# Providing Direct Answers in Search Results: A Study of User Behavior

Zhijing Wu[†], Mark Sanderson[‡], B. Barla Cambazoglu[‡], W. Bruce Croft[‡], Falk Scholer[‡]
† Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
‡ Computer Science, School of Science, RMIT University, Melbourne VIC 3000, Australia
wuzhijing.joyce@gmail.com,{mark.sanderson,barla.cambazoglu,bruce.croft,falk.scholer}@rmit.edu.au

## ABSTRACT

To study the impact of providing direct answers in search results on user behavior, we conducted a controlled user study to analyze factors including reading time, eye-tracked attention, and the influence of the quality of answer module content. We also studied a more advanced answer interface, where multiple answers are shown on the search engine results page (SERP). Our results show that users focus more extensively than normal on the top items in the result list when answers are provided. The existence of the answer module helps to improve user engagement on SERPs, reduces user effort, and promotes user satisfaction during the search process. Furthermore, we investigate how the question type – factoid or non-factoid – affects user interaction patterns. This work provides insight into the design of SERPs that includes direct answers to queries, including when answers should be shown.

## CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; *Web interfaces*; *Search interfaces.*

## KEYWORDS

answer module, question answering, user interaction, web search, search behaviour, controlled user study, eye-tracking

**ACM Reference Format:**
Zhijing Wu[†], Mark Sanderson[‡], B. Barla Cambazoglu[‡], W. Bruce Croft[‡], Falk Scholer[‡]. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3412017

## 1 INTRODUCTION

Commercial web search engines increasingly feature *answer modules* as part of their Search Engine Result Page (SERP). The module, which usually appears at the top of the SERP, provides a concise answer to certain types of informational queries issued by users (Figure 1 shows an example). Like the commonly displayed knowledge panel, the answer module provides information directly within the SERP. Hopefully the module increases user engagement with the SERP and reduces the time that a user needs to spend inspecting remote web resources linked from the SERP.

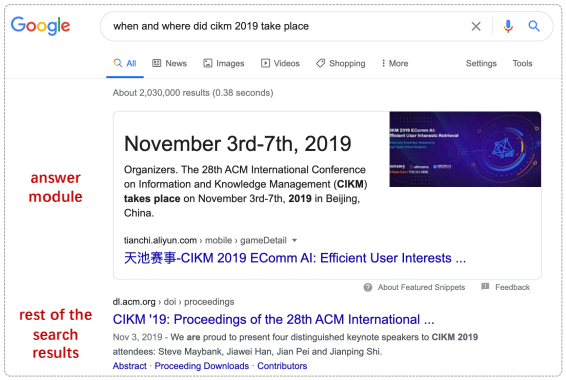


**Figure 1: The answer module displayed as a featured snippet in a SERP.**

The growing popularity of the answer module in modern web search is due to the ongoing shift from short bag-of-word queries to longer, well-formed, question-like queries [39], as well as the advance of deep learning techniques that enable better understanding and handling of such queries. Most research dealing with answers to web search questions has had a system-centric point of view and focused on issues such as the extraction of answers from various forms of text content, the ranking and triggering of answers, and the generation of answers in natural language [32, 41]. With respect to user-centric research, there are several works that looked into the usability of modules such as health cards [13] in a SERP. However, there has been no previous work that adopted a user-centric approach and aimed to understand the way users interact with answers presented in a SERP.

Our work attempts to give insight into the impact of the answer module on users' overall search experience in information-seeking search tasks, giving rise to the following research questions:

**RQ1:** How do users interact with a search engine when a direct answer to the query is provided on the SERP, and how does the quality of the answer affect these interactions?

**RQ2:** How do interaction patterns change when a more advanced multi-answer module is presented?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*CIKM '20, October 19–23, 2020, Virtual Event, Ireland*
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412017

Users interact with search engines in different ways for different types of queries [30]. Some types are associated with exploratory, browsing-style user behavior, while for other queries users exhibit a more focused search. This motivates our third research question:

**RQ3:** How does question type (factoid versus complex) affect interaction patterns during answer seeking tasks?

To understand the impact of the answer module on users' search experiences, we carried out a controlled lab-based user study in which participants interacted with SERPs in different settings (a SERP with no answer, a traditional single answer, or a browsable multi-answer module). We gathered self-reported feedback from study participants through questionnaires, and tracked and recorded user interaction signals including mouse clicks, eye and cursor movements. Our key contributions are:

- We conduct and eye-tracking experiment that included 600 search sessions to investigate users' interactions during answer seeking tasks.
- We provide a thorough analysis of how users interact with a SERP when an answer module is shown. To the best of our knowledge, this is the first published work to focus on users' interaction patterns on such an interface, although some commercial web search engines have started using answer modules.
- We also show the effect of a more advanced answer module, and question type, on users' interaction patterns.

In the rest of the paper, Section 2 provides a survey of key related work. The details of the lab study are presented in Section 3. We provide an analysis of the collected data in Section 4. Conclusions and avenues for future work are discussed in Section 5.

## 2 RELATED WORK

Our review is grounded in related work on SERP and individual result item design, and eye-tracking analysis.

### 2.1 Design of Search Engine Result Pages

A range of studies have investigated aspects of the design of SERPs, including what should be included as a result item, how individual items should be formatted, and the layout and design of the result page overall. The most traditional SERP design, popularised by early web search engines, is often referred to as "10 blue links". Here, information about the top 10 documents that are identified as being relevant to a query are shown in a list, ordered by decreasing likelihood of relevance. Each *result item* in the list represents a document, which typically includes: a clickable title to open the underlying document; the URL of the document; and a short text summary of the document content (often also referred to as a snippet). Tombros and Sanderson [38] established that query-biased summaries (their name for snippets) – where sentences that most closely match the terms of a user's query are extracted from the underlying document and displayed – can improve a searcher's speed and correctness in identifying the relevance of an underlying resource compared to static summaries that use the opening sentences of a document; this approach continues to be used in modern search engines. The length of summaries was studied by Cutrell and Guan [7], who demonstrated that providing more information is beneficial for informational searches, but can be harmful for navigational searches. A broad study of other features of result items was carried out by Clarke et al. [4], showing that the presence of more query terms in the summary, the length of the shown URL, and the readability of the summary are key factors that influence web search behavior.

