Users, Queries, and Bad Abandonment in Web Search

by

Mustafa Abualsaud

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Falk Scholer Professor, School of Computing Technologies, RMIT
Supervisor(s):	Mark D. Smucker Professor, Management Sciences, University of Waterloo
Internal Member:	Charles L. A. Clarke Professor, School of Computer Science, University of Waterloo Gordon V. Cormack Professor, School of Computer Science, University of Waterloo
Internal-External Member:	Mark Hancock Associate Professor, Management Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapter 3 of this thesis describes the first user study conducted to understand query abandonment and the first paper published (Zhang, Abualsaud, and Smucker, 2018). The work was done by myself and in collaboration with Haotian Zhang, another PhD student at the time, with guidance from Mark Smucker. The implementation of the user study system, as well as the logging of user behavior was done by myself. My contribution in the work involved all stages of the research project including initial discussions, participant recruitment, topic creation, data analysis, and reporting of results. Implementation of the web application was done by myself. The work was first presented by myself at Google in December 2017, prior to the CHIIR'18 conference.

Part of the introduction written in Chapter 3 was taken from the paper (Zhang, Abualsaud, and Smucker, 2018), which Mark Smucker has contributed to. Chapters 4, 5, and 6 describes work that was done by myself, with guidance from Mark Smucker and Charles Clarke.

This thesis includes peer-reviewed material that has appeared in conference and journal proceedings published by the Association for Computing Machinery (ACM). The ACM's policy on reuse of published materials in a dissertation is as follows:

> *"Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included."*

The following list of papers serves as a declaration of the Versions of Record for works included in this thesis:

Zhang, H., M. Abualsaud, and M. D. Smucker (2018). A study of immediate requery behavior in search. In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, CHIIR '18. ACM. doi: 10.1145/3176349.3176400

Abualsaud, M. and M. D. Smucker (2019a). Patterns of search result examination: Query to first action. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19. ACM. doi: 10.1145/3357384.3358041

Abualsaud, M. (2020). The effect of queries and search result quality on the rate of query abandonment in interactive information retrieval. *In Proceedings of the 2020*

Conference on Human Information Interaction and Retrieval, CHIIR '20. ACM. doi: 10.1145/3343413.3377951

Abualsaud, M., M. D. Smucker, and C. L. Clarke (2021). Visualizing searcher gaze patterns. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '21. ACM. doi: 10.1145/3406522.3446041

Abualsaud, M. and M. D. Smucker (2021). The dark side of relevance: The effect of non-relevant results on mobile search behavior. Under review

Abstract

After a user submits a query and receives a list of search results, the user may abandon their query without clicking on any of the search results. A *bad query abandonment* is when a searcher abandons the Search Engine Result Page (SERP) because they were dissatisfied with the quality of the search results, often making the user reformulate their query in the hope of receiving better search results. As we move closer to understanding when and what causes a user to abandon their query under different qualities of search results, we move forward in an overall understanding of user behavior with search engines. In this thesis, we describe three user studies to investigate bad query abandonment.

First, we report on a study to investigate the rate and time at which users abandon their queries at different levels of search quality. We had users search for answers to questions, but showed users manipulated SERPs that contain one relevant document placed at different ranks. We show that as the quality of search results decreases, the probability of abandonment increases, and that users quickly decide to abandon their queries. Users make their decisions fast, but not all users are the same. We show that there appear to be two types of users that behave differently, with one group more likely to abandon their query and are quicker in finding answers than the group less likely to abandon their query.

Second, we describe an eye-tracking experiment that focuses on understanding possible causes of users' willingness to examine SERPs and what motivates users to continue or discontinue their examination. Using eye-tracking data, we found that a user deciding to abandon a query is best understood by the user's examination pattern not including a relevant search result. If a user sees a relevant result, they are very likely to click it. However, users' examination of results are different and may be influenced by other factors. The key factors we found are the rank of search results, the user type, and the query quality. For example, we show that regardless of where the relevant document is placed in the SERP, the type of query submitted affects examination, and if a user enters an ambiguous query, they are likely to examine fewer results.

Third, we show how the nature of non-relevant material affects users' willingness to further explore a ranked list of search results. We constructed and showed participants manipulated SERPs with different types of non-relevant documents. We found that user examination of search results and time to query abandonment is influenced by the coherence and type of non-relevant documents included in the SERP. For SERPs coherent on off-topic results, users spend the least amount of time before abandoning and are less likely to request to view more results. The time they spend increases as the SERP quality improves, and users are more likely to request to view more results when the SERP contains diversified non-relevant results on multiple subtopics.

Acknowledgements

I would like to thank all the people who made this thesis possible. My sincere gratitude to my supervisor Mark Smucker, whom I had the pleasure to work with during both my Master's and Ph.D. degrees. Mark was a great mentor throughout my time at the University of Waterloo, providing invaluable guidance and advice. His motivating ideas within information retrieval made me enjoy the research area even more. I would also like Charles Clarke for valuable feedback and discussions regarding some of the topics in this thesis, and for advice on writing academic papers. I would additionally like to thank my dissertation committee members: Falk Scholer, Charles Clarke, Gordon Cormack and Mark Hancock, for serving on my committee and for their valuable feedback on the thesis.

My experience in graduate school substantially changed the way I think and approach problems. I would like to thank all of my friends and colleagues that made this journey enjoyable. I am very thankful to have met great friends and colleagues, whom I had great discussions that I enjoyed immensely.

I also acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [Grants CRDPJ 468812-14, RGPIN-2014-03642, RGPIN-2020-04665], Google, and the University of Waterloo.

Dedication

To my family

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis investigates a specific type of query abandonment in web search. In particular, we study *bad* query abandonment as being when a user does not click on any search results and instead decides to reformulate their query to continue their search.

1.1 Preamble

Understanding the nature of people's search tasks and the behavior of people's search, comprehension, and interaction with information can offer important implications on both the design and evaluation of search systems. In turn, it can help designers of search engines build better algorithms and interfaces that deliver successful searches to users.

Search engine algorithms and interfaces have progressed significantly over the years. Hearst (2009, Chapter 1) showed a screenshot of Google's search interface in 2007 for the query "darter habitat". A user seeking more information about darter habitats (e.g., where do darter habitat live?) may be presented with an ordered list of search results summerized by a title and a short snippet (Figure 1.1 in Hearst (2009, Chapter 1)). To find an answer to their information need, the user would need to examine the search results and click on the document(s) they believe contain the answer. Such a decision would require the user to determine which document(s) to click, e.g., the first document, the second document, a document on the next page of results, or even reformulate the search query to find other and possibly better search results. White (2016, Chapter 1) notes that fact-finding in the web, as in the mentioned example, "may require only a single resource or direct answer, and strong system performance may be evidenced by low searcher engagement and short

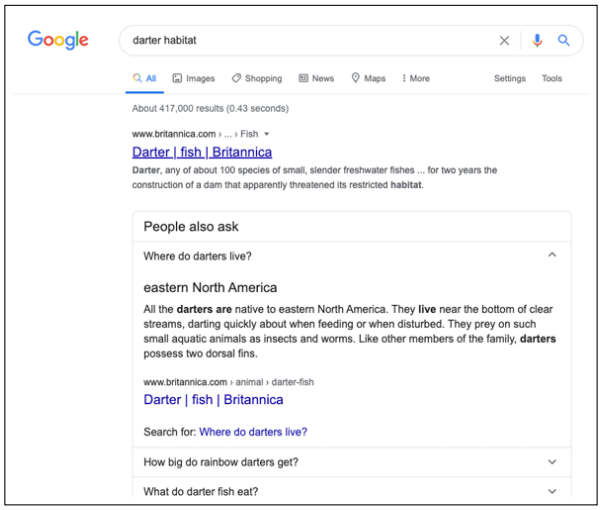


Figure 1.1: Google search results in 2020. Screenshot taken on June 26th, 2020. Google has added direct answers into the search result page, shortening task completion time for users trying to find answers about a certain topic.

task completion times". Knowing such information, whether by experiments, analysis of search logs, or other methods, would benefit Information Retrieval (IR) system designers in designing targeted search support tools that address what users and their behavior indicate (e.g., shortening task completion times for fact-finding tasks). Searching for the same query today returns targeted search support, such as the direct answer box shown in Figure 1.1, that users can use to achieve the same goal without leaving the search page and with a shorter amount of time.

The implications of understanding search behavior can help in many aspects of the search process beyond the mentioned example. In return, it can help build better search systems that significantly improve search satisfaction of the millions of users that use it.

User-centered aspects in IR

Dedicated academic venues exist to discuss research on the many user-centered aspects of information interaction and IR. We briefly explain some aspects while describing a traditional search process.

Searchers often use web search engines to seek information regarding a particular task in mind, whether exploratory in nature, finding facts, or learning a new topic of interest.

As users are typically responsible for creating their search queries, they start by formulating a query they believe is relevant to their information need. For example, a user interested in knowing the height of Mount Everest may choose to enter the query "mount everest","mount everest height", or any other query the user believes relevant. This process of formulating a query can be straightforward to a user familiar with the topic of their information need and who knows the right terms to include in their search query. The terms "mount everest" and "height" combined, for example, seem as a well-formed query for someone looking to find the height of Mount Everest, as it includes important terms relevant to what the user is trying to learn. Query formulation can be more difficult when users are less familiar with their search topic and unaware of the correct vocabulary to use. Supporting the searcher in formulating their queries opens up a whole user-centered aspect of IR identified as *query suggestions* that focuses mainly on ways to help searchers come up with more relevant and better queries.

Another user-centered aspect of IR, identified as *query reformulation*, starts right after a user submits a query and is presented with search results. For example, following a user's initial query, the user may be presented with search results considered irrelevant to what they are looking for and not worthy of further exploration. In other words, the search results do not satisfy the task the user is trying to achieve. In today's search engines, when this occurs, users are faced with the following choices: examine more search results on the next search page(s), submit a reformulated query to the search engine, or quit the search process. Reformulating and submitting a new query can be seen as a partial failure by the search engine. It is partial because the user has not yet reached the point of quitting the search process, but is considered a failure because the search engine failed to deliver satisfying information in their initial query. When a user decides to submit a reformulated query and not click on any search results, their action is considered an abandonment of the search results. This behavior is commonly known as *query abandonment*. The other option of quitting the search process is obviously undesirable, as it would indicate complete failure of the search system to satisfy the user. If a user quits, their action is also considered an abandonment of the query, except this time, the user has completely given up on the search engine.

In the two examples of abandonment mentioned above, we assumed the user action is driven by dissatisfaction with the search results. An abandonment that is driven by dissatisfaction is commonly known as *bad query abandonment*. This type of abandonment is undesirable because it is associated with user dissatisfaction, and search engines have dedicated efforts to reduce it (Das Sarma et al., 2008). Query abandonment, however, can also be driven by users' satisfaction with the search results, e.g., when direct answers are presented in the SERP like in Figure 1.1. When a user abandons search results because they

have found what they are looking for directly in the SERP, their behavior can be considered a desirable and positive signal by the search engine. In other words, the search engine successfully returned the information the user is looking for without requiring the user to click on any search result. This behavior is commonly known as *good query abandonment*.

1.2 Query Abandonment

A deeper look into query abandonment

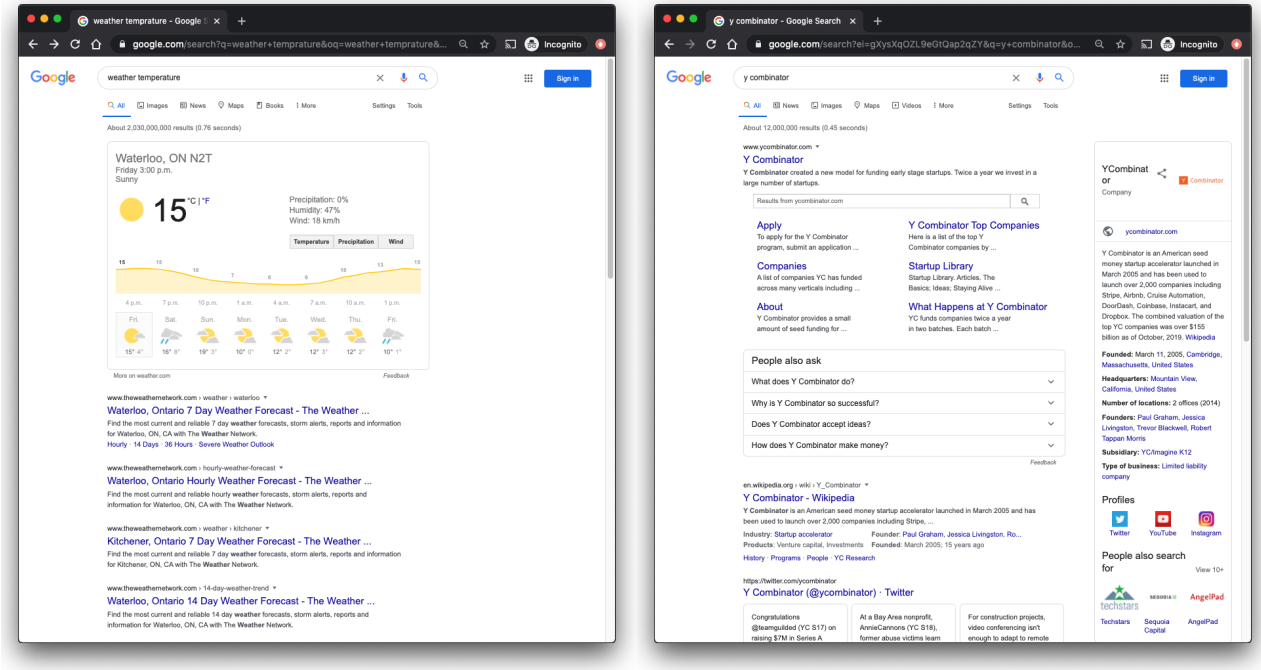
As search results are presented to the users, it has been recognized that some search users will decide to reformulate their query without clicking on any search results, therefore abandoning the query's search results. While the terminology describing this behavior varies, it is commonly referred to as *query abandonment*. Joachims and Radlinski (2007) termed "abandonment" to be "the user's decision to not click on any of the results." Likewise, Radlinski et al. (2008) defined *abandonment rate* to be "the fraction of queries for which no results were clicked on." White (2016) defines it as *search abandonment* and describes it as "abandonment [that] occurs when searchers do not click on any of the results returned by the search engine". Unfortunately, "query abandonment" also sounds similar to what a user does after clicking on a result and deciding to reformulate a query. Indeed, Wu and Kelly (2014) defined query abandonment as "the point at which a person decides to stop his/her current query and enter a new one." As White (2016) explains, it is difficult to define abandonment because there are many ways in which searchers can abandon a SERP, e.g., closing the browser window, clicking on a query suggestion, or others.

There is a host of reasons why people may abandon their queries. For example, a search query on the current local weather temperature in Google returned answers presented directly in the search result (e.g., Figure 1.2a). If an answer addressing the user's information need is presented directly in the SERP, it is not uncommon for users to abandon the search result. Their information need has been satisfied without the need to open any of the search results. This behavior is considered a positive and desirable interaction, and is commonly known as "good query abandonment" (Li et al., 2009a).

