. An enterprise search paradigm based on extended query auto-completion. do we still need search and navigation? In *ADCS '13*, 2013.

[18] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05*, pages 154–161, 2005.

[19] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3:1–224, 2009.

[20] E. Kharitonov, C. MacDonald, P. Serdyukov, and I. Ounis. User Model-based Metrics for Offline Query Suggestion Evaluation. In *SIGIR '13*, 2013.

[21] O. Komogortsev, D. Gobert, S. Jayarathna, D. H. Koh, and S. Gowda. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Trans. Biomedical Engineering*, 57(11):2635–2645, 2010.

[22] A. Kuznetsova, R. Christensen, and P. Brockhoff. lmertest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). *R package version 2.0-6*, 2014.

[23] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[24] Y. Liu, R. Song, Y. Chen, J.-Y. Nie, and J.-R. Wen. Adaptive query suggestion for difficult queries. In *SIGIR '12*, pages 15–24, 2012.

[25] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan. Eye tracking and online search: Lessons learned and challenges ahead. *JASIS&T*, 59(7):1041–1052, 2008.

[26] B. Mitra, M. Shokouhi, F. Radlinski, and K. Hofmann. On user interactions with query auto-completion. In *SIGIR '14*, pages 1055–1058, 2014.

[27] A. Olsen and R. Matos. Identifying parameter values for an i-vt fixation filter suitable for handling data sampled with various sampling frequencies. In *Symp. Eye Tracking Research and Applications*, pages 317–320, 2012.

[28] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3): 372–422, 1998.

[29] J. Salojärvi, I. Kojo, J. Simola, and S. Kaski. Can relevance be inferred from eye movements in information retrieval? In *WSOM'03*, pages 261–266, 2003.

[30] C. Shah, J. Liu, R. González-Ibáñez, and N. Belkin. Exploration of dynamic query suggestions and dynamic search results for their effects on search behaviors. *ASIST*, 49(1):1–10, 2012.

[31] M. Shokouhi. Learning to personalize query auto-completion. In *Proc. SIGIR*, pages 103–112, 2013.

[32] M. Shokouhi and K. Radinsky. Time-sensitive query auto-completion. In *SIGIR '12*, pages 601–610, 2012.

[33] A. Strizhevskaya, A. Baytin, I. Galinskaya, and P. Serdyukov. Actualization of query suggestions using query logs. In *WWW '12*, pages 611–612, 2012.

[34] E. Toms, H. O'Brien, T. Mackenzie, C. Jordan, L. Freund, S. Toze, E. Dawe, and A. MacNutt. Task effects on interactive search: The query factor. In *INEX'08*, pages 359–372, 2008.

[35] I. Weber and C. Castillo. The demographics of web search. In *SIGIR '10*, pages 523–530, 2010.

[36] R. W. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.*, 43:685–704, 2007.