Studies of the linear ranked list-based layout of SERPs have identified possible sources of bias that may be introduced when searchers examine the results. Through experiments that manipulated the ordering of result items returned by a web search engine, Joachims et al. [14] demonstrated both trust bias (searchers are more likely to select answers that are presented higher in a ranked results list) and quality bias (selections are influenced by the overall quality of other items in the results). Subsequent work investigated the modelling of user click behavior to take these sources of position bias into account [6]. As an alternative to the typical linear list presentation, Kammerer and Gerjets [15] studied a grid-based layout and demonstrated that trust bias can be mitigated when users are unsure whether the retrieval system has attempted to rank the search results. Two-dimensional or matrix layouts are popular for presenting search results in particular areas such as image or online shopping search [3].

Direct alternatives to the typical text-based summaries that are displayed as result items on SERPs have also been proposed in the literature, including enhanced thumbnails [42] where web pages are represented by a small screenshot with additional highlighting of key information components. Visual snippets that include a salient image, logo, and document title were also examined [35]. Analysis indicated that such representations reduces user search time for some types of tasks. However, they are not widely used by modern web search services.

While the traditional query-biased summary representation of text-based documents continues to be widely used, SERPs have evolved to dynamically incorporate additional types of information. The results of vertical search, where a row of images or videos are directly embedded into the SERP [1, 40], are a key example. Others include advertisements, direct inline answers, knowledge panels, maps, recent news stories, tweets, and specific information such as weather data or exchange rates [24, 36].

### 2.2 Eye Tracking Analysis of Search Behavior

Tracking the movements of a person's eyes has a long history of application in the study of human attention, as people shift their eyes to focus on regions or objects of interest. To track a user's gaze on a screen, modern hardware typically employs a video-based corneal reflection approach to capture fixations (specific points at which the gaze is briefly maintained) and saccades (fast eye movements to the next point of fixation) [8]. Given the relationship between gaze position and attention, eye tracking has been used in a range of studies aiming to better understand user behavior when interacting with IR systems, and in particular with SERPs. Foundational work by Granka et al. [12] examined the relationship between eye gaze and click behavior and established that users on average read a SERP from top to bottom, and that the total time spent viewing a result decreases with rank, but the decrease is slower than the drop in click frequency. The result helped advance

the development of click models, since when a user clicks on an item, it can be inferred that they have also viewed those that are higher ranked.

Eye tracking analysis has been used to study patterns of behavior when searchers interact with retrieval systems. Aula et al. [2] were able to divide searchers into two groups, economic and exhaustive evaluators, based on their information seeking strategies. Economic evaluators proceeded to their next action (reformulating a query, or clicking on a document) by examining a smaller number of result list items compared to exhaustive evaluators. Cole et al. [5] studied gaze and user activity patterns as searchers proceeded through tasks of different complexity and types (including journalism-style activities, and searching for biomedical information), finding that differences in tasks can be inferred using representations derived from eye tracking data. Thomas et al. [37] also found differences in gaze patterns for tasks of differing complexity levels, and observed that user attention appears to move down a SERP in a band of attention, with gaze moving more freely between the individual result items within this band. Zheng et al. [44] used eye tracking analysis to develop a two-stage reading model, which is then used to improve the performance of an answer retrieval system.

Analysing viewing behavior using eye tracking has also enabled the detailed comparison of how the distribution of user attention changes when elements of the SERP are altered, including different versions of individual result list items, such as varying lengths of summaries [7], or including different types of result items, such as advertisements [9]. One key limitation regarding eye tracking analysis is that specialised and often expensive hardware is required; as an alternative, Lagun and Agichtein [18] developed an approach whereby a SERP is modified so that only one result item is clearly visible at a time while the rest of the page is blurred; the user can change the viewport using a mouse or trackpad as they read the results. A comparison with unrestricted viewing data gathered using eye tracking showed a high degree of similarity between the data obtained using the viewport approach.

## 3 USER STUDY

To investigate our research questions, we conducted a controlled lab-based study to collect data on user interaction with answers. The study was reviewed and approved by the RMIT University Human Research Ethics Committee.

### 3.1 Task Design

Since we focus on question-type queries in this work, we designed our tasks based on the publicly available MS MARCO [26] dataset, which contains about one million questions sampled from real, anonymized queries submitted to the Bing search engine. We manually selected 20 questions, composed of 10 factoid questions and 10 non-factoid questions, as our search tasks. Factoid questions can be solved with simple entity-based answers, while non-factoid questions require several answer sentences or passages [29]. Examples of the search questions include "what are the largest cities in Australia?" (factoid) and "what can cause a rash?" (non-factoid).

To study the impact of providing direct answers in SERPs on user behavior, we generated SERPs with or without an answer module for each search task. To gather search results, we submitted the 20 questions to the Google search engine in October 2019. Vertical and sponsored results were removed because they have been shown to affect user behavior on SERPs [22]. The top ten organic search results were retained for each question. We then asked three assessors to assign two-grade snippet usefulness labels for each of the top five search results. An inter-assessor agreement of 0.827 was obtained (using Fleiss' $\kappa$ [11]), indicating almost perfect agreement [19].[1] We used the majority vote across the three judgments as the final *usefulness score* of each item. There are 2.65 useful snippets out of the top 5 snippets on average.

To gather direct answer passages, we obtained the direct answers from the Passage Ranking MS MARCO dataset. For each question, this dataset includes passages selected by assessors to indicate if they were helpful for writing natural language answers to the question. From this dataset, we manually sampled useful passages (referred to as *useful answers* below), and useless passages (referred to as *useless answers* below) according to the provided usefulness labels. Useful answers can provide meaningful information for the search task, while useless answers are relevant to the topic of the search task but do not provide useful information. For example, for the question "what is the difference between medicare and medicaid", a useless answer is "...the Center for Medicare and Medicaid Services and the Social Security Administration do not call you to ask you to disclose financial information in order to get a new card. If you receive such a call, you should...". It is relevant to medicare and medicaid but is not useful for answering this question.

Considering that the quality of the direct answer may affect user behavior, we controlled the usefulness of answers to generate SERPs that simulate a range of system configurations: each SERP included ten organic search results shown as traditional snippets, and could additionally include one direct answer, or a set of five multiple answers displayed in a carousel (we chose the number five, because around five search results are shown on SERPs with an answer module before users need to scroll). As a result, for this study, we generated five SERP configurations for each search task, where the quantity and quality of answers vary:

- **S1:** Ten organic search results.
- **S2:** One *useful* answer and ten organic search results.
- **S3:** One *useless* answer and ten organic search results.
- **S4:** Five answers and ten organic search results, where the first answer is *useful* and the same as the answer in S2. The usefulness of the second to fifth answers are the same as those of the second to fifth search result snippets.
- **S5:** Five answers and ten organic search results, where the first answer is *useless* and the same as the answer in S3. The second to fifth answers are the same as those in S4.