Abandonment can also be due to dissatisfaction with the search results. To illustrate this with an example, let us picture the following imaginary, yet possibly common scenario. A user might be interested in learning more about the Y combinator function[1],[2] and decides

[1]In programming languages, the Y combinator is a higher-order function that allows us to do recursion in a programming language that does not have any recursion mechanisms implemented.

[2]Credit to Prof. Gordon Cormack for showing this query example during a lecture on web search.



(a) "weather temprature" search results.

(b) "y combinator" search results.

Figure 1.2: Screenshots of search results from Google, a commercial search engine. Search results captured on March 27th, 2020.

to use a search engine to find more information about the topic. The user enters the query "y combinator" in the search engine and examines the search results shown in Figure 1.2b, which are all related to the Y Combinator startup accelerator. Unsatisfied with every search result because of their irrelevance to the user's information need, the user abandons the SERP without clicking at any search results and decides to reformulate their query. This behavior of abandoning search results due to unsatisfactory results is common and accounts for 41% of all query abandonment (Diriye et al., 2012). Obviously, this behavior is considered negative and undesirable, because the search engine has failed to return information that would help the user satisfy their information need. This behavior is commonly known as "bad query abandonment", and is the focus of this dissertation.

1.3 Research Purpose

In addressing and learning more about bad query abandonment, we can reveal insight on user behavior during the search process in which IR designers can use to build a better and successful search experience. For example, with measurements such as how much time and how far people examine before deciding to abandon their query, we can produce insight into users' willingness to continue or discontinue examination of search results and understand what search results people examine before making their decision to reformulate. In this dissertation, we conducted three user studies to investigate the following:

Search result quality and query abandonment

- To investigate the rate of abandonment when users are presented with search results of various levels of quality.
- To determine how much time users spend before making their decisions to abandon search results.

Search results examination prior to query abandonment

- To determine the number and the order of search results people examine before abandoning their query.
- To test whether different types of searchers exhibit different examination patterns of search results.
- To investigate other factors that can influence search result examination and user's decision to abandon search results.
- To visualize users' gazing patterns under different types of SERPs.

The effect of non-relevance to query abandonment

- To observe the effect of different low quality search results on search interactions.
- To investigate the need for a broader definition of non-relevance.

1.4 Contributions

In this dissertation, we make the following contributions:

- By conducting a carefully controlled user study to investigate query abandonment, we found that users make quick decisions to click on a search result or abandon their query. We also found that the probability of query abandonment increases as the user has to search further down the ranked list to find a relevant document. (Chapter 3)

- Further analysis of the data indicates the possibility of two classes of users that behave differently. One group, which contains most users, seems to be focusing on the top of the ranked list to decide whether to abandon or not. The other group appears to be more likely to examine the whole ranked list. The group more likely to abandon their query is able to find answers quicker than the group less likely to abandon their query. (Chapter 3)

- Using eye-tracking, we conducted another experiment to determine what users examine prior to making their action to either click or abandon the results. We found that a user deciding to reformulate a query rather than click on a result is best understood as being caused by the user's examination pattern not including a relevant search result. If a user sees a relevant result, they are very likely to click it. However, users do not look at all search results, and their examination may be influenced by other factors. (Chapter 4)

- Besides search result quality and user type, we investigate how query quality (ambiguous and non-ambiguous) can play a role in influencing examination and users' decision to whether or not to abandon the search results. We found that type of query can be a factor in whether or not a user will abandon their query, and it influences different types of users differently. For example, if the query is considered somewhat ambiguous, some type of users stop their examination after determining the top three search results to be non-relevant. (Chapter 4)

- In mobile search, users are likely to scroll to view the first five results, but if a relevant result is not seen, they are more likely abandon their query. (Chapter 4)

- With data representing the rank of the top-most relevant result, user type, and query type, we build a decision tree model to interpret the search process. The decision tree provides a holistic view of the search process and abandonment, encompassing three important parts, users, queries, and search results, and shows the influence

of users and queries to each other at specific ranks in the search result. This has important implications on designing more comprehensive effectiveness measures that also include users and queries into the evaluation. (Chapter 4)

- Using the collected eye-tracking data, we demonstrate how time-series heat maps can help understand multiple searchers' behaviors over time. We show how the visualization can be useful in communicating differences between searchers' gaze patterns and complement traditional eye-tracking heatmaps. (Chapter 5)

- In a separate user study, we turn our attention to understanding user behavior when the search result page has no relevant documents. We show that users' interactions are influenced differently by the type and quality of the SERP presented to them. While every SERP shown to the user contained only non-relevant documents, the coherence and the nature of the non-relevant document in the SERP can influence how far down the ranked list users are willing to examine. (Chapter 6)

1.5 Thesis Overview

This thesis is organized as follows:

- **Chapter 2 – Background and Related Work:** This chapter comprises background in information retrieval and some of the related work in search behavior and query abandonment that our work is situated within or builds on.
- **Chapter 3 – Search Results Quality and Query Abandonment:** In this chapter, we explain our first experiment on studying query abandonment, where we looked into the effect of various degrees of *SERP quality* on the rate and time on people's decision to abandon queries in fact-finding tasks.
- **Chapter 4 – Patterns of Search Result Examination:** Our first experiment also left us with some unanswered questions, particularly what search results people examine, why people decide to examine more results, and what could influence their decision to process more search results or abandon the query. This chapter describes our second experiment where we used an eye-tracker to understand the examination behavior before users abandon their search queries.
- **Chapter 5 – Visualizing Searcher Gaze Patterns:** Eye-tracking can generate large amounts of data points. This chapter extends our previous work by proposing a visualization technique that incorporates timing information to quickly visualize and better understand what search results users examine at different time periods. Using the collected eye-tracking data, we show how the visualization can help communicate search examination behavior for different types of users and queries.
- **Chapter 6 – The Effect of Non-Relevant Results on Search Behavior:** In this chapter, we describe our third experiment. Instead of presenting users with either relevant or non-relevant search results, as in our two previous experiments, we broaden our notion of what it means for a document to be non-relevant. In particular, we investigated how SERPs with different types of non-relevant results affect examination, the rate of abandonment, and the time users spend to make their first action.
- **Chapter 7 – Conclusion and Future Work:** We conclude by summarizing the findings of our research and discussing possible future work.

Chapter 2

Background and Related Work

This chapters reviews the background and related work this dissertation is situated within or builds on. The review includes an examination of previous research in measuring and understanding user search behavior (Section 2.1), and previous work that specifically addresses query abandonment (Section 2.2).

2.1 Studying User Search Behavior

In this section, we review some of the methods used to study search behavior and provide a brief summary of their findings.

2.1.1 Mining Search Logs for Studying Search Behavior

Search engines are used by million of users, and have become a valuable tool for users to search for information. As such, search engine logs contain large amounts of data representing different types of user interactions. Analyzing these logs can provide useful information into the ways people interact with search systems to find information. Analyzing query logs is not uncommon, and have been used to understand search behavior in web search (Silverstein et al., 1999; Jansen and Pooch, 2001; Broder, 2002; White and Morris, 2007; Buscher et al., 2012), email search (Ai et al., 2017), job search (Spina et al., 2017) and other domains.

One of the earlier works in analyzing web search logs is the work of Silverstein et al. (1999). In their work published in 1999, Silverstein et al. used query logs of the 90's era

web search engine, AltaVista. The authors analyzed approximately 1 billion entries of search requests representing 285 million user sessions. With these query logs, the authors reported many descriptive statistics about users search behavior, e.g., the topics and queries the majority of people search for, the number of queries per session, the number of terms and operators in a query, etc. Some of the main findings of Silverstein et al.'s work is that web users typically use short queries when searching, mostly examine the first page of search results that contain 10 items, and rarely modify their initial query.

Like Silverstein et al. (1999), Broder (2002) also used AltaVista query logs for analysis on user search behavior, particularly in the type of searches people submit to search engines. Broder defined a taxonomy of queries based on users' intent that included three types called navigational, informational, and transactional. Navigational queries are associated with the intent of reaching a particular website, informational queries are to acquire information that the user feels is present on a particular page, and transactional queries involve an intent to complete a transaction, such as making a purchase or finding a map. Broder selected 1000 random queries from AltaVista's daily log, and after further post-processing and removal of non-English queries, 400 queries remained. These queries were manually inspected to determine the type of query. From this subset of queries, they found that navigational, informational, and transactional queries account for 20%, 48% and 30% of the queries, respectively. The three types of queries have become widely used as part of various user studies investigating user search behavior.

White and Morris (2007) used interaction log data from consenting users to understand search behavior of different types of users. In particular, the author examined differences in behavior between advanced and non-advanced searchers. The interaction logs consisted of 586,029 unique users who submitted millions of queries to three search engines – Google, Yahoo!, and MSN Search. In their work, advanced and non-advanced users were identified by their usage of search engines advance searching operators. An example of these operators are query modifiers such as '+' (plus), '−' (minus), and ' "" ' (double quotes) that are used to emphasize, deemphasize, and group query terms. The authors found significant differences in the behavior of advance and non-advanced users. For example, advanced users are more successful in their search and consistently visit more relevant pages than non-advanced users.

Buscher et al. (2012) used interaction log data from the Bing search engine to investigate how user and task differences impact users' examination behavior of the search result page. This data is more extensive than Silverstein et al.'s and includes more interaction types such as mouse movements, scroll, text-selections, mouse clicks, and others. With this data, clustering algorithms were used to identify groups of users who shared similar search interaction behaviors. Buscher et al. identified three meta-clusters centered on the amount

of time users spend inspecting the search result page: long, medium, and short. Interactions under the long cluster included detailed examinations of the results, high number of hovers and clicks on search items, lots of scrolling, and signs of reading behavior using the mouse cursor. Interactions in the medium cluster mainly differ from the other cluster in the number of abandonment. In the short cluster, the interactions include a shorter time on the search results page, quick mouse movements in a focused way, and inspection of few search results. Buscher et al. also looked into user clustering specifically under non-navigational search tasks, and found three distinct clusters of users. These were named economic, exhaustive-active, and exhaustive-passive users. Overall, economic users spend less time on the search page than exhaustive users. They also click quickly, and on average, click on less than one result per query. On the other hand, exhaustive users examine the search page in detail, exhibit more clicking behavior (both on hyperlinks and other areas of the page), and have a lower rate of abandonment. The difference between exhaustive-active and passive is that exhaustive-passive users spend even more time on the search result page, have longer cursor idle time, and abandon more often than the exhaustive-active users. The percentage of users in the economic, exhaustive-active, and exhaustive-passive clusters were 75%, 16% and 9%, respectively. These behavioral differences in user searching behavior, particularly the notion of economic and exhaustive users, were also identified by Aula et al. (2005), which we describe in details in Section 2.1.3.

One common theme in the research described above is that the researchers used logs of users interacting with real search engines. AltaVista, Bing, and the other search systems that were used were not intentionally modified to investigate specific search behaviors. Instead, the researchers used the interaction data of real searchers using real search engines. Access to these interaction data is often restricted and not publicly available for many reasons, e.g., user privacy concerns. While large-scale search log data can reveal a lot of useful knowledge, smaller scale search interaction data collected in user studies can also be used. We describe some of these user studies in the next section.

2.1.2 Manipulated SERPs for Studying Search Behavior

Collecting user search interaction with manipulated search results can be a useful method for understanding how users interact with different search results. This method was previously used by many researchers. Briefly, these manipulation methods work by changing certain aspect(s) of the search system that the researcher wants to investigate. In the context of search, this can be the number of relevant documents in the SERP, the length of search result snippets, the order of relevant search results, and so on. These manipulations are often unknown to the user while conducting their search. While users interact with

the search system and is shown the manipulated results, user data is being recorded for further analysis. We discuss some of the prior work that is relevant to this thesis below.

Many researchers have suggested using *Information Scent (IS)* (Pirolli and Card, 1999) to understand how users seek information on the web. Information scent is part of *Information Foraging Theory (IFT)* (Pirolli and Card, 1999). IFT implies humans' information-seeking behavior is similar to how animals use environmental cues to identify the most useful places to forage for food. In this theory, humans during information seeking "look for information from sources they believe are the most cost-effective by making predictive judgments using proximal cues" (Wu et al., 2014). Wu et al. (2014) used IS to study search behavior. In particular, the authors manipulated the number of the relevant documents in the search results of users first three queries', and asked users to search for relevant documents to open-ended question. They manipulated SERPs according to two within-subject variables: Information Scent Level (ISL) and Information Scent Pattern (ISP). ISL was defined as the number of relevant documents appearing in the first SERP of the task, and ISP as the distribution of four relevant documents in the SERP. For example, users are sometime shown a single relevant document positioned at the top of the SERP. In other cases, the user might be shown multiple relevant search results placed at different ranks. Both ISL and ISP included three categories. Low, medium, and high for ISL and persistent, disrupted, and bursting for ISP. These categories addressed different qualities of the SERPs. The authors found that around 42% of users abandoned their queries without any click on low ISL SERPs (where only the first document is relevant), and 13% of users requery on medium ISL SERPs (where only the top 3 documents are relevant). Only 1.6% of users requery on high ISL SERP (where only the top 5 documents are relevant). For tasks under ISP, they found no big difference in SERP abandonment between persistent ISP (relevant documents at rank 1, 2, 5, and 8) and disrupted ISP (relevant documents at rank 1, 2, 3, and 4). Persistent ISP and disrupted ISP had 10% and 12% SERP abandonment rate, respectively. Bursting ISP (relevant documents at rank 4, 5, 6, and 7) had 20% rate of SERP abandonment. Wu et al. (2014) found some factors that may influence query abandonment. The first factor was the properties of search results. The proportion and relative location of relevant results determines the quality of SERPs and further affect query abandonment. The second factor was the properties of the query. Users can learn new vocabulary from the current query result, and as a result, they issue a new query.

Ong et al. (2017) conducted a similar experiment to Wu et al. (2014) that primarily focuses on understanding differences in web search behavior for mobile and desktop users. In their experiment, users were shown SERPs with the same type of manipulation as in Wu et al. (2014). The authors found that certain search behaviours, such as query reformulation or number of document clicks, are less on mobile than on desktop environment, and desktop

users submitted more queries and saved fewer documents in lower positions. The authors attributed the differences in search behavior to the search environment in desktop and mobile. For example, because of different screen sizes in both desktop and mobile, search results visibility was affected. In their experiment, the mobile search interface allows three search results to be visible above the fold (i.e., visible without scrolling), whereas the desktop allows eight results. The authors also noted that mobile users may have a lower information need threshold, e.g., the number of relevant documents is restricted by the environment. While Ong et al. (2017)'s work is focused on comparing differences between search environments, their work emphasizes the importance of visibility of search results and its effect on search behavior, including query abandonment.

Joachims et al. (2005) conducted a user study where users were provided Google results to answer informational and navigational questions. Two examples of each type of question are "*Find the homepage of Michael Jordan, the statistician*" and "*Where is the tallest mountain in New York located?*". In their study, subjects were instructed to start their search with any query they would like and search as they would normally do while using Google or other commercial web search engines. The search result presented to the subjects was based on one of three experimental conditions. Either the results were not manipulated, manipulated by swapping the first two results, or by reversing the results order. They found that users are likely to click on higher ranking items irrespective of relevance and the performance of the search engine. The number of relevant results in the search list, however, is not controlled and could contain multiple relevant documents, which could be a possible influence to which document a user clicks.

Cutrell and Guan (2007) looked at how varying the amount of information in the search snippet affects user examination. They used both informational and navigational tasks similar to Joachims et al.'s work. In Cutrell and Guan's work, the authors manipulated the snippet length of search results to either be short, medium, and long. Short snippets contained about one single line of words, medium snippets about two to three lines, and long snippets typically six to seven lines of words. In their user study, participants were asked to use a custom search engine to find answers to navigational and informational questions. For each search task, an initial query was launched, and the manipulated search results for the query was presented to the participants. After launching the initial query, participants were free to use the search engine in any way they like. Cutrell and Guan (2007) found that increasing the amount of information in the snippets helps with informational queries but can hurt performance for navigational tasks. In another related work (Guan and Cutrell, 2007), the same authors conducted a study where they manipulated the search results to include what the authors described as "best" search result item and investigated the fraction of times participants looked at it. The placement of the "best" search result was

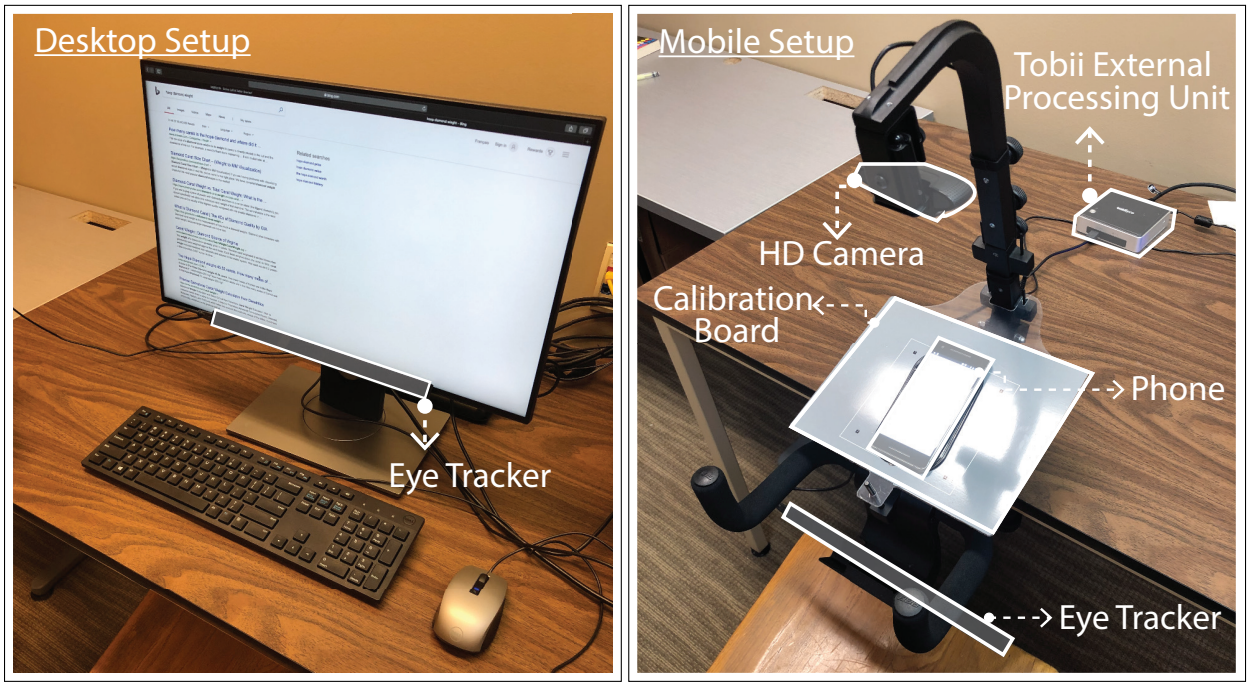
either at the top, middle, or bottom of the list. They report that as the rank of the "best" search result decreases from the top to the bottom of the list, the chances of users clicking at it decreases and may be related to their probability of examining it.

In mobile, Kim et al. (2017) conducted a user study where they manipulated snippet size to determine the appropriate size of snippets for mobile. Like Cutrell and Guan (2007), they had three types of snippets sizes (short, medium, and long) and two types of tasks (informational and navigational). Unlike Cutrell and Guan (2007), Kim et al. (2017) focused on mobile devices instead of desktop. The authors found that longer snippets resulted in longer search times with no better search accuracy for informational tasks. This was due to multiple reasons, including longer reading time to read longer snippets and more frequent scrolling. Overall, they suggest that it is best to serve snippets of two to three lines of text for mobile devices.

Other researchers have also manipulated SERP results to study user behavior with search entity cards (Bota et al., 2016) and images in aggregated search (Arguello and Capra, 2012). Somewhat relevant to our work in Chapter 6 is the work of Arguello and Capra (2012). Arguello and Capra looked into diversification in aggregated search (i.e., the task of combining search results from multiple search services such as images, news, and web documents in a single SERP). In particular, the authors focused on the coherence between two search components: images and web results. They conducted a crowd-sourcing experiment where participants accessed a custom search engine and were instructed to find answers to simple questions (e.g., *What is the latest album released by Seal?*). There were two layouts of the search engine they designed. One layout contained no images (i.e., no image-based search results), and the other layout contained images that were either all relevant to the search task, non-relevant, or mixed. The authors called these target and off-target results. For the example question above, the target sense is about the musician, whereas the non-target sense is about the animal. With these representations, the author looked into how the senses represented in the web results affect user interaction. They indeed found that senses represented in the web results affect user interaction. They also found that when web results are diversified, image results in the SERP has a significant effect on user interaction with the web results.

2.1.3 Eye-tracking and User Search Behavior

Studying where users look at provides insights into user attention and the information they process, which can help understand what affects user behavior and decision-making. Eye-tracking is a sensor technology that enables a computer to know where a person is looking



(a) Screen-based eye-tracker on desktop.

(b) Eye-tracker mobile device stand.

Figure 2.1: Example of screen-based eye-tracker

at a certain point in time. While many types of eye-trackers exist, we focus on video-based eye-trackers, which is the type most suited for online and usability experiments.

Eye-tracking Hardware

Video-based eye-trackers work by capturing images of the subject's eye. The sampling rate of eye-trackers, often between 50-250 Hz, indicates the number of images the eye-tracker can register per second. The larger the sampling rate, the more accurate the eye-tracker is in its ability to estimate the true location of where the subject is looking.

Two different types of video-based eye-tracker exists: wearable and screen-based. Wearable eye-trackers are suitable for experiments where the researcher tries to understand how subjects view and interact in the real world. Screen-based eye-trackers are mounted on a computer screen, and are commonly used for online experiments or usability testing to understand how subjects interact with interfaces.

Screen-based eye-trackers like the one shown in Figure 2.1a integrate with monitors and

laptop screens. The eye-tracker can also be mounted on tripods or mobile stands coupled with a scene camera for user studies involving tablets or mobile devices. Figure 2.1b shows an example of an eye-tracker used in a mobile stand.

For eye-trackers to work as accurately as possible, the user must first go through a calibration procedure. The purpose of the calibration procedure is to collect characteristics of the user's eyes and use them with calibration algorithms to calculate gaze data. This step is usually done by asking the subject to look at different gaze points placed at different locations within the screen. Recent eye-trackers are usually occupied with automated calibration that makes it easier for practitioners and researchers to quickly set-up and use. On recent eye-trackers, this process usually takes 3-5 minutes. It is important to note that the calibration process may not always be successful. In certain situations, the eye-tracker may not be able to capture the user's eyes. This can be due to various reasons such as the lighting environment, subject's height, wearing eyeglasses, having long eye-lashes, or wearing mascara.

The two basic elements of eye movement are called fixations and saccades. Fixations are the most common eye movements that researchers analyze to make inferences about the subject's cognitive process. In short, fixations are when a user's gaze stops scanning and focuses in a certain area, typically for 200-300 milliseconds, to process what is being seen. Saccades, on the other hand, are the movement of the subject's eyes between fixation points, and can be used for visualizing the eye's scanning path.

When using an eye-tracker, the eye-tracker generates a dataset of coordinates that can be visualized and interpreted to expose user behavior. The dataset typically includes:

- An order list of eye movements and their coordinates. This can be useful to determine the subject's sequence of examination.
- The type of eye movement and a timestamp of when a particular eye movement occurred. This data can be helpful to determine the time to the first examination of a particular area in the screen.
- The time length of a particular fixation. Timing information can be useful to determine how long people read or examine particular elements.
- Total fixations per element or area of interest (AOI). The number of fixation can be useful to determine engagement or distractions.

Figure 2.2: Heatmaps of user eye-tracking study on search engines.

Eye-tracking in Web Search User Studies

One of the benefits of using an eye-tracker in information retrieval is the ability to know what the user is looking at before performing an action. Eye-tracking enables us to determine what the user has examined or skipped in the order of occurrence. This data can be useful for understanding the process in which people reach their decisions while searching, e.g., to either click on a document, examine more documents, reformulate their query, or quit the search process. Several researchers have used eye-tracking as part of their studies on how users interact with search results (Joachims et al., 2005; Guan and Cutrell, 2007; Cutrell and Guan, 2007; Liu et al., 2014; Eickhoff et al., 2015; Aula et al., 2005; Dumais et al., 2010; Klöckner et al., 2004a; Granka et al., 2004; Hofmann et al., 2014). Granka et al. (2008) reports that two of the main research findings in the eye-tracking and information retrieval literature falls under two categories: overall viewing pattern and individual user differences. We review these findings next.

Overall Viewing Pattern

In an eye-tracking user study done by the usability consulting company Nielsen Norman Group, researchers found that the majority of people read web pages in an F-shape manner. The F-shape consists of three components: reading the information on the top of the page in a horizontal manner, then reading horizontally in a slightly lower position on the page, and lastly scanning the left side of the page's content vertically. This F-shape pattern is also sometimes called the Golden Triangle. Figure 2.2 shows an example of this viewing pattern.

This viewing pattern shows that the majority of people do not read the information on a web page word-by-word, and instead focus their attention on information positioned at the top of the webpage. In the context of search engines, the results show the importance of placing relevant information in the top search results. For example, Granka et al. (2004) used eye-tracking and showed that users spend more time and attention to top-ranked results and that they generally work top to bottom when looking for relevant documents. In addition to spending more time on top-ranked results, researchers have also found that users are biased towards clicking on top results (Joachims et al., 2005, 2007; Lorigo et al., 2008).

While the F-shape viewing pattern can be considered a generalization of how people view SERPs, other research indicates that users' viewing pattern on SERP is actually more complex and is influenced by the type of user and search task. For example, internal research by Google suggests that tasks and users can impact the way in which a results page is viewed. In their work (summarized in Granka et al. (2008)), 32 users were provided with a SERP for the query "*tallest active player NBA*", and were instructed to find the answer to the question. The figures in Granka et al. (2008) show a heatmap of the aggregated users reading pattern, and two different styles of examining SERPs. The heatmap figure in Granka et al. (2008) shows a slight resemblance to the F-shape. The two scan path figures show the scanning path of two different users on the same SERP, clearly indicating different searching styles. One scan path shows a searching style of a user examining less than two results and the other scan path is of a user examining the SERP more exhaustively and spending time viewing more than three results.

Differences in Users Viewing Pattern

Using scan patterns visualization, Granka et al. (2008) show show how different users have different examination patterns, even when users are presented with the same page. The figures suggest that users may employ different searching strategies while processing the

SERP. For example, one user in Granka et al. (2008) is quickly scanning the results before making their decision, whereas another user in processes the SERP more exhaustively and carefully. These user differences have been noticed by other researchers as well.

Using eye-tracking, Klöckner et al. (2004a) classified users into two groups based on how they processed search results. One group followed a "strictly depth-first" strategy where they work down the ranked list one result at a time. The remaining participants followed either "partially breadth-first" or "extreme breadth-first" strategies. A partial breadth-first strategy is reflected by looking ahead at a few results and making comparisons between the results to determine what result to click. The extreme breadth-first approach involves studying all of the search results before deciding which result to click.

Aula et al. (2005) conducted an eye-tracking experiment to study how people evaluate search result pages. Aula et al. selected 10 query results such that three of them had no relevant documents, three had more than five relevant documents, and the remaining four were mixed. They recruited 42 students to participate in the study. To analyze how people examine the SERP, they developed a unique static visualization that presents the order in which each search result was visited. Aula et al. printed out the visualizations and manually inspected them to determine any patterns in how people evaluate SERPs, and to group the visualizations accordingly. The visualization show the order in which search results were examined, with circles that corresponds to the time a has user has spent on each search result. Examples of these static visualization are shown in Aula et al. (2005). Like Klöckner et al. (2004a), Aula et al. (2005) found users to follow either an "economic" or "exhaustive" strategy for processing search results. In Aula et al. (2005)'s study, about 6-7 summaries were visible at a time on the computer screen, and *economic* users would scan at most the first three results before acting. The *exhaustive* users would examine more than half of the visible summaries and sometimes even scroll to see the remaining summaries before acting. Aula et al. (2005) found that the *economic* searchers had more computer experience and would fixate for shorter periods on each result.

Similarly, Dumais et al. (2010) found three groups of users, and following the convention of Aula et al. (2005), named the groups: "economic-results", "economic-ads", and "exhaustive". Dumais et al.'s study involved a commercial search engine and two economic groups that differed in how they examined advertisements. A significant difference between the economic and exhaustive groups was the amount of time spent examining result summaries. The economic users spent between 8.7 and 9.9 seconds while the exhaustive users spent 14.6 seconds on average.

Some users may display exhaustive behavior as a result of being dyslexic, for MacFarlane et al. (2017) have found that dyslexic users are more likely to backtrack and reread the

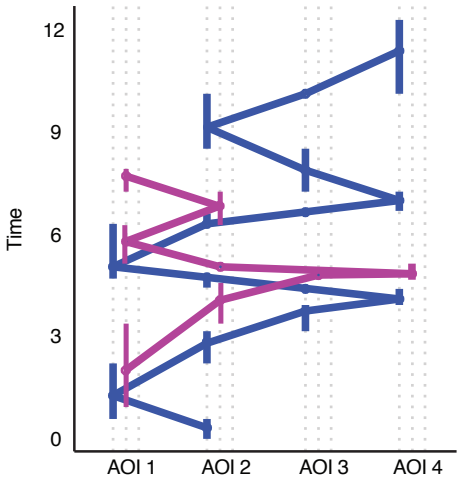


Figure 2.3: Example of two users AOIs examination behaviors. Y-axis indicates time. Based on Raschke et al. (2012)

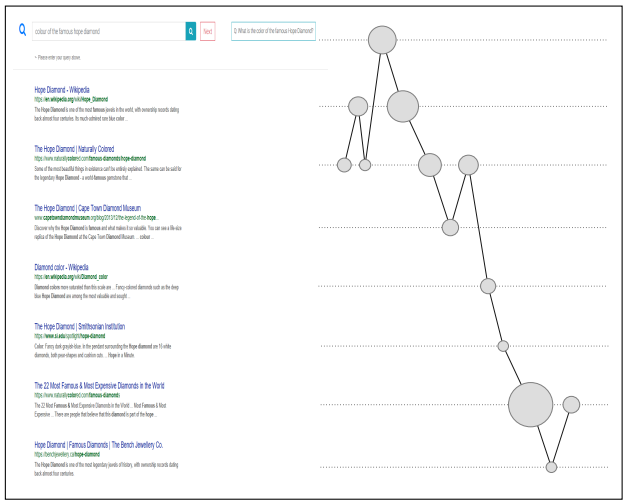


Figure 2.4: An example visualization of single user examining behaviour on a SERP. X-axis indicates time. Based on Räihä et al. (2005)

material. Palani et al. (2020) conducted an eye-tracking study of web search by people with and without dyslexia. Palani et al. confirm that searchers with dyslexia have different gaze patterns and search behavior that reflects their struggle during at different stages during the searching process.