In total, six answers are kept for each question: one useful answer for S2, one useless answer for S3, four additional answers as the second to fifth answers for S4 and S5. There are 52.55 words per answer on average (54.70 for useful answers, 50.85 for useless answers). Figure 2 shows SERP examples for the different settings. The broad layout is similar to the SERPs of search engines such as Google. The interfaces of S2 and S3 are the same as each other, and include an answer module containing one direct answer above search results.

[1] slight: 0.0-0.2, fair: 0.2-0.4, moderate: 0.4-0.6, substantial: 0.6-0.8, almost perfect: 0.8-1.0.

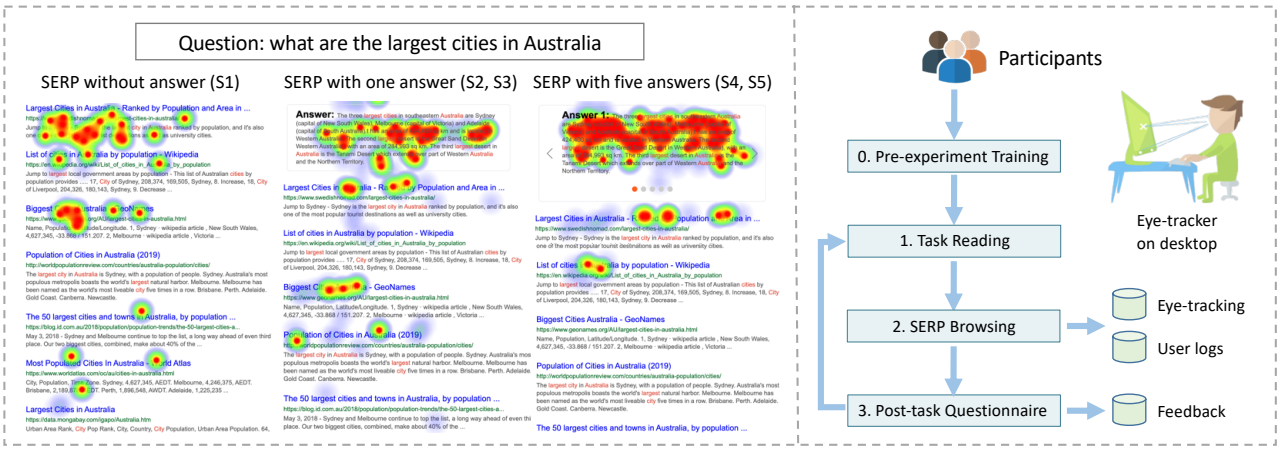


**Figure 2: SERP examples in different settings and user study procedure.**

There is a more advanced answer module in S4 and S5, comprised of five answers in a horizontal carousel; users can click on previous, next, and spot buttons to switch among the answers.

## 3.2 Participants

We recruited 30 participants through on-campus posters and emailing volunteer groups. There were 8 males and 22 females, comprised of graduate and undergraduate students and university staff. Ages ranged from 18 to 46. Areas of study varied from humanities and sociology to natural science and engineering. All participants were fluent in English and indicated that they use computers and search engines at least once a day. Each was asked to perform all 20 of the search tasks. For each task, they were told that they should find answers to the question by browsing one SERP. It took participants around 80 minutes to complete these tasks. Each was paid $55 (USD) as compensation.

Participants were randomly allocated into one of three groups without repetition, such that the first group had six participants while the second and third groups both had 12 participants each. The participants in the first group were shown SERPs with no answer module (20 S1 SERPs). Those in the second group were shown SERPs with one direct answer, either with a useful answer (10 S2 SERPs) or with a useless answer (10 S3 SERPs). Similarly, the participants in the third group were shown SERPs with five direct answers, where the first answer was useful (10 S4 SERPs) or useless (10 S5 SERPs). Therefore, each SERP was viewed six times for each of the system settings. The tasks were shown to each participant in a randomized fashion to avoid ordering effects.

Summary statistics of the tasks and user sessions are shown in Table 1. We designed 20 tasks in total and generated 5 SERPs for each task. There were 120 answers and 200 search results. In total, we collected 600 search sessions from 30 participants, with 120 sessions per setting (S1 to S5).

## 3.3 Procedure

The procedure of the user study is shown in Figure 2. Participants were invited to our lab and used a desktop computer equipped with an eye tracker to perform their activities. After collecting pre-experiment demographic information and carrying out eye-tracker calibration, the study began with two warm-up tasks where we introduced the procedure of the user study. Participants were asked to perform the second task on their own, to ensure they were familiar with the experimental system. Next, the 20 experimental tasks were carried out. They were presented in a random order to avoid learning or order effects. Participants were permitted to take a rest if they felt tired during the study. The procedure of each individual search task is as follows:

**Table 1: Statistics of the dataset in our user study.**

| #Tasks | #SERPs | #Answers | #Results | #Users | #Sessions |
|---|---|---|---|---|---|
| 20 | 100 | 120 | 200 | 30 | 600 |

**Task reading.** First, participants were shown a question. Their task was to find answers to this question. For example, "what are the largest cities in Australia?" This question was also shown at the top of the SERP in the next step.

**SERP browsing.** After participants read the question, a pre-generated SERP was shown. They were instructed to find useful information for the question of this task. They could examine and interact with the SERP as they normally would, including clicking on a result to read a landing page or clicking on links on the landing page if desired. For SERPs including answers, participants could choose whether to read them. Once they felt that they had enough information, or they could find no more, they could leave the SERP by pressing the space bar. During this step, we collected participant eye movements via the eye tracker, and other interactions through our interface, including fixations, timestamps, click behavior, mouse movements, and so on.

**Post-task questionnaire.** After participants left the SERP, they were required to answer three questions: 1) How satisfied do you feel with the search engine result page overall? (five-level rating); 2) Can you successfully answer the question in this search task? (five-level rating); 3) Please briefly write down the answer you found. This explicit feedback from participants helped us better understand their search experience. If they had not found the answer, participants could leave the third question blank. Finally, we

**Table 2: Dwell time and click behavior in different settings. "*/**" indicates that the differences among five settings are statistically significant at $p < 0.05/0.01$ level (Kruskal–Wallis test). "†" indicates that the result is significantly different from that in S1 (Dunn's test).**

| | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Session dwell time (DTime) (s)** | 103.870 | 86.469† | 88.626 | 81.916† | 92.268 |
| DTime on landing pages (s)** | 84.545 | 54.545† | 60.320† | 39.986† | 50.388† |
| Avg. DTime per page (s)** | 35.485 | 31.184† | 32.107 | 26.936† | 28.188† |
| #Clicked results** | 2.433 | 1.625† | 1.775† | 1.300† | 1.633† |
| Lowest clicked results position** | 4.158 | 3.017† | 3.458† | 2.608† | 3.158† |
| Sessions without click (%)** | 1.667 | 14.167 | 5.000 | 12.500 | 9.167 |

showed the answers and search results on the screen again, and participants were asked to assign two-grade usefulness labels for each of the items that they had read.

## 3.4 Experiment System and Platform

We conducted the user study on a desktop computer with a screen resolution of 1920 x 1080. We used a Tobii Pro X2-60 eye tracker to record eye movements. The tracker can detect the presence, attention, and focus of the user. It is common to filter out short duration fixations. Because we found users read faster when browsing a SERP, we adopted a shorter threshold than the 200ms suggested by Lorigo et al. [23], instead using 60ms which was sufficient to reproduce participants' browsing process and preserve most of the eye tracking data. An initial calibration process was carried out for each participant to ensure that the eye tracking data would be recorded accurately. In order to obtain 30 participants, we had to test 40 people, of which ten were eliminated as the system failed to calibrate. This is not an uncommon occurrence in eye-tracking studies. We developed a user study system using Django, through which participants could log in and complete their search tasks. We used a backend database to record participants' explicit feedback and interaction behavior including mouse movements, scrolling, selection behavior, click behavior, and timestamps.

# 4 RESULTS

We first investigated users' interaction patterns on SERPs without and with a direct answer. To address RQ1, we also analyzed the influence of the quality of the answer in users' interactions. Then we studied how users' interactions change when a more advanced answer module containing multiple answers was provided on the SERP to address RQ2. Finally, we considered two types of questions: *factoid* and *non-factoid*, and compared the differences in users' interactions between different question types to address RQ3. The Kruskal–Wallis test [17], which is a non-parametric method used in IIR analysis [16], and does not assume normal distributions, was conducted using the setting of SERPs as the experimental factor. Dunn's test [10] was further used to identify which particular pairs differed significantly. In Section 4.3, the Scheirer–Ray–Hare test [31] was used to test for signficant differences using both the setting of SERPs and types of questions as factors.

## 4.1 RQ1: Interaction Patterns on SERPs with a Single Direct Answer

We compared users' interaction patterns for SERPs, which provide no answer (S1) and those that provided a single answer (S2 and S3) from three perspectives: click behavior, fixation distribution, and examination sequence. We also analyzed users' explicit feedback including satisfaction and success to better understand user behavior.

**Table 3: Examination behavior on SERPs of different settings, including fixation duration (in second), number of examined answers and results. "*/**/†" have the same meaning as in Table 2.**

| | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Fixation duration on the SERP** | 12.520 | 19.242† | 17.223† | 28.130† | 28.104† |
| #Examined answers** | - | 1.000 | 1.000 | 4.133 | 4.308 |
| Lowest examined answers position** | - | 1.000 | 1.000 | 4.158 | 4.317 |
| Duration on answers** | - | 9.503 | 6.504 | 20.795 | 19.500 |
| Avg. duration per answer** | - | 9.503 | 6.504 | 4.991 | 4.436 |
| #Examined results* | 6.200 | 5.150† | 5.517 | 5.625 | 5.508 |
| Lowest examined snippets position | 7.675 | 6.633 | 7.075 | 7.075 | 6.958 |
| Duration on snippets** | 12.520 | 9.739† | 10.719 | 7.336† | 8.609† |
| Avg. duration per snippet** | 1.950 | 1.776 | 1.980 | 1.236† | 1.440† |
| #Clicked results** | 2.433 | 1.625† | 1.775† | 1.300† | 1.633† |
| Lowest clicked result position** | 4.158 | 3.017† | 3.458† | 2.608† | 3.158† |
| Duration on landing pages** | 51.233 | 31.711† | 35.206† | 23.320† | 28.527† |
| Avg. duration per landing page** | 21.461 | 18.317† | 18.444 | 15.705† | 16.422† |

*4.1.1 Click Behavior.* Implicit signals, such as clicks and dwell time, are commonly used to improve search result quality [27]. In Table 2, we report our findings related to these signals. According to the table, users spent about 104 seconds on average browsing SERPs containing only organic search results (S1). When an answer was shown on the SERP, users spent less time completing the search task, especially when the answer is useful (S2). Users also spent significantly less time reading landing pages, even when the answer was not useful. This means that the presence of the answer module leads to shorter sessions.

Users clicked on significantly different numbers of search results in each setting. They clicked on more results on SERPs without a direct answer, and the likelihood of clicking on lower-ranked results is higher. It is worth noting that there was no click in about 14% sessions when we provided useful answers, while this percentage in S1 is less than 2%. This shows that users seldom find sufficient information only through snippets because in about 98% of SERPs users click on at least one search result. Providing a direct answer helps users get more useful information without a click, reducing user effort during the search task. In the S3 setting, there is no click for 5% of sessions. This is higher than the rate of the S1 setting. Given that the S3 setting shows a useless answer, this may mean that the users sometimes over-trust the answer module.

*4.1.2 Fixation Distribution.* Using eye tracking data we analyzed users' examination patterns on the SERP. For each SERP, we considered the fixations on direct answers and search result snippets. We used the sum of fixation duration associated with an answer/snippet as its fixation duration, which has been regarded as important implicit feedback for improving result ranking in search engines [28]. Table 3 shows fixation durations on the SERP, examined answers, examined results snippets, clicked result snippets, and landing pages. A longer fixation duration indicates that users paid more attention to the item. We found that when a direct answer is shown, users paid more attention to the SERP than the landing page, especially when the answer was useful. This indicates that the direct answer helps improve user engagement on the SERP and reduces user effort on the landing page. Within the SERP, users read fewer search

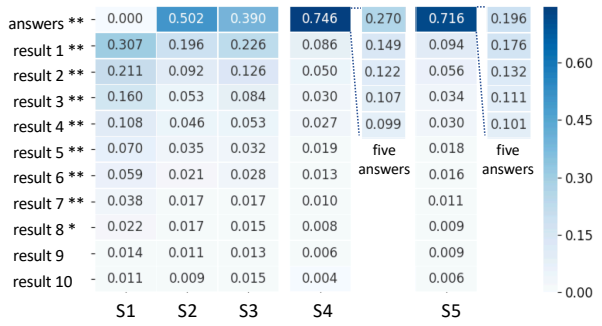


**Figure 3: Vertical distributions of fixation duration on different SERP settings. The proportions of fixation duration for the five answers shown for S4 and S5 are also shown. "*/**" indicates that the differences among the five settings are statistically significant at $p < 0.05/0.01$ level (Kruskal–Wallis test).**

result snippets and paid less attention to snippets when a direct answer is shown. The average number of examined answers was 1 in S2 and S3, which shows that users always read the provided answer, but the fixation duration was longer when the answer is useful.