Visualizing User Search Behavior

Many of the work using eye-tracking include some visualization of fixations data (Dumais et al., 2010; Wang et al., 2016; Balatsoukas and Ruthven, 2012; Hofmann et al., 2014). Eye-tracking heat maps overlaid on thumbnail images of SERPS are widely used to visualize searcher gaze patterns and understand search behavior. Often these heat maps only show fixations for individual searchers and do not provide timing information (e.g., Figure 2.2). For example, Dumais et al. (2010) use heat maps to illustrate individual differences in gaze patterns. Both Liu et al. (2015) and Wang et al. (2016) use heat maps to provide examples of individual searchers interacting with search verticals, such as images, news, shopping, and maps. Similarly, the heat maps in Wang et al. (2016) illustrate whole-page interactions of individual searchers, including verticals and other elements. Balatsoukas and Ruthven (2012) overlay SERPs with fixations and other information, similar to heat maps.

In addition to providing specific examples of searcher behavior for illustrative purposes, heatmaps can be used to summarize outcomes from an experiment by overlaying fixations from multiple searchers. For example, Buscher et al. (2009) used a heatmap to display fixations from 20 participants in their experiments. Papoutsaki et al. (2017) used heat maps both to provide examples of individual interactions and to summarize the interactions of many searchers. Although heat maps provide an overall understanding of gaze patterns, they do not provide timing information.

Besides heatmaps, eye-tracking scan paths consist of an arbitrary number of fixations overlaid on top of the interface. Each fixation point is depicted by a circle, connected by saccades, which are depicted by lines. These fixations points are numbered and described by their coordinates on the screen. The number on each fixation point indicates the order in which the fixation occurred. Scanpaths are useful to visualize a single user's eye-tracking data, but can be hard to visualize when adding multiple users' data on top of each other. Nonetheless, scan paths are widely used as a method for visualizing and analyzing eye-tracking data and have been used as part of different IR research (Clark et al., 2012; Bhattacharya et al., 2020).

There exist other methods of visualizing eye-tracking data besides heatmaps (Räihä et al., 2005; Raschke et al., 2012). Raschke et al. (2012)'s visualization technique can be used to display a visual scan path of multiple users while incorporating time into the visualization. The y-axis indicates time, and the x-axis indicates the list of areas of interest (AOIs) being investigated. The color of the line indicates users. The scan path of the user changes as time passes and the duration of the fixation at each AOI is indicated by the vertical length of the line. While the visualization is useful for visualizing a few users' scan paths, a larger number of searchers increases the number of scan paths and could introduce visual clutter. Figure 2.3 shows an example of the visualization technique.

Räihä et al. (2005) proposed a static technique for visualizing gaze data from single users while incorporating some aspect of time. With the AOIs displayed on the left as the y-coordinate, and the x-coordinate denoting a relative point in time, the points in the plot indicate the fixation length and the visiting order of the AOIs. Räihä et al. (2005)'s technique works when AOIs are built in linear order. This technique is useful for visualizing a single user data, but will clearly result in a visual clutter as more users are added to the plot. As described earlier, this visualization technique was used in Aula et al. (2005) to uncover the different examination behavior of different users. Figure 2.4 shows an example of the visualization technique.

2.2 Query Abandonment

2.2.1 Types of Query Abandonment

Although the word "abandonment" has negative connotations, query abandonment can sometimes be a desirable user action. As Li et al. (2009b) highlights, good query abandonment occurs when users, for example, find the answer they were looking for in the search results summaries or located somewhere on the SERP. Good query abandonment can, therefore, be considered a signal of success. In other words, the search engine has succeeded in finding and presenting the relevant information directly in the SERP, without requiring the user to click at any of the search results. Good query abandonment is especially common in mobile search and on queries potentially indicating good abandonment, such as queries seeking a weather report or a listing of local address (Li et al., 2009b). As a proxy for increasing user satisfaction, commercial search engines aim to increase the rate of good abandonment (Williams et al., 2016), with some efforts to incorporate this behavior as part of search metrics to indicate success (Khabsa et al., 2016).

The other type of query abandonment is termed "bad query abandonment", and is associated with the user being dissatisfied with the search results and therefore abandoning the SERP. After the user submits a query and the SERP is displayed, the user begins processing the search results. Influenced by the irrelevance of the examined search results, the user may abandon the search result without further examining any search results and decide to reformulate their query in the hopes of receiving a better SERP. This behavior of abandonment resembles user dissatisfaction (i.e., dissatisfied for not finding the relevant documents), and is the most common scenario behind all query abandonment (Diriye et al., 2012). Stamou and Efthimiadis (2010) show that approximately 50% of abandoned queries are queries with non-relevant results that have negatively influenced users. In this dissertation, we focus on bad query abandonment.

2.2.2 Query Abandonment Rationales and Prediction

To understand why people abandon their queries, Stamou and Efthimiadis (2009) employed a survey to study search tasks with query abandonment. The authors categorized the causes of abandonment as *intentional* and *unintentional*. Intentional causes are encountered with a deliberate intention to look for answers in the search results' snippets, and unintentional causes can be due to irrelevant results, already seen results, or interrupted search. Diriye et al. (2012) extended Stamou and Efthimiadis (2009) work by conducting a much larger

user study that collected abandonment rationals at abandonment time using a browser plugin that prompts participants with survey questions right after a query is abandoned. Participants in the survey included 186 people from within the Microsoft Corporation's campus. Diriye et al. (2012) found that the majority of abandonment were caused by dissatisfaction of the search results (41%), followed by satisfactory reasons to abandon, e.g., relevant information presented directly in the search result (31%). The authors also found 27% of abandonment is not due to satisfaction nor dissatisfaction with results. The reasons of abandonment were: users came up with a better query before they viewed the SERP (13%), users found search results not sufficiently important (3%), and the user got interrupted by some factor (1%) (e.g., network failed and tab closed). Some 10% of the reasons fell into a catch-all "other" category. Both Diriye et al. (2012)'s and Stamou and Efthimiadis (2010)' employed survey questionnaires as their methodology to uncover reasons why people abandon their queries.

Diriye et al. (2012); Song et al. (2014); Brückner et al. (2020) investigated methods for predicting abandonment rationales. Being capable of accurately predicting abandonment rationales has implications for the design and evaluation of search engines. For example, the rate of query abandonment and its predicted rationale can be used as a supplement metric to evaluate the performance of the search engines, along side other existing metrics. Diriye et al. (2012) generated multiple features set that are then used to build different binary classifiers to predict whether an abandonment falls under satisfactory (SAT), dissatisfaction (DSAT), unintentional, and other. Around 2,000 features were generated and were divided into five categories: (1) session, (2) query, (3) search result, (4) hyperlink-click and dwell, and (5) cursor. The authors showed a breakdown of the impact of each feature category on the classifier performance. For example, using cursor-based features, which capture aspects of how people examine the SERP, yield reasonable prediction performance, especially for DSAT abandonments. The authors also show that accurate prediction of SAT and DSAT abandonments is achievable with only session, query, and search result-based features and excluding post-query features such as clicks, dwell time and cursor features.

Using the same data collected by Diriye et al. (2012), Song et al. (2014) used contextual information from user search sessions to build an Support Vector Machines (SVM) based classifier. The information include query features (i.e., the length of the query), SERP features (i.e., the total number of answers shown in the SERP) and session features (i.e., the total session length in terms of queries or query dwell time). Unlike Diriye et al. (2012), the authors do not include historical features such as overall query frequency, which can be obtained by having search logs with a longer-period. The results show that their SVM models substantially outperformed the boosted decision tree classifier which Diriye et al. (2012) reported as the best of all classifiers they tried.

More recently, Brückner et al. (2020) used mouse movement data to train recurrent neural networks for predicting good and bad abandonment. The author used data from a previous crowdsourcing experiment where participants where asked to search for answers to simple questions (e.g., "*How old is Brad Pitt?*") and were shown knowledge graph[1] (Navalpakkam et al., 2013) in the SERP. Knowledge graphs are often presented on the right side of the SERP and help users discover new information quickly and easily.[2] To distinguish between good and bad abandonment, the authors considered a query to be good abandonment if the user noticed the knowledge graph and marked it as useful, otherwise it is considered as a bad abandonment. Using mouse coordinates collected while participants interact with the SERP, the author show that predicting the type of abandonment can be efficiently done using recurrent neural networks that take mouse coordinates as input. Their experiment illustrate that distinguishing between the type of abandonment can be done with good accuracy without engineering many or sophisticated features.

2.2.3 Studying Query Abandonment

One of the difficulties in studying bad query abandonment arises from the nature of this behavior. Users initially start their search process with the intention to succeed in their search rather than fail or quit. As a result, users may be driven to click on a search result that appear somewhat promising, even when it is ultimately considered not helpful. Bad query abandonment, therefore, may be considered less common than other types of search behaviors. While it may be difficult to drive users to naturally make this type of behavior, it can still be studied in different ways. We list and briefly explain some of the existing approaches to study query abandonment in previous literature.

- **Surveys:** Survey questions can be designed to study query abandonment in particular. Stamou and Efthimiadis (2009) employed a survey to study query abandonment, with a questionnaire to understand the causes of query abandonment and when it happens. While surveys can be an excellent method to get responses quickly and easily, it can have few drawbacks. First, it depends on users' memory to remember when and how their query abandonment has occurred, and second, it lacks important data on user behavior while interacting with the search engine.

- **In-situ questionnaires:** An alternative method to surveys is to develop an in-situ questionnaire, i.e., questionnaires that are asked right after a user abandons their

[1] Knowledge graph are also called by other researchers as entity cards or knowledge module.
[2] https://blog.google/products/search/introducing-knowledge-graph-things-not/

query while using a search engine. Diriye et al. (2012) employed this method using a browser extension that tracks search behavior and pops up with questions once the user abandons a query. This method allows for collecting user behavior data and search interactions before the user abandons their queries. In Diriye et al. (2012) work, participants were told to install the browser extension and perform search tasks as they would normally every day. Participants search tasks and search results were not controlled. While their method is useful, it does not allow a more refined understanding of abandonment under certain types of search tasks or quality of search results, e.g., users' actions under a certain level of search results quality.

- **Controlled experiments:** Rather then employing questionnaires to study abandonment, some researchers conducted controlled experiments to study user search behavior, including query abandonment. For example, Wu et al. (2014) conducted controlled user study in which participants had to complete several search tasks. In their experiment, search results of users queries were intentionally manipulated to show controlled types of search results predetermined before the study. The manipulation technique Wu et al. (2014) used was designed to understand how participants interact with search results with varying amount of relevant items. Similarly, in Joachims et al. (2005) study, search results were manipulated such that each participant would be shown search results in reverse order, or in the standard order but with the first two ranks swapped.

2.3 Summary

In this chapter, we provided relevant background on conducting research on user search behavior. We described some of the work related to query abandonment and the research methods used to study this behavior. Prior work studied query abandonment using surveys, in-situ questionnaires, and controlled experiments. The controlled experiments, such as Wu et al. (2014); Joachims et al. (2005); Guan and Cutrell (2007) and other, looked into query abandonment as part of their work. However, to the best of our knowledge, there are not any work that solely focused on studying query abandonment under SERPs of varying qualities. Unlike Joachims et al. (2005); Wu et al. (2014) and others, where many of the search tasks in their user studies include multiple relevant documents in the SERP, our work in this thesis is focused on understanding possible causes to how far users are willing to examine SERPs with either no relevant documents or one relevant document placed at different ranks.

Chapter 3

Search Results Quality and Query Abandonment

Our first study investigates the effect of SERPs of different quality on the rate and time to abandon search results. In this work, we look into answering the following questions: When a user enters a query and is presented with a SERP that contains a relevant search result placed at the top of the list, what action would they make? Would the user click on the relevant search result or abandon the results? What if the top most relevant search result is placed at a lower rank? How much time does it take for people to make a decision to either click on a search result or abandon their queries? We set up a user study to understand the behavior of query abandonment under controlled SERPs of different qualities.

3.1 Introduction

Today's search engines are typified by interfaces that allow a search user to issue a text query and then receive a list of search results. The moment the search engine results page (SERP) is displayed, the user begins processing that page with a goal of making one of three decisions:

1. Click a search result to navigate to its page for viewing.
2. Abandon the query, but continue the search by reformulating the query to produce a new search results page.

3. Abandon not only the query but also the search. The next interaction with the search engine will not be a continuation of the current search.

Modern web search engines not only return organic search results, but also advertisements and other possible interaction mechanisms, for example, other suggested queries. In this work, we limit our discussion to an abstract search engine that only returns organic search results in a ranked list, and where each search result is displayed with a summary to aid the user in deciding on the result's relevance.

While both choice 2 and 3 can be considered an abandonment of search results, our work in this chapter focuses on choice 2, i.e. a query reformulation without any clicks on search results. While a user performing a query abandonment does not click on any search results, the user will spend some time to view the search results and reformulate the query.

Query abandonment means that the user effectively places zero value on the search results. Even if the search results may contain relevant results, the query abandonment means that the user has spent time on the page but remains unsatisfied. If a user found significant value in the search result summaries, we assume the user would either click on a search result or abandon the query satisfied. Given the apparent loss in value to the user that results from a query abandonment, it is important to understand what conditions make abandonment likely. In particular, how good do search results need to be to have at least one click and avoid being treated as worthless with a query abandonment?

We conducted a controlled user study to investigate the relationship between search results quality and the behaviour of query abandonment. In our study, we asked participants to find the answers to a set of questions. The questions were selected to be simple to answer given a good search engine, but unlikely for our study participants to already know the answers. For example, one question was *"How long is the Las Vegas monorail in miles?"* We varied the quality of the search results by placing one relevant document at varying ranks. We selected the non-relevant search results to appear somewhat plausible as search results for the given question, but to also be clearly non-relevant on inspection.

3.2 User Study

In this section, we describe details of our user study. That includes: the search tasks used in our user study, how we control the quality of search results, the study design and procedure, the search interface, and information about our participants.

3.2.1 Search Tasks

We asked each participant in our user study to search for answers to 12 factoid questions. The list of questions used in the study are shown in Table 3.1. For each search task, we provided participants with a single question and asked them to use our search engine to find an answer to the question using our custom search engine. Participants could enter as many queries as they wanted and spend as much time as needed to find the correct answer using our search engine.

We designed the questions to meet the following requirements:

- Most participants should not already know the answer, and thus, participants would be forced to search to find an answer. While it is difficult to determine which questions might be known to participants, we choose the question based on what we believe is uncommon to most people. We also included a pre-questionnaire to ask participants if they knew the answer to a particular question before conducting their search.
- The question should be straightforward and answered easily with the help of a modern search engine. Complex question are known to have different search behaviors, but in this work, we focus on factoid questions.
- Each question should only have one standard correct answer. The reason for this is to not confuse people of different possible answers while they are searching, which could be a confounding variable.
- The question should allow easy retrieval of plausible non-relevant search results and a relevant web page containing the answer. The reason for this point is to be able to construct SERPs of different qualities. How these different qualities are constructed is mentioned in the next section.

3.2.2 Controlling Search Results Quality

Manipulating SERPs can be a useful method to study search behavior. Figure 3.1 illustrates some of the possible ways SERP can be manipulated. In this work, we manipulated the number and the order of relevant documents in the SERP.

For each search task a participant performed, we returned a manipulated SERP, i.e., treatment. Each treatment consists of a different manipulation of SERP quality:

Table 3.1: The 12 search task questions and their corresponding answers and trigger query terms. The first query for the task that contains any of these terms will elicit the manipulated SERP to be presented to the participant. Question with ID "P" is used as the practice question shown to participants in the practice interface of the user study.

ID	Question	Answer	Triggered Query Words
P	What is the weight of Hope Diamond in carats?	45.52	N/A (practice question)
1	How long is the Las Vegas monorail in miles?	3.9/4 miles.	Las, Vegas, monorail
2	Find out the name of the album that the Mountain Goats band released in 2004.	We Shall All Be Healed	Mountain, Mountian, Goats, Goat, album
3	Which year was the first Earth Day held?	1970	Earth, Day
4	Which year was the Holes (novel) written by Louis Sachar first published?	1998	Holes, hole, louis, sachar, Novel
5	Find the phone number of Rocky Mountain Chocolate Factory located in Ottawa, ON?	(613) 241-1091	Rocky, Mountain, Chocolate, Factory, Ottawa
6	What is the name of opening theme song for Mister Rogers' Neighbourhood?	Won't You Be My Neighbor?	Mister, Rogers, Roger, Roger's, Neighbourhood, opening, theme, song
7	Which album is the song Rain Man by Eminem from?	Encore	Rain, Man, Eminem
8	How many chapters are in The Art of War book written by Sun Tzu?	13	Art, War, Sun, Tzu
9	What is the scientific name of Mad cow disease?	Bovine Spongiform..	Mad, Cow, Disease
10	How many campuses does the University of North Carolina have?	17	University, North, Carolina, Campus, campuses, UNC
11	Which Canadian site was selected as one of United Nations World Heritage Sites in 1999?	Miguasha National Park	United, Nations, World, Heritage, UN
12	How many times did Michael Jordan play the NBA All-Star Games?	13	Michael, Jordan, NBA, All-Star, Star

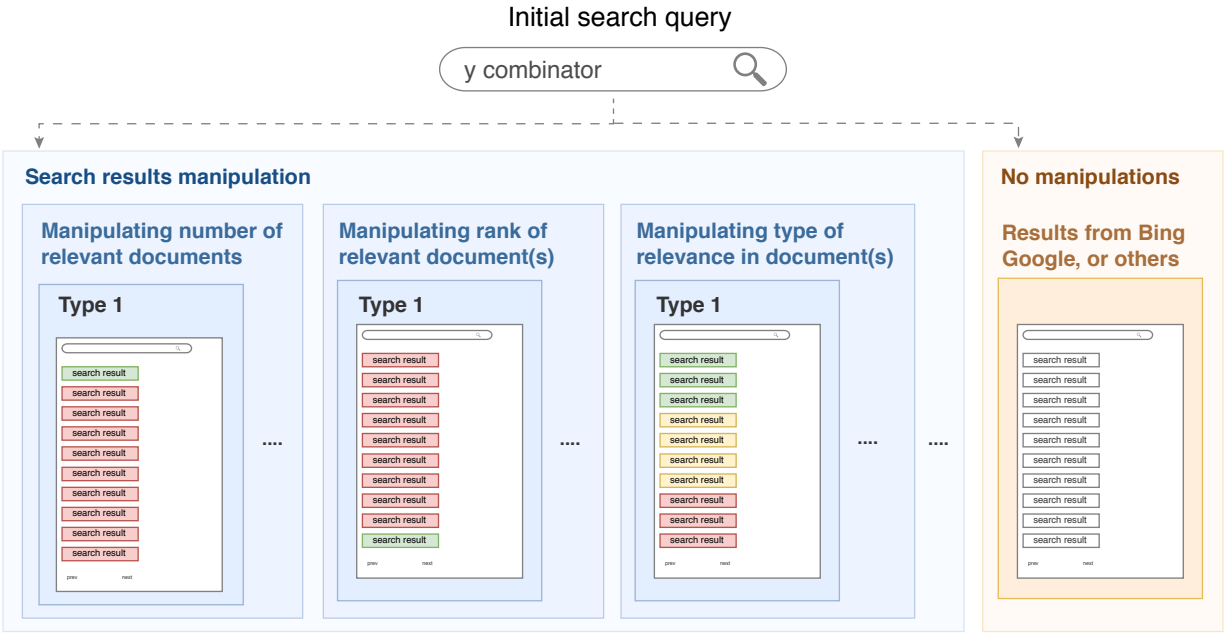


Figure 3.1: Possible manipulation techniques to search results after a user submits a query. In this work, we focus on the first two types: controlling the number of relevant documents shown to the user, e.g., 1 relevant document is placed at the top of the SERP, and controlling the rank of relevant documents, e.g., placing relevant documents at the bottom of the SERP.

- For ten of the treatments, the SERP contained 1 relevant result and 9 non-relevant results. A relevant result contains the correct answer on the corresponding web page. We placed the relevant results at ranks 1-10 and denoted these tasks as **Correct@1, ... Correct@10**.
- For one treatment, the SERP contained 10 non-relevant results and we denote this task as **NoCorrect** (`NC` for short).
- For one treatment, the SERP result contained results returned by the Bing API[1] without any manipulation, denoted as **Bing**.

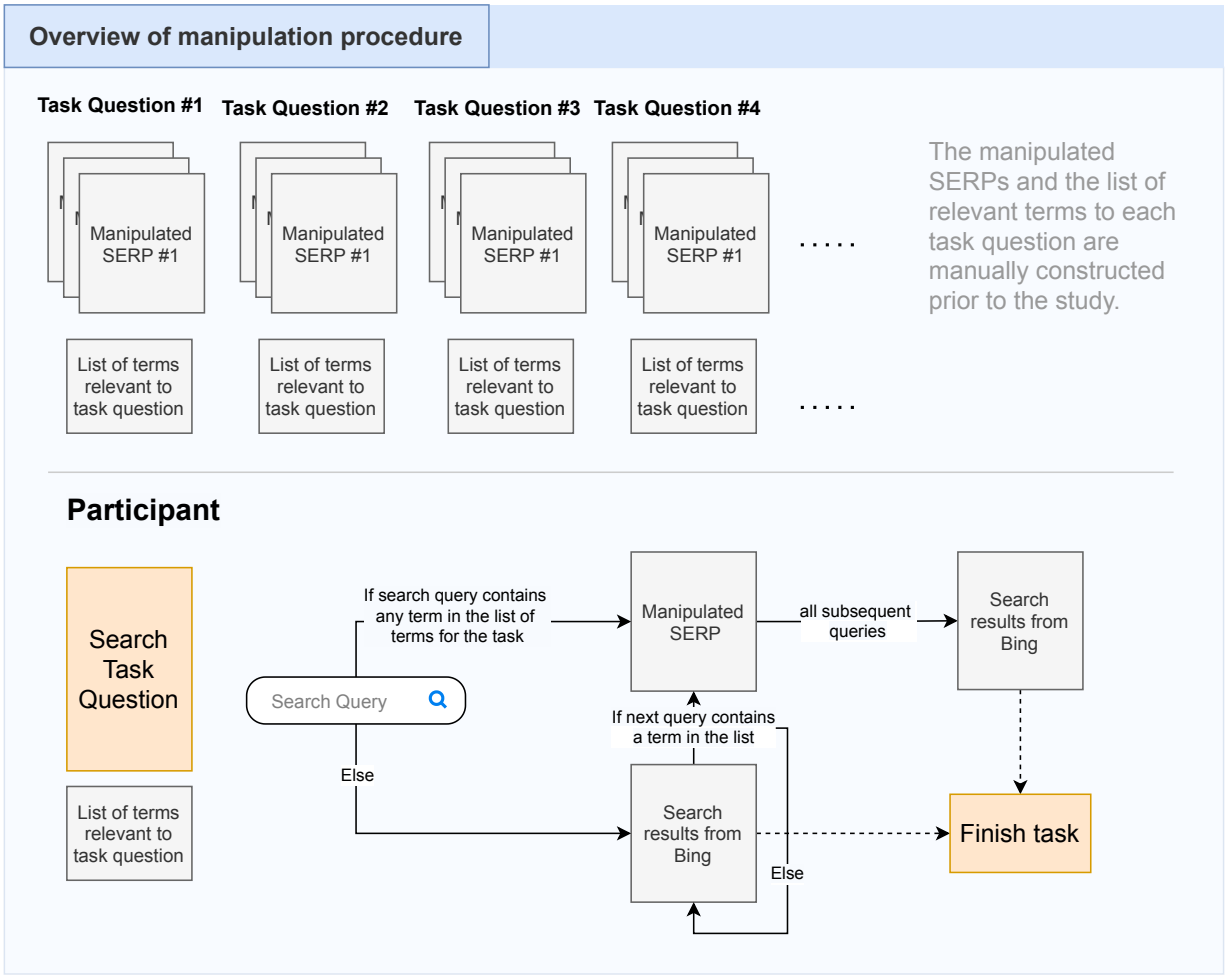


Figure 3.2: An overview of how the manipulated SERP are constructed and shown to participants in our user studies.

3.2.3 When are Manipulated SERPs Shown?

We wanted to only show the study participant the manipulated SERP if the participant entered a query that could reasonably be an attempt to use a search engine to find an answer to the given question. For each search task question, we constructed a list of terms that if any of them were entered by the participant as part of their query, the interface would trigger the manipulated SERP. If the participant entered a query lacking all of the selected keywords, we would use the query to request original search results from the Bing search API and present them to the user.

For each search task, the participant can only trigger the manipulated SERP once, and only after the participant submits a query with any terms that we deemed to be relevant to the current search task's question. All further queries will not trigger the manipulated SERP to show, regardless of the query terms. All queries following the display of a manipulated SERP produce live, original results from the Bing search API. Figure 3.2 shows an overview of how the manipulated SERP are constructed and shown to participants in our user studies.

For example, question #8 which asks for the number of chapters in the Art of War book by Sun Tzu has the following relevant query terms: `Art`, `War`, `Sun`, `Tzu`. We constructed relevant terms for each question prior to the study. If the participant entered a query lacking all of the selected keywords, we would send the query to the Bing search API and return original results. The list of trigger terms for each question is shown in Table 3.1.

For each search task, the participant can only prompt the manipulated SERP once. All further queries will not prompt the manipulated SERP, regardless of the query terms. All queries following the display of a manipulated SERP produce results from the Bing search API.

For the control search task, all queries are sent to the Bing search API, and the results are then shown to the participant.

3.2.4 How are Manipulated SERPs Constructed?

Our search engine only provided 10 search results in response to a query. With 10 search results and simple binary relevance, there are 1024 (2^{10}) possible ways to construct search results to vary their quality. In this work, a relevant document contains the answer to the user's question and a non-relevant document does not contain the answer. To simplify our

[1] `http://www.azure.microsoft.com/services/cognitive-services/bing-web-search-api/`

study, we decided to focus on the placement of a single relevant document in a ranked list of 10 search results. Placing the single relevant document at ranks 1 through 10 yields 10 different rankings, assuming that the relevant document placed in lower ranking would results in a lower search quality for the user. We also produced a ranking where all 10 documents were non-relevant. Finally, we also had a control condition where the search results were the original results produced by the Bing search API in response to the user's query.

With our single relevant document and our set of non-relevant documents, we constructed manipulated SERPs as follows:

- For treatments **Correct@1, ...,** and **Correct@10**, we placed the relevant document at the corresponding rank and randomly filled the rest of the results with our non-relevant documents.
- For the **NoCorrect** treatment, we randomly positioned the 10 non-relevant documents in the SERP.

In order to reduce the chances of participants noticing the manipulations as they are completing their tasks, we included search tasks (denoted as **Bing** treatment) that have no search result manipulation. In these tasks, we use the Bing API to return results to the queries submitted by the user. The purpose of these tasks is to have the participants feel like the search engine being used in the experiment is reasonable, and to have it used as a comparison with other experimental conditions. Throughout this work, we use the term relevant and correct SERP result interchangeably to indicate the relevant document with the correct answer.

All search results shown in manipulated SERPs contained at least one keyword from the task's question. Relevant, or *correct*, documents provided a straightforward answer to the user's question that should be easy for the user to find. Non-relevant, or *incorrect*, documents contain keywords from the question, and may be related to the question in some way, but their overall topic is clearly non-relevant. A non-relevant document does not contain the answer.