Previous work has demonstrated that users exhibit a position bias when examining results on a SERP [6, 43]. To investigate the situation when a direct answer is provided, we examined the distribution of fixation durations for each answer and search result snippet. The results are shown in Figure 3. Values in the figure are the proportions of fixation duration. For example, "0.502" in the first row indicates that 50.2% of fixations on the SERP in S2 are located in the answer module. The values in the top left cell are zero since there is no answer module in S1. We can see that users tend to pay more attention to the answer module and top-ranked search results. Higher-ranked results received more user attention and were more likely to be examined. This is consistent with the findings of Joachims et al. [14] and Liu et al. [21]. The answer attracted the most attention among all items on the SERP in S2 and S3. The first item attracted more attention when it is an answer rather than a search result (30.7%). Especially when the answer is useful, more than 50% of the attention was paid to it. This indicates that users tend to first carefully read the provided answer. Fixation duration on the same result was significantly different when settings vary in the top seven results. It is longer when there is no answer or the answer is useless. When the answer is useful, the attention decreased faster with the vertical position. This is because users can access useful information faster in S2.

The above observations illustrate that users spend more time carefully reading the SERP and less time reading landing pages. The position bias affects users' attention distribution on the SERP. When a useful answer is provided, users tend to first carefully read it to get useful information, and their attention decreases faster with the vertical position.

*4.1.3 Examination Sequence.* To further analyze the temporal sequence of users' examination behavior, we plotted the arrival time

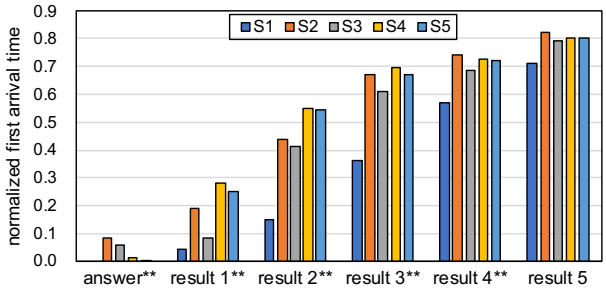


**Figure 4: Normalized first arrival time at different vertical positions of SERPs. "*/**" indicates that the differences among five settings are statistically significant at $p < 0.05/0.01$ level (Kruskal–Wallis test).**

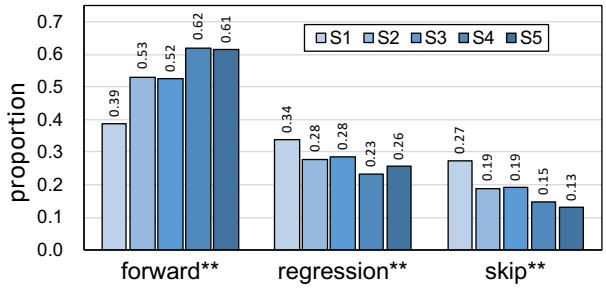


**Figure 5: Examine transitions among items (answers and search results) on SERPs. "**" indicates that the differences among five settings are statistically significant at $p < 0.01$ level (Kruskal–Wallis test).**

distribution of different ranked positions in different settings as Figure 4 shows. The y-axis is the first arrival time of eye fixations for a given position, which is normalized by the whole session length [33]. We only plotted the first arrival time of the answer module and the top five search results, since only the first five search results are shown on the SERP before users need to scroll. This shows that the first arrival time increases with the rank, which is consistent with the findings of Granka et al. [12] that the user usually browses the SERP from top to bottom on average. We can also observe that when there is an answer on the SERP, the first arrival time to search results is larger. The differences decrease as rank increases. The first arrival time to the same search result is smaller when the direct answer is useless. If users found that the answer cannot provide useful information, they chose to move on to the search results below more quickly.

We next analyzed the fixation transitions among items (i.e. the answer and search results) on the SERP. Since items were listed in one-dimension, the transitions are in two directions: up and down. We split users' fixation transitions into three categories based on past work [25]:

- **Forward:** users' fixations go down to the next search result snippet on the SERP.
- **Regression:** users' fixations go up to the direct answer or search result snippets with higher ranks.
- **Skip:** users' fixations skip some results and go down to search result snippets which are at lower ranks.

Percentages of the three transitions are shown in Figure 5. In all settings, most transitions are forward, followed by regressions, and then skips. Users preferred short-distance transitions between two contiguous items, which is consistent with past work [34]. It also indicates that users tend to read items on the SERP from top to bottom, one by one, but sometimes they went back to revisit some items or skip others. The result is aligned with past findings in document reading [20]. On SERPs with an answer, there are more forward transitions, fewer regressions, and fewer skip transitions. The differences among settings are significant, and indicate that a direct answer helps users by reducing the need to go back or go deeper to find more useful information.

*4.1.4 Explicit Feedback.* From users' explicit feedback, we can gain additional insight into the way that a direct answer affects the user experience. Table 4 shows some results about user satisfaction and task success in different settings. Compared to S1, users who are shown SERPs with a direct answer reported a higher level of satisfaction and success, especially when the answer is useful. Despite these differences in task success, no significant differences were detected. SERP setting had a statistically significant effect on user satisfaction.

*4.1.5 Summary.* We investigated users' interaction patterns on a SERP interface containing a direct answer to the search query above the search results. Our findings related to RQ1 are:

- Users focused more extensively on the top-ranked items in the SERP. The answer module attracts a lot more attention from users than a regular search result placed at the same position attracts. It also leads to a decrease in attention given to the rest of the SERP, even when the provided answer is not useful.
- Providing a direct answer (even when the provided answer is not useful) helps users get more useful information without a click and shortens the time spent on completing search tasks, reducing user effort.
- Users tend to read items on the SERP from top to bottom, one by one. A direct answer helps users by reducing the need to go back to previous results or to go deeper to find more useful information.
- Providing a direct answer helps users perceive a higher level of satisfaction. Regarding user perception of task success, the improvement is not significant, but an increasing trend is observed.

## 4.2 RQ2: Interaction Patterns on SERPs with Multiple Answers

To investigate the impact of providing multiple answer items on users' interaction patterns, we focused on SERPs with five answers presented in a carousel, where only one answer is visible at a time. Users can click on previous, next, and spot buttons to switch to other answers. We compared users' interaction patterns on such an interface (S4 and S5) with those described in Section 4.1 to answer RQ2.

*4.2.1 Click Behavior.* Table 2 shows users' click behavior and dwell time on search results. Users spent only 82 seconds to finish the search task on average in the S4 setting, where multiple answers were provided with the first being useful. The dwell time was even shorter than that of the S2 setting, where only one direct useful answer was provided. The dwell time on landing pages was 40 seconds on average, which is half of that in S1. Results on the number of clicks and sessions without a click when multiple answers were provided have the same trends as those in S2 and S3. Even if the first answer is useless, there are 9.167% of sessions without a click. Compared to showing one answer, multiple answers provide more choices for users. Providing multiple direct answers further helps to reduce user effort during the search task.

**Table 4: User satisfaction and task success in different settings. "*" indicates that the differences among five settings are statistically significant at $p < 0.05$ level (Kruskal–Wallis test). Differences between different setting pairs are not statistically significant (Dunn's test).**

| | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| Satisfaction* | 4.308 | 4.508 | 4.425 | **4.555** | 4.318 |
| Task success | 4.342 | 4.417 | 4.392 | **4.527** | 4.464 |

*4.2.2 Fixation Distribution.* Statistics of fixation durations on the SERP, answers, search result snippets, and landing pages are shown in Table 3. We find that users pay significantly more attention to the SERP and less attention to landing pages when multiple answers are provided. Multiple direct answers further help improve user engagement on the SERP and reduce user effort on the landing page. The total duration of answers is slightly shorter when the first answer is useless, but the number of examined answers is larger. This indicates that users tend to read more answers when they found the first answer cannot satisfy their information needs. They were patient with the advanced answer module and are willing to explore it. The gaps between S4 and S5 are smaller than those between S2 and S3, suggesting that failing to provide a useful first answer has less influence on users when multiple additional answers are present.

The vertical distributions of fixation duration on SERPs in S4 and S5 are shown in Figure 3. Direct answers attracted more than 70% of fixations on the SERP. We also plotted the duration distribution of the five answers. Users paid more attention to the fifth answer (9.9% in S4 and 10.1% in S5) than the first search result (8.6% in S4 and 9.4% in S5). Fixation duration on the first answer is longer when it is useful, while durations on other answers are shorter. We can also observe that users' attention decays the fastest with the vertical position when multiple answers are provided. Users paid more than 80% attention to the first two items on the SERP, while the cumulative duration reaches 80% in the fifth item in S1, in the fourth item in S2 and S3. This suggests that users already get most of the useful information from a few top-ranked items. Therefore, they pay less attention to the lower-ranked search results.

*4.2.3 Examination Sequence.* Figure 4 shows the normalized first arrival time at different vertical positions. We can observe that the first arrival time of the first and second search results in S4 and S5 is larger than that in S2 and S3. It shows that the number of provided answers only affects the arrival time of those top-ranked search results. When the vertical position increases, differences between the first arrival time across the five settings become smaller. Figure 5

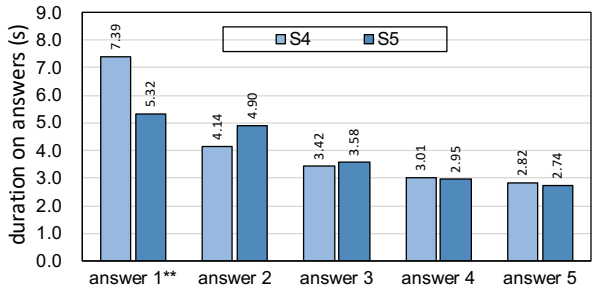


**Figure 6: Average fixation duration on the first to fifth answers. Durations on the first answer are significantly different in S4 and S5 at $p < 0.01$ (Kruskal–Wallis test).**

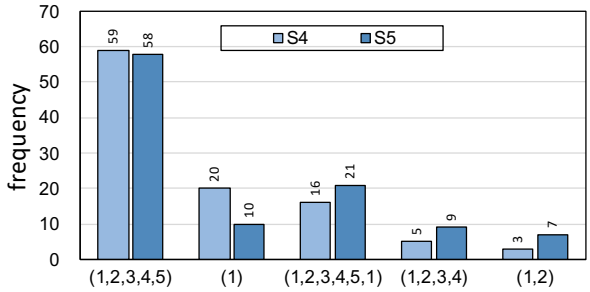


**Figure 7: The five most frequent answer examination sequences.**

shows the percentages of three types of examination transitions on the SERP. There are more forward transitions, and fewer regressions and skip transitions on SERPs with multiple answers, indicating that multiple answers can help users reduce the need to go back or go deeper to find more useful information, even more than when just providing one answer.

*4.2.4 Interactions within the answer module.* We further analyzed users' interactions within the advanced answer module, including click behavior, fixation distribution, and examination sequence. Table 5 shows that users tend to browse answers from the first to the fifth by clicking on the next button. They seldom used spot buttons to locate a certain answer. There are also some clicks on the previous button, which shows that users revisit answers sometimes. When the first answer is useful, there is more revisiting behavior. Users had more interactions with answers when the first answer is useless. The average fixation duration on the first to fifth answers is shown in Figure 6. We find that when the first answer can provide useful information, users spend more time reading it. As to the second answer, users spent more time reading it when the first answer cannot provide useful information. There is no significant difference in fixation duration on the third, fourth, and fifth answers when the usefulness of the first answer varies.

We extracted user examination sequences in the advanced answer module for each search session, and show the frequencies of the five sequences with the highest frequency in Figure 7. For example, the sequence $(1, 2)$ indicates that users first read the first answer, then read the second answer, and then no longer read answers in the answer module. It is observed that about half of the users read all five answers, from the first one to the fifth one. Some users reread the first answer after they read all five answers. When the first answer can provide useful information, the frequency of only reading the first answer is twice of that when the first answer is useless, showing that users are highly likely to leave the answer module if they already found useful information in the first answer.

*4.2.5 Explicit Feedback.* We analyzed users' satisfaction and task success on SERPs with multiple answers. From Table 4 we observe that users get the highest level of satisfaction when the first answer is useful. Results on search success are similar. Even if the first answer is useless, users perceive a higher level of task success than just providing one direct answer. It indicates that providing multiple answers on the SERP may help improve user satisfaction and task success.