We found all documents and their snippets by manually issuing queries to the Bing search API. For documents with the correct answer in their snippets, we manually removed the answer from the snippets to influence the user to click on the document and find the answer from its content. If the snippet contained the answer, the user might abandon the query because they have already found the answer (e.g., good abandonment). We only

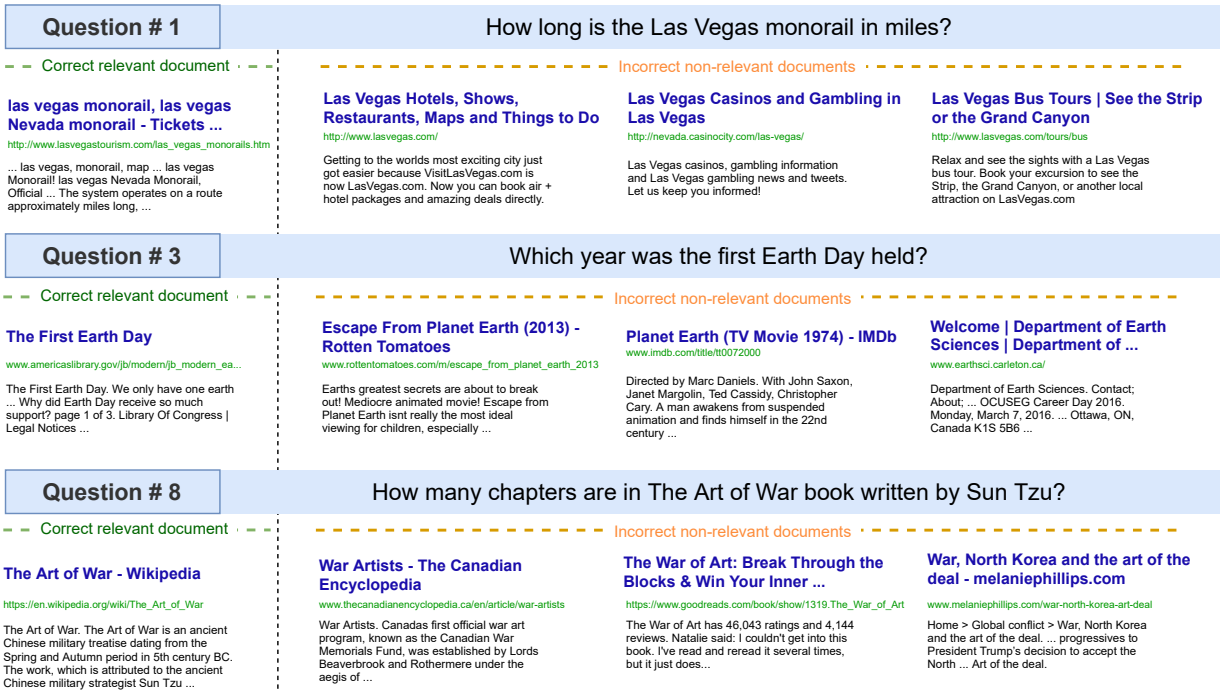


Figure 3.3: Example of search results in our manipulated SERPs for three search tasks.

controlled the snippet content for the manipulated SERPs. The control SERP (Bing) used snippets directly from the Bing API, and can contain direct answers in the search results snippets.

In order to make the manipulated SERPs look realistic and reasonable, and to prevent participants from having any suspicion or confusion regarding the SERP, the incorrect documents were selected from queries with terms in the corresponding factoid question, for example, the "Las Vegas Monorail" question shown in Figure 3.6 (ID 1 in Table 3.1). For this question, a somewhat realistic but unrelated query would be "Las Vegas Casino" or "Las Vegas Hotel". Both queries have the phrases "Las Vegas" but are not relevant to Las Vegas's monorail. For the question on the the Art of War chapters, non-relevant documents can be about books with similar titles and different authors. Such documents contain relevant words but their content is not relevant to the question. Figure 3.3 shows examples of search results in our manipulated SERPs for both the correct and incorrect documents. We used such queries to retrieve incorrect documents for all 12 questions.

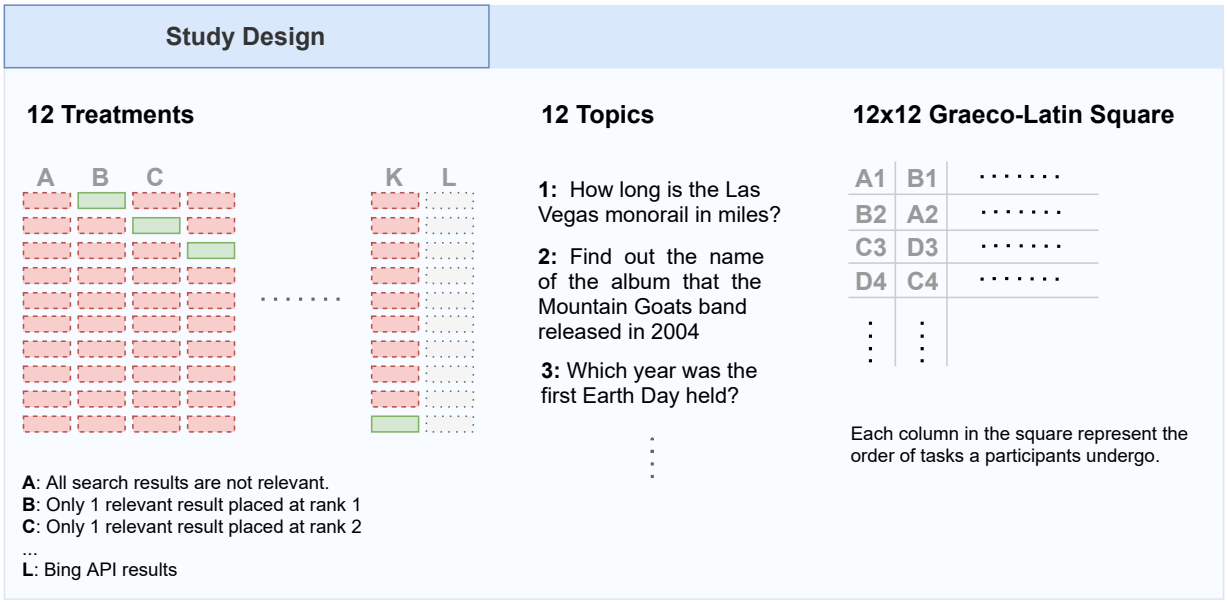


Figure 3.4: User study design.

3.2.5 Balanced Design

Figure 3.4 shows an overview of the study design. In total, there are 12 different treatments and 12 different topics. We used a 12×12 Graeco-Latin square to balance search topics and treatments across task order. The 12×12 Graeco-Latin square forms a single block where each row represents the order of tasks a participant undergoes, as shown in Figure 3.4. Each block contains all possible treatment-topic pairs. In other words, after recruiting 12 participants, our data will include interaction behavior of each topic under each treatment. Each participant saw each search topic and treatment once. We created 6 different blocks to account for the number of participants we were planning to recruit.

3.2.6 Procedure

Figure 3.5 shows an overview of the study procedure. The study was run in a closed computer laboratory using desktop machines with the same monitor size and specifications. The computer monitors had a screen resolution of 1680 × 1050 pixels. Google Chrome browser was used to access the website where the study is hosted.

After receiving participants' informed consent (more details in Appendix B), we col-

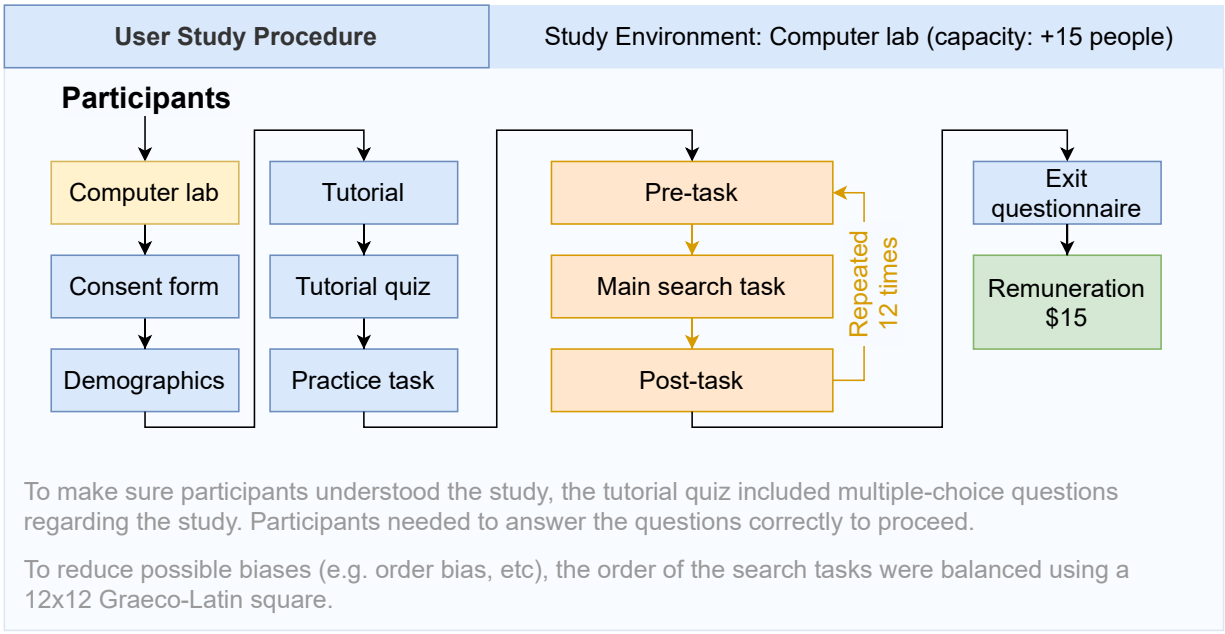


Figure 3.5: User study procedure.

lected participants' demographics and information on their search engine usage and experience before starting the study. A tutorial on the study tasks and expectations were provided before the study. To make sure participants read the tutorial, we included a short multiple-choice quiz after the tutorial page. Participants needed to answer the questions correctly to proceed to the next part of the study. The purpose of this quiz is to make sure people have read the tutorial and understand the study requirements.

We provided a practice page of the search interface and asked all participants to familiarize themselves with the interface by searching for an answer to a practice question. All search results returned by the system during the practice were Bing results. Participants proceeded to their first task after completing the practice task. Completion of the practice task and all further tasks were done by providing a written answer to the task's question.

Each search task included a pre-task and a post-task questionnaire. During the pre-task, we showed the current question and asked participants about their prior knowledge of the current question topic. The post-task questionnaire asks the participants about their confidence in their answers. We asked participants on their feedback and overall experience with an end-of-study questionnaire.

3.2.7 Tutorial

In our study, we explained that users will be asked to complete several brief questionnaires and to search for and save answers towards given search questions for 12 topics using a search engine. Our tutorial included screenshots of the search interface, with annotations to indicate where and how people can submit their queries. Users were told that they can submit as many queries as they would like to complete the search task. Quiz questions and answers are shown Appendix B.1.

3.2.8 Search Interface

We designed an interface similar to that of common commercial search engines (Figure 3.6), except our interface only permitted ten results per query. Participants could enter their search queries using the search bar and trigger the query by either clicking on the "Search" button or pressing "Enter" keystroke. The search box does not provide query suggestions. The question of the current task that participants need to search for was always visible and shown next to the search bar. The question was also shown during the pre-task. Clicking on the help button would trigger a pop-up showing the help information on how to use the interface. Clicking on the answer button will redirect the user to a page with a text box where users can submit their answers.

Participants were asked to use this search interface to find an answer for each question and were allowed to submit multiple queries and click on multiple documents if they wished.

To accurately measure clicks and time spent in the SERP and reading the documents, we disabled right-clicks and opening documents in new tabs. Participants needed to use the back button on the browser to return to the SERP after clicking and viewing a document.

The web application that displays the search engine interface was implemented in Python and JavaScript. JavaScript was used to record various user behavior such as clicks and mouse moves. The web server was hosted locally and accessed with a web browser.

3.2.9 Participants

After receiving ethics approval from University of Waterloo's office of research ethics, we recruited participants through posters placed in different departments of the university. The study took place in a computer lab with more than 20 computers. The study involved 73 participants in total, but only 60 participants' data was used for our analysis. We

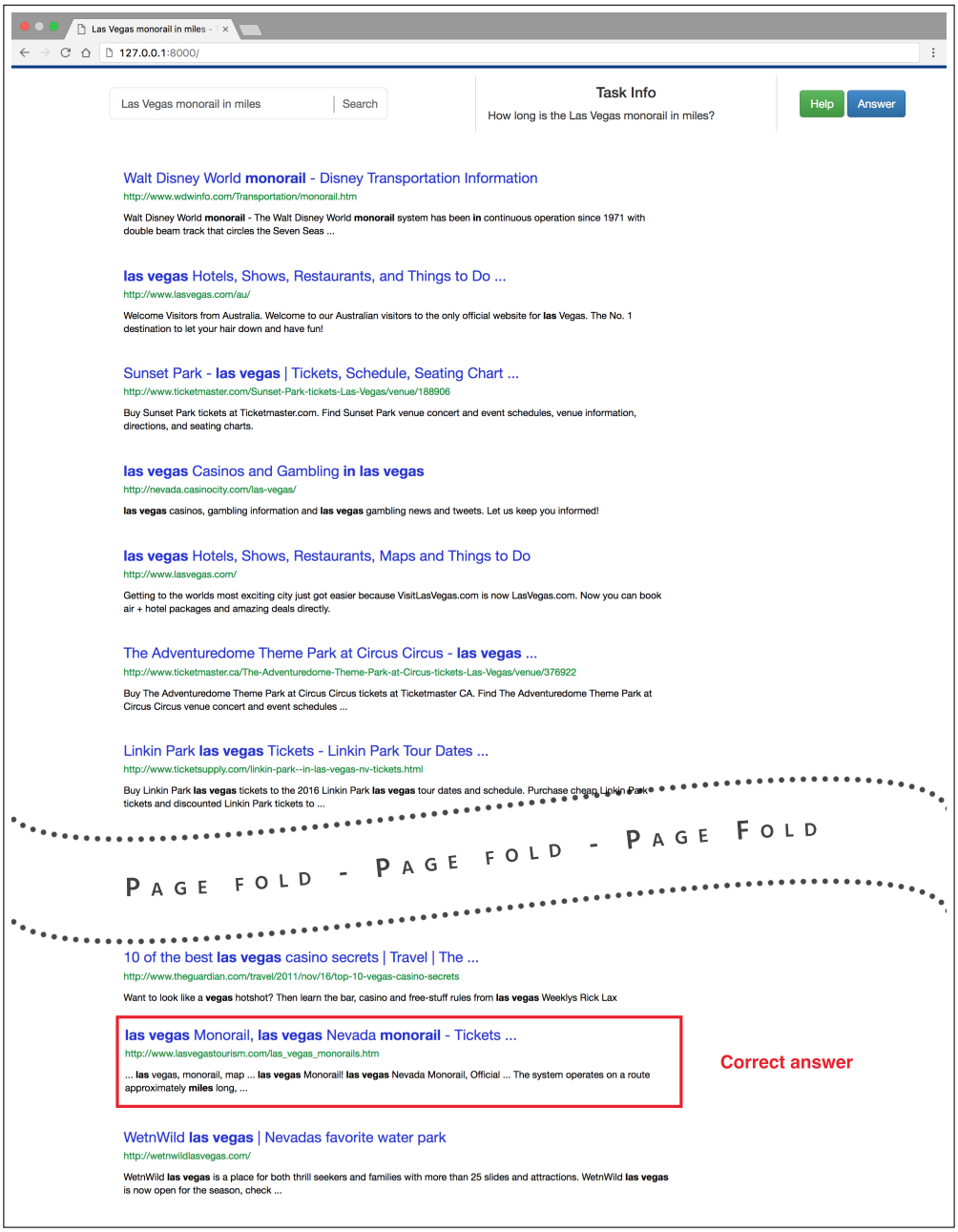


Figure 3.6: The search interface for all tasks. The interface has a search bar, help button and answer button. The SERP shows a maximum of 10 documents with no further results available. Here, a manipulated SERP is presented and the correct document is placed at the rank 9. In general, the results at ranks 8-10 were not visible without scrolling.

removed data of 13 participants due to pilot testing and technical issues. After careful examination of the 60 participants' data, we did not find any irregularities and thus did not clean or modify their data before the analysis. Each of the 60 participants completed their 12 tasks in a balanced order, yielding a total of 720 tasks, 660 were manipulated SERP tasks, and 60 were non-manipulated Bing SERP tasks (control).