**Table 5: Statistics of users' clicks within the answer module in S4 and S5. The differences between S4 and S5 are not statistically significant.**

| | S4 | S5 |
|---|---|---|
| #Clicks on previous button | 0.158 | 0.067 |
| #Clicks on next button | 3.442 | 3.642 |
| #Clicks on spot button | 0.092 | 0.042 |
| #Clicks | 3.692 | 3.750 |

*4.2.6 Summary.* We compared users' interaction patterns on SERPs that show multiple answers with those that show only one answer. To answer RQ2, we summarize our findings as follows:

- Providing multiple direct answers helps further reduce user effort, improves user engagement on SERPs, and improves user satisfaction.
- Even when the first answer of the multi-answer module is not useful (the S5 setting), users prefer to explore the answer module, rather than moving to the regular result list.

## 4.3 RQ3: Interaction Patterns in Questions of Different Types

To further understand users' interactions on SERPs for different types of questions, we analyzed interaction patterns for both factoid and non-factoid questions. In Section 4.1 and 4.2, we mentioned that users perceive the highest level of satisfaction and task success in S4, where multiple answers are provided and the first answer is useful. Therefore, in this section we focus on S1 and S4, and investigate how these two factors, setting and question type, affect users' interactions. The Scheirer–Ray–Hare test, a non-parametric alternative to multi-way ANOVA, was conducted to test the significance of the influences caused by these two factors.

We analyzed dwell time, click behavior, fixation, user satisfaction, and task success. The results are shown in Table 6. Users spent a longer time completing non-factoid questions than factoid questions. Question type and setting both had a significant influence on the session dwell time. Providing direct answers helps reduce users' time effort in both types of questions. When we provide direct

**Table 6: Dwell time, click behavior, fixation, and explicit feedback in factoid and non-factoid questions. Time is measured in seconds. "*/†/‡" indicates the question type/setting/their interaction has a statistically significant effect on the variable at $p < 0.05$ (Scheirer–Ray–Hare test).**

| | factoid | | non-factoid | |
|---|---|---|---|---|
| | S1 | S4 | S1 | S4 |
| Session dwell time (DTime) * † | 89.625 | 73.362 | 118.114 | 90.469 |
| DTime on landing pages * † | 71.359 | 33.861 | 97.731 | 46.112 |
| Avg. DTime per page * † | 33.186 | 24.279 | 37.783 | 29.592 |
| #Clicked results * † | 2.150 | 1.133 | 2.717 | 1.467 |
| Lowest clicked results position * † | 3.583 | 1.867 | 4.733 | 3.350 |
| Sessions without click (%) * † | 3.333 | 18.333 | 0.000 | 6.667 |
| Fixation duration on the SERP † | 12.099 | 27.058 | 12.941 | 29.203 |
| #Examined answers † | - | 4.033 | - | 4.233 |
| Lowest examined answers position † | - | 4.050 | - | 4.267 |
| Duration on answers † | - | 19.443 | - | 22.147 |
| Avg. duration per answer † | - | 4.708 | - | 5.274 |
| #Examined results * | 5.733 | 5.217 | 6.667 | 6.033 |
| Lowest examined snippets position * | 7.167 | 6.750 | 8.183 | 7.400 |
| Duration on snippets † | 12.099 | 7.615 | 12.941 | 7.056 |
| Avg. duration per snippet † | 2.026 | 1.296 | 1.873 | 1.176 |
| #Clicked results * † | 2.150 | 1.133 | 2.717 | 1.467 |
| Lowest clicked results position * † | 3.583 | 1.867 | 4.733 | 3.350 |
| Duration on landing pages * † | 44.222 | 20.355 | 58.243 | 26.284 |
| Avg. duration per landing page * † | 20.211 | 14.645 | 22.71 | 16.765 |
| Satisfaction * † ‡ | 4.417 | 4.691 | 4.200 | 4.418 |
| Task success * † ‡ | 4.533 | 4.673 | 4.150 | 4.382 |

answers, the session dwell time had an 18.1% and 23.4% reduction under factoid and non-factoid questions, respectively. For factoid questions, users clicked on fewer search results. The lowest ranks of clicked results in S1 and S4 are 3.583 and 1.867, respectively. Adding direct answers to SERPs of factoid questions reduced the lowest position of clicked results by around 48%. In this configuration, there was also no click in 18.3% of search sessions, indicating that users' click behavior is affected more by direct answers under factoid questions than non-factoid questions. For non-factoid questions, there was at least one click in each search session in S1.

With respect to the fixation distribution in factoid and non-factoid questions, there is no significant difference in fixation duration on the SERP and answers, while durations on snippets and landing pages are significantly longer for non-factoid questions. However, users perceived a significantly lower level of satisfaction and task success for non-factoid questions. This can be explained by non-factoid questions being more complex than factoid questions, making it more difficult for the search engine to return satisfying search results and therefore reducing users' confidence in answering such questions correctly.

To answer RQ3, we summarize our findings as follows:

- Providing direct answers changes users' click behavior more in the case of factoid questions than non-factoid questions.
- Non-factoid questions lead to increased user effort and significantly lower levels of satisfaction.

## 5 CONCLUSION

In this paper, we investigated how users interact with the SERP when direct answers to question-style queries are provided with search results. To the best of our knowledge, this is the first work to study user behavior on SERPs containing an answer module. Through a user study with 30 participants, we find that the answer module helps users complete search tasks more quickly, and reduces user effort. It attracts more fixations and improves users' engagement with the SERP. These results indicate that the answer module has a positive effect on users during answer-seeking tasks. We further study the effect of direct answer quantity and question type on user behavior. Results show that users tend to explore more in the answer module and less in other search results when they are provided with multiple answers in a carousel interface. Users' click behavior is affected more for factoid questions than non-factoid questions. No results were clicked in 18% of search sessions when answering factoid questions with multiple direct answers. This shows that users can often get enough useful information from the SERP without having to view a landing page.

Our findings provide new insight for the design of SERPs, such as informing under which situations web-based information retrieval systems should show an answer module, and have a number of implications. First, we showed that the presence of the answer module on the SERP has a stronger impact on the search metrics than the content of the module. In particular, we observed a decrease in the number of clicks issued and the duration of search sessions when an answer module is present on the SERP, even when the displayed module contains a poor answer. This points at the importance of the answer module triggering problem. That is, deciding when to display the answer module should be considered a problem at least as important as deciding what to display in the module. Second, as demonstrated by our eye-tracking experiments, we observed increased user engagement with the multi-answer module, compared to the single-answer module. This implies that users are willing to inspect more direct answers before they start inspecting potentially longer landing pages. Interestingly, none of the major commercial search engines currently support multi-answer modules in their SERP designs, and this remains as a potential venue for exploration. Third, we observed a significant decrease in the number of traditional search results inspected by users when the answer module is present, i.e., users do not go too deep in search result lists. This may imply that the presence of the answer module calls for different SERP designs. For example, the search engine may display fewer search results to users, allocating the available space to other kinds of assets, such as advertisements, question suggestions, or other vertical search results.