Participants' age ranged between 18 and 48 years old (mean = 23.6). There were 34 male and 26 female participants. Of these participants, 54 of them were from science, technology, engineering, or math, 1 from arts, and 5 did not specify their major.

Each participant was compensated $15 with an advertised payment of $10 for participation and a $5 bonus for answering at least 10 out of 12 questions correctly. However, regardless of participant performance, we paid all participants the full $15. This payment structure was designed to motivate good performance while not harming any person who might not have been able to answer 10 questions correctly. 58 participants answered 10 or more questions correctly. One participant answered 9 questions correctly, and one participant only answered 8 questions correctly.

3.2.10 Data Post-processing

After analyzing the search logs for manipulated SERP tasks, we found that only two participants on two different tasks failed to trigger the manipulated SERP with their first query. The first user entered "canadian heratige site 1999" as their first query for task #11, with the wrong spelling of the word "heritage". None of the query terms are triggers. The second user entered an empty query for task #3 and our system returned an empty SERP. Both of these two users successfully triggered a manipulated SERP with their second query. For both of these two users, we skip their first query and analyze their data from the query that triggered a manipulated SERP.

3.3 Result and Discussion

In our study, participants used a search engine to find answers to 12 questions. For 11 search tasks, we manipulated the search results quality. For one of the search tasks, which acted as a control, participants received results directly from the Bing search API. For the manipulated SERPs, any queries that followed the manipulated SERP provided results from the Bing search API. As explained in Section 3.2.2, the manipulated SERPs included 1 single correct document, placed in different ranks from 1 to 10, or 0 correct documents.

In total, we collected user interactions data for 720 search tasks (12 tasks $\times$ 60 users). In the next sections, we present our result after analyzing the data. In this work, we focused mainly on two questions: the probability of abandonment and the time it takes for users to make a query abandonment under different SERPs. We also did some analysis on users searching strategy to understand the results more clearly. At the end of the section, we discuss our findings and the limitations of the work.

3.3.1 Probability of Query Abandonment

Figure 3.7 and Table 3.2 show the probability of abandonment under each of our study treatments. In Figure 3.7, we clearly see that as the rank of the relevant document goes from rank 1 (top of page) to rank 10 (bottom of page), the probability of an abandonment increases. The highest probability for an abandonment, 0.92, occurs when all of the search results are non-relevant (`NC`). The non-overlapping confidence interval indicates that this rate is a statistically significant difference from the other conditions. The control condition's search results, which are Bing API search results, have a probability of abandonment of only 0.18, which is, for all purposes, the same as we saw for a relevant result at rank 1. The probability of abandonment at rank 1, 0.17, is significantly different than at rank 2, 0.42. There is an increase in the probability of abandonment after the 7th result, which we believe is due to the page fold. In order to view the 8th, 9th, and 10th result, the user would need to scroll down the page.

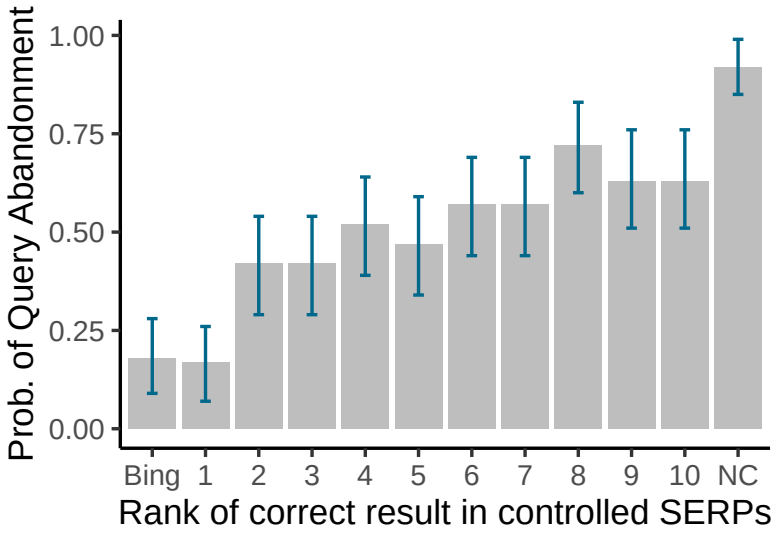


Figure 3.7: The probability of abandonment for the 12 different SERP conditions. The error bars are 95% confidence intervals.

Correct Document Rank	Frequency	Probability of Abandonment [95% CI]
Control (Bing API)	11	0.18 [0.09, 0.28]
1	10	0.17 [0.07, 0.26]
2	25	0.42 [0.29, 0.54]
3	25	0.42 [0.29, 0.54]
4	31	0.52 [0.39, 0.64]
5	28	0.47 [0.34, 0.59]
6	34	0.57 [0.44, 0.69]
7	34	0.57 [0.44, 0.69]
8	43	0.72 [0.60, 0.83]
9	38	0.63 [0.51, 0.76]
10	38	0.63 [0.51, 0.76]
No Correct	55	0.92 [0.85, 0.99]

Table 3.2: The frequency and probability to query abandonment with corresponding 95% confidence interval on the different SERPs (cf. Figure 3.7).

3.3.2 Time to Query Abandonment

Figure 3.8 and Table 3.3 show the time to query abandonment under each of our study treatments. Figure 3.8 shows that the time it takes a user to decide to abandon their query appears to be independent of the search results quality. Figure 3.9 shows the distribution of all times to query abandonment. The median time for a query abandonment is 7.7 seconds, and the average time is 9.2 seconds. A log-normal distribution fitted to this data has a mean of 2.0 and a standard deviation of 0.68.

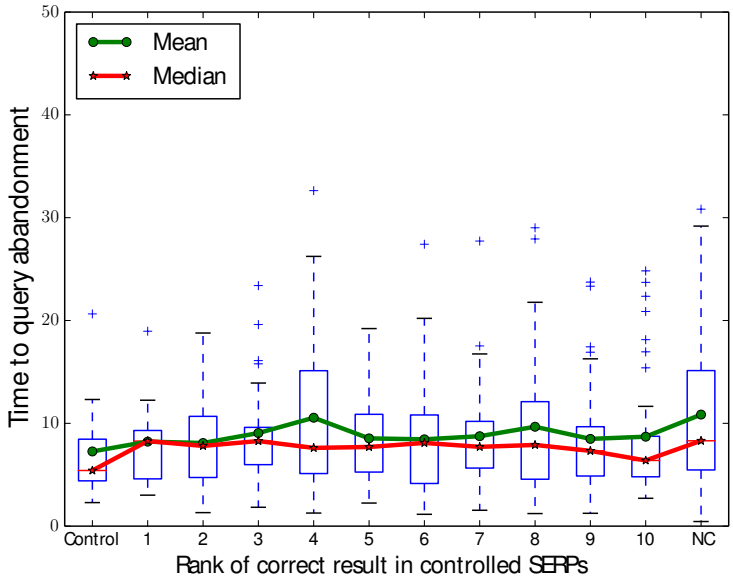


Figure 3.8: Time to query abandonment on each condition.

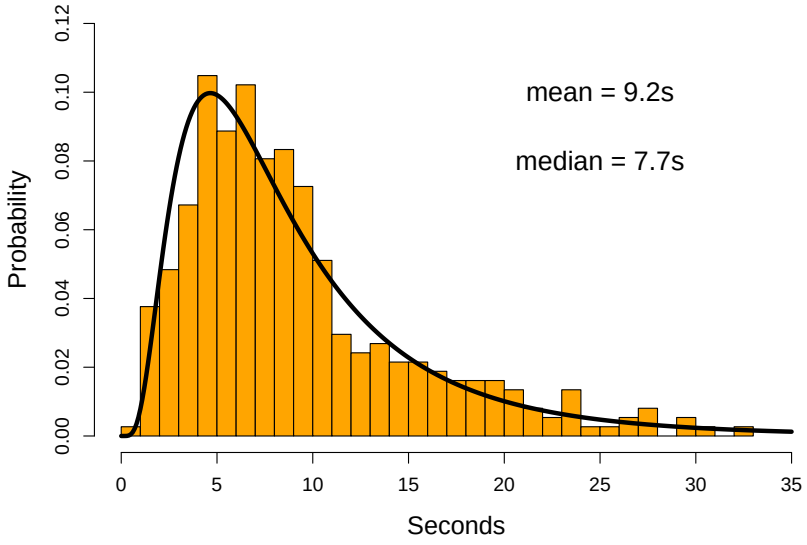


Figure 3.9: The distribution of time to query abandonment on all SERPs. A log normal curve fit to the data is also shown.

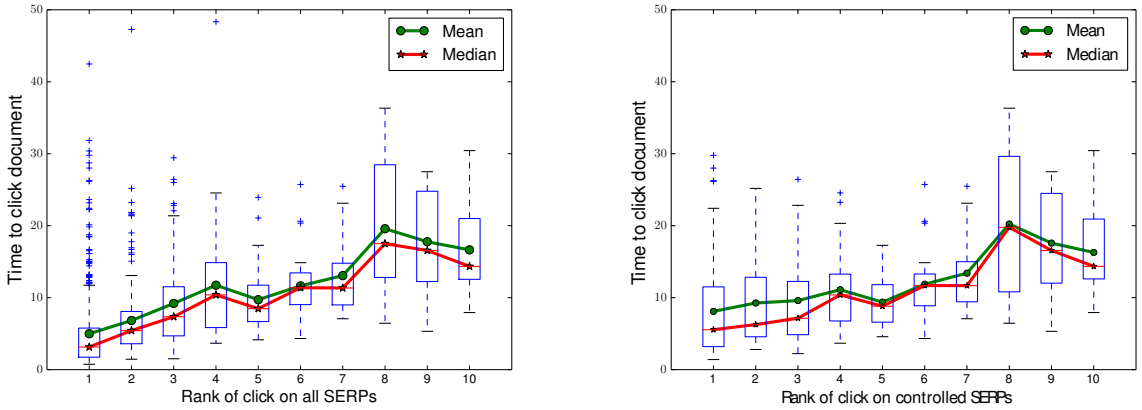
Correct Document Rank	Frequency	Seconds to Abandonment [95% CI]
Control (Bing API)	11	7.3 [4.1, 10.5]
1	10	8.2 [5.2, 11.2]
2	25	8.1 [6.2, 10]
3	25	9.1 [7.0, 11.1]
4	31	10.5 [7.9, 13.2]
5	28	8.5 [6.9, 10.2]
6	34	8.4 [6.5, 10.4]
7	34	8.7 [7.0, 10.5]
8	43	9.7 [7.7, 11.6]
9	38	8.5 [6.8, 10.2]
10	38	0.63 [0.51, 0.76]
No Correct	55	10.9 [8.9, 12.8]

Table 3.3: The frequency and mean time to query abandonment with corresponding 95% confidence interval on the different SERPs (cf. Figure 3.8).

3.3.3 Time to Document Clicks

We also measured the time from a query to a participant's first click on the search results. Figure 3.10 and Table 3.4 show the time from a query to the first result click for ranks 1-10.

We can see a linear increase in the time it takes participants to scan the ranked list of results from rank 1 to rank 4. The median time from query to a click on rank 1 is only 3.1 seconds, and then it takes approximately 2 seconds more for each rank up to rank 4, which takes 10.4 seconds to reach. Participant's behavior on ranks 5-7 is different with these median times taking 8.5, 11.4, and 11.3 seconds. Finally, for the ranks that require the participant to scroll to reach, ranks 8-10, we see that participants appear to scan these upward from rank 10 to 9 to 8 with median times of 14.4, 16.6, and 17.5 seconds, respectively.



(a) (b)

Figure 3.10: Time from query to the first result click at different ranks on all SERPs combined (a), and on manipulated SERPs (b)

Rank of Correct Document	Median Time To Click	Mean Time To Click [95% CI]
1	3.1	5.0 [4.5, 5.5]
2	5.4	6.8 [6.0, 7.7]
3	7.3	9.2 [8.0, 10.4]
4	10.4	11.7 [9.5, 13.9]
5	8.5	9.7 [8.6, 10.9]
6	11.4	11.6 [10.1, 13.2]
7	11.3	13.1 [10.8, 15.3]
8	17.5	19.6 [14.6, 24.5]
9	16.6	17.8 [14.5, 21.0]
10	14.4	16.6 [14.0, 19.3]

Table 3.4: Time in seconds to first click on a result at different ranks (cf. Figure 3.10a).

3.3.4 Analysis of Users and Search Strategies

Given past eye-tracking research that has shown there to be two different classes of searchers, i.e. economic and exhaustive searchers (see Section 2.1.3), we looked closer at the individual behavior of the study participants.

Figure 3.11 shows the distribution of the number of abandonments per participant. While our analysis is limited by the number of participants and the number of search tasks,

it appears that we have one group of participants who have a low rate of abandonments ($\leq$ 3 abandonments), and another group that abandon their queries much more frequently ($\geq$ 4 abandonments). Our threshold criteria was based on visually inspecting Figure 3.11. As such, we label each participant as either having a low or high probability of abandonments and looked at the behavior of each group.

Figure 3.12a shows the probability of abandonments for the *low* vs. *high* groups. As can be seen, the *low* group's probability of abandonments stays low until they are faced with search results that are all non-relevant. In contrast, the *high* group's probability of abandonments grows quickly as the rank of the relevant document goes from 1 to 10. It appears that the *low* group are *exhaustive* searchers while the *high* group are likely *economic* searchers.

Azzopardi (2011) suggests that users try to optimize their search behavior to find answers as quickly as possible. If this is to be the case, then we should see the majority of *economic* users find answers regardless of their probability of query abandonment. We computed the time to from the start of the task to the point where users submitted their answer, and indeed, we found that participants who are more likely to abandon their query are able to find answers faster. The mean time to answer for the participants likely to abandon (*high*) is 85.9 seconds and the mean time for the participants with *low* probability of query abandonment is 111.6 seconds, and this difference is statistically significant by a two-tailed, Student's t-test ($p = 0.0005$). While this difference is significant, it is possible that the *high* group's performance is the result of many additional factors that correlate with a higher probability for query abandonment.

Figure 3.12b shows the median time to answer a question for the *low* and *high* groups of users across the 12 search conditions. While the data is noisy because of the limited participants in the *low* group, the data shows that for the control condition, and conditions where the relevant document is at ranks 1-4 and 8-10, the *low* participants take longer than the *high* group. We also see that for the mid-ranks of 5-7, the *low* users have slightly faster times to answer than the *high* group.