As with any research, there are potential limitations to our experiments. First, the answer modules shown in commercial search engines usually include a combination of text, images, and links. The answer modules displayed in our study contain only textual information. Second, although we conducted two warm-up search tasks, the participants may have had a bias to explore the direct answer module more as this is a relatively novel interface. Third, to control the answer quality, the interaction of participants with the search interface was limited in certain ways. For example, although the participants could interact with the search results freely, they were not allowed to reformulate their queries.

Our results reveal certain user behavior not observed in published literature. In the future, we plan to study user behavior further, examining a more dynamic search setup. We believe such studies can provide a better understanding of users' answer-seeking process and provide new insights for SERP design.

## ACKNOWLEDGMENTS

This research was partially supported by the Australian Research Council (Project DP180102687).

## REFERENCES

[1] Jaime Arguello, Fernando Diaz, Jamie Callan, and Jean-Francois Crespo. 2009. Sources of Evidence for Vertical Selection. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Boston, MA, USA, 315–322.

[2] Anne Aula, Päivi Majaranta, and Kari-Jouko Räihä. 2005. Eye-tracking reveals the personal styles for search result evaluation. In *IFIP conference on human-computer interaction*. 1058–1061.

[3] Flavio Chierichetti, Ravi Kumar, and Prabhakar Raghavan. 2011. Optimizing Two-Dimensional Search Results Presentation. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. 257–266.

[4] Charles L. A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. 2007. The Influence of Caption Features on Clickthrough Patterns in Web Search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam, The Netherlands, 135–142.

[5] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2015. User Activity Patterns During Information Search. *ACM Trans. Inf. Syst.* (2015).

[6] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the international conference on web search and web data mining*. Palo Alto, CA, 87–94.

[7] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for? An eye-tracking study of information usage in Web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 407–416.

[8] Andrew Duchowski. 2007. *Eye Tracking Methodology Theory and Practice* (2 ed.). Springer.

[9] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *Proceedings of the Third Symposium on Information Interaction in Context*. New Brunswick, New Jersey, USA, 185–194.

[10] Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics* 6, 3 (1964), 241–252.

[11] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[12] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-Tracking Analysis of User Behavior in WWW Search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[13] Jimmy Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. 2019. Health Cards for Consumer Health Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

[14] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*. Salvador, Brazil, 154–161.

[15] Yvonne Kammerer and Peter Gerjets. 2014. The Role of Search Result Position and Source Trustworthiness in the Selection of Web Search Results When Using a List or a Grid Interface. *International Journal of Human–Computer Interaction* 30, 3 (2014), 177–191.

[16] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and trends in Information Retrieval* (2009).

[17] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952).

[18] Dmitry Lagun and Eugene Agichtein. 2011. ViewSer: Enabling Large-Scale Remote User Studies of Web Search Examination and Interaction. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Beijing, China, 365–374.

[19] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[20] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding Reading Attention Distribution during Relevance Judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery.

[21] Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From Skimming to Reading: A Two-Stage Examination Model for Web Search. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*.

[22] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of Vertical Result in Web Search Examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 193–202.

[23] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *JASIST* 59 (05 2008).

[24] Emma Lurie and Eni Mustafaraj. 2018. Investigating the Effects of Google's Search Engine Result Page in Evaluating the Credibility of Online News Sources. In *Proceedings of the 10th ACM Conference on Web Science*.

[25] Scott A McDonald and Richard C Shillcock. 2003. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research* 43, 16 (2003), 1735–1751.

[26] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268

[27] Neil O'Hare, Paloma de Juan, Rossano Schifanella, Yunlong He, Dawei Yin, and Yi Chang. 2016. Leveraging User Interaction Signals for Web Image Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 10.

[28] Bing Pan, Helene A. Hembrooke, Geri K. Gay, Laura A. Granka, Matthew K. Feusner, and Jill K. Newman. 2004. The Determinants of Web Page Viewing Behavior: An Eye-Tracking Study. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications (ETRA '04)*. 147–154.

[29] Chen Qu, Liu Yang, W Bruce Croft, Falk Scholer, and Yongfeng Zhang. 2019. Answer interaction in non-factoid question answering systems. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM.

[30] Daniel E. Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web*.

[31] C James Scheirer, William S Ray, and Nathan Hare. 1976. The analysis of ranked data derived from completely randomized factorial designs. *Biometrics* (1976).

[32] Aliaksei Severyn and Alessandro Moschitti. 2013. Automatic feature engineering for answer selection and extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 458–467.

[33] Shanu Sushmita, Hideo Joho, Mounia Lalmas, and Robert Villa. 2010. Factors Affecting Click-through Behavior in Aggregated Search Interfaces. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. 519–528.

[34] Benjamin W. Tatler and Benjamin T. Vincent. 2009. The prominence of behavioural biases in eye guidance. *Visual Cognition* 17, 6-7 (2009), 1029–1054.

[35] Jaime Teevan, Edward Cutrell, Danyel Fisher, Steven M. Drucker, Gonzalo Ramos, Paul André, and Chang Hu. 2009. Visual Snippets: Summarizing Web Pages for Search and Revisitation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Boston, MA, USA, 2023–2032.

[36] Paul Thomas, Alistair Moffat, Peter Bailey, Falk Scholer, and Nick Craswell. 2018. Better Effectiveness Metrics for SERPs, Cards, and Rankings. In *Proceedings of the 23rd Australasian Document Computing Symposium*. Dunedin, New Zealand, 8.

[37] Paul Thomas, Falk Scholer, and Alistair Moffat. 2013. What Users Do: The Eyes Have It. In *Asian Information Retrieval Symposium*. 416–427.

[38] Anastasios Tombros and Mark Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*. 2–10.

[39] Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. 2015. What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries. 1571–1580.

[40] Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. 2013. Incorporating Vertical Results into Search Click Models. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland, 503–512.

[41] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 189–198.

[42] Allison Woodruff, Ruth Rosenholtz, Julie B. Morrison, Andrew Faulring, and Peter Pirolli. 2002. A Comparison of the Use of Text Summaries, Plain Thumbnails, and Enhanced Thumbnails for Web Search Tasks. *Journal of the American Society for Information Science and Technology* 53, 2 (2002), 172–185.

[43] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating Examination Behavior of Image Search Users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 275–284.

[44] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human Behavior Inspired Machine Reading Comprehension. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Paris, France, 425–434.