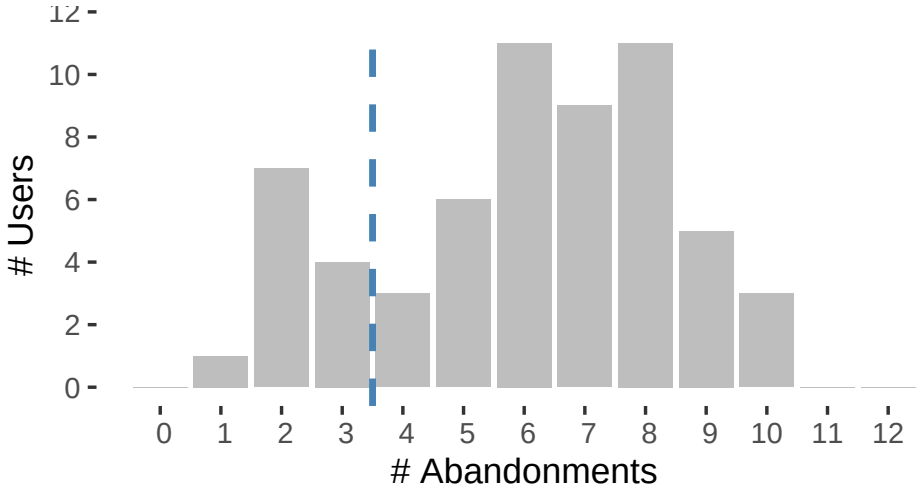
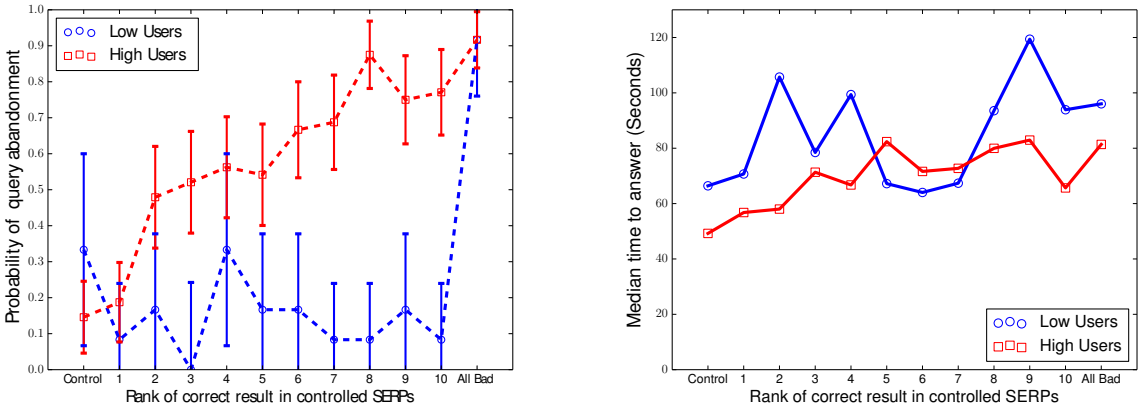


Figure 3.11: Distribution of the number of abandonment per participant. While our analysis is limited by the number of participants and the number of search tasks, it appears that we have one group of participants who have a low rate of abandonment (≤ 3 abandonment), and another group that abandonment much more frequently (≥ 4 abandonment).



(a) (b)

Figure 3.12: Analysis of query abandonment and time to task completion based on user type. (a) The probability of abandonment for the 12 different SERP conditions for two different groups of participants (cf. Figure 3.7). The "Low Users" abandoned their queries for 3 or fewer of the 12 search tasks. The "High Users" each had 4 or more query abandonments. The error bars are 95% confidence intervals. (b) The median time from starting a task to answer a question for different groups of users.

3.3.5 General Discussion

Query abandonment can be good and bad. In this work, we focused on abandonment where the user abandons the query unsatisfied by the poor quality search result presented to them. From a search engine point of view, this particular type of abandonments should be minimized, as it provides no value to the user and incurs an additional cost in terms of the total time searching for relevant information.

When we reduced the quality of the search results in the SERP, we expected that users will be less satisfied and their abandonment rate would increase. Indeed, Figure 3.7 mirrors this expectation. As the rank of the top-most relevant documents moves down the rank list, the higher the chance that the user will abandon the results. Looking closely at the figure, it seems that there are three groups where rank appears important. The first is ranks 1–3, which was shown by eye-tracking heatmaps to be highly visible more likely to examined more thoroughly than other ranks, as indicated by the wider and more visible attention in the heatmaps (e.g., Figure 2.2). The next is ranks 4–7, which represents the area in the screen that is within the page fold. And 8–10, where the results are below the page fold and would require the user an additional scrolling action to be view-able. Based on our results and others' eye-tracking studies, it appears that abandonment in web search is largely caused when the topmost relevant search results appears at ranks lower than 3 or 4.

One of the interesting findings that we did not expect is that being quick to abandon can be an efficient strategy. In our work, we found that the group of users that abandoned their queries quickly after determining the search results to be not helpful were able to find answers and complete their tasks faster than participants who stayed with the search results. In other words, being quick to abandon may actually be an efficient strategy which many of our participants have employed. These result mean is that modifying the search engine to minimize the rate of abandonment can actually hurt user performance if the modification forces users to stick with bad results rather than quickly move to better results.

Indeed, it would seem that an important function of web search engines is to help users quickly find a query that delivers relevant documents at ranks 1 to 3. The faster a search engine can guide a user's query reformulations to the "right query", the faster the user will find relevant results.

Traditional evaluation of search engines focuses on the single list of search results produced by a query. Unfortunately, looking only at the quality of a search engine's ranking, focuses attention primarily on the minority of users who have a low probability of query

abandonment. In our study, it does not seem to matter to the majority of participants if a relevant document is at rank 5 or rank 10, both are considered to be worthless. It is important to keep in mind that for different or more complex search tasks, we might expect user behavior to differ from what we observed.

If only the top 3 or 4 results matter to a majority of users, as information retrieval researchers, we should help users zero-in on the right query and to find ways to evaluate a search engine's ability to help users with this process of querying and repeated reformulation.

3.3.6 Limitation

A limitation of our work is that we only studied one type of search task. Our study participants needed to find answers to simple questions. Other search tasks may result in different behavior. For example, when our study participants experienced a SERP with only 1 relevant document at rank 1, we only saw a 17% query abandonment rate, which is considerably different than the 42% that Wu et al. (2014) found. Likewise, when our topmost relevant document is at rank 4, we found that 52% of participants would abandon their query while Wu et al.'s "bursting" pattern had only a 20% rate. We think these differences in results are likely the result of the different types of search tasks that our two studies used. Our study had participants search for a single answer to a simple question. On the other hand, Wu et al. had many search tasks that would involve attempting to find many relevant documents. It appears that the search task can change query abandonment behavior.

We choose simple factoid questions because we wanted to capture search behavior of participants searching for answers as they would normally do for topics that are familiar. Asking participants to search for answers to more complex questions or questions they are unfamiliar with may result in different patterns of SERP examination, which could be interesting to further study but is not in the scope of our experiment. Our participant are young university students. Certainly older people might be different in their behavior in some way.

A potential concern of our study would be if participants noticed the manipulation of search results. Our study provided a means for participants to supply open ended feedback after each search task as well as at the end of the study. Some participants commented that they were surprised that our search engine would not return Wikipedia search results at rank one when they included keywords such as "wiki" in their queries. One participant noted that our search engine seemed to be sensitive to the order of words in the query.

Thus, while participants may have noticed some behavior different from commercial search engines, they did not specifically make mention of our manipulated behavior, and we did not notice any behavior that would indicate that they understood how the results were manipulated.

3.4 Summary

In this work, we conducted a controlled user study to investigate the relationship between search results quality and query abandonment.

We found that in our study:

- Users make their decisions to abandon or click quickly. The median time from the moment users submit their query to the moment they abandon their query was 7.7 seconds.
- The probability of a query abandonment increases as the user has to search further down the ranked list to find a relevant document. In particular, the probability of abandoning doubles when the topmost relevant document is at rank 2 rather than at rank 1.
- The time it takes users to decide whether to abandon or not appears to be independent of search results quality.

We also found that there may be two classes of user behavior for the examination of search results. One group, the majority, focuses on the top of the ranked list to make their decision about whether to requery or not. The other group appears to be more likely to examine the whole ranked list. The group more likely to abandon is able to find answers faster than the other group.

From this experiment result and other eye-tracking studies, the top 3 or 4 search results appear to be important to most users. As IR researchers, we should plan our search systems to return relevant information at those ranks, help users in focusing their attention on the right query for their information need, and find ways to evaluate a search engine's ability to help users with this process of querying and repeated reformulation.

Chapter 4

Patterns of Search Result Examination

Our previous work in Chapter 3 investigated the the rate in which people abandon their queries and the time it takes people to make their decision to either click on a search result or abandon their queries. It also left us with some unanswered questions. In this chapter, we describe our work where we used an eye-tracker to investigate some properties of query abandonment, including: what search results people examine before making their action, why people may decide to examine more results, and what influence their decision to process more search results or abandon the query.

4.1 Introduction

Given a set of search results, our user study in Chapter 3 shows that as the rank of the topmost relevant result increases, the probability increases that a user will not click on the relevant result and will instead reformulate and requery to get fresh search results. In this chapter, we discuss our work on using eye-tracking to better understand the underlying causes of these requeries without clicks and direct our study to user behavior from the query to the user's first action: either a click on a search result or a requery.

Our study is motivated by the work described in Chapter 3. Similar to that work, we allow users to freely query our search engine and control the search results to allow only one relevant result at ranks 1-10 or no relevant results in response to a user's first query. If a user abandons their query, the search engine defaults to a commercial search engine's results.

We include eye-tracking in our study to be able to know what search results users do and do not examine. Unlike other work (Joachims et al., 2005; Wu et al., 2014; Ong et al., 2017; Maxwell and Azzopardi, 2018) where many of their search tasks include multiple relevant documents in the SERP, our focus in this work is not in investigating which document among those that are relevant the user clicks on, but on understanding possible causes to how far are users are willing to examine SERPs with either no relevant document or one relevant document placed at different ranks, and what motivates users to continue or stop their examination prior to making their decision on whether to abandon or not. We aim to understand how different reasons to query abandonment can affect examination and vice-versa, i.e., how examination patterns can influence a person to abandon their query. Our work is different from Guan and Cutrell (2007) which concluded that the low click probability on what the authors described as the "best" search result is caused by their probability of examining it. In this work, we investigate factors that influence their examination and cause them not to look at the "best" search result when it is placed in any of the 10 ranks of the search result, as opposed to two ranks from the top, middle and bottom areas of the search results as in their study.

We also investigate an important part that was missing from the work in Chapter 3 and other related research, that is, the quality of user queries and their influence to examination behavior and decisions to abandon their queries.

While it is well known that users are less likely to examine lower ranked search results, we show that regardless of rank, if a user sees a relevant result, the user will click it with high probability. We confirm our hypothesis that the *exhaustive* and *economic* user types as characterized by Aula et al. (2005) play a significant role in understanding requeries without clicks. What drives a user's examination to end their search process at certain ranks in the search result? We found that certain ranks and display issues affect user examination patterns, but most interestingly we found that the quality of a user's query appears to be known to the user and the user will modify their examination pattern based on query quality. This gives us an understanding of how likely people are to examine certain ranks under different types of query quality and can be seen as motivation to design effectiveness measures that include factors other than the relevance of search results.

4.2 User Study

This study follows the same study design as our previous study in Chapter 3. In this section, we briefly describe the changes implemented in this study.

4.2.1 Lab Setup

Unlike the previous user study, this study was conducted in a small computer lab space with eye-tracking capabilities. Only the participant and the research coordinator were allowed to be in the room while the study was taking place. We used both a desktop computer and a mobile device for this study. The desktop computer was running Windows 10 operating system. We used a 24 inch monitor with 1920 × 1080 resolution. For mobile, we used a Google Pixel 2[1].

To reduce the risk of participants moving their eyes out of the eye-tracker hardware, participants were seated in a stationary chair that does not swivel. The participant's desk was positioned perpendicular to the researcher's desk. In the desktop setup, the researcher and the participant were using the same computer, but with two different monitors, keyboards and mice. The researcher can monitor the eye-tracking fixation data in real-time from his monitor screen while the participant is completing their task from their own monitor. Figure 4.1 shows the setup of the computer lab.

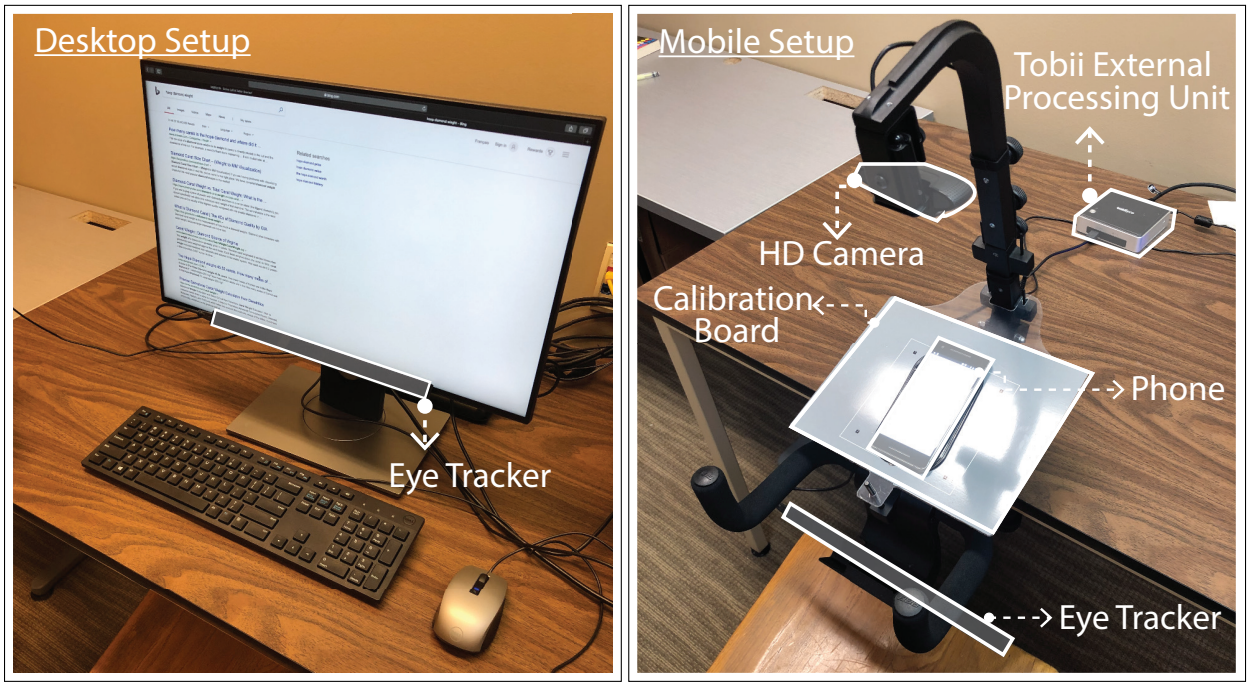
4.2.2 Tutorial

Like the previous study, participant needed to complete a tutorial before completing their tasks. In the tutorial, we told participants that they needed to complete 12 search task and that fr each task, they needed to search for the answer using our search engine. Participants were also instructed that *"Once you are confident that you found the answer, please say it out loud immediately. For example: I found the answer! it is ..."*. Participants were also told *"We want you to search for answer as you would normally using your phone/PC. Please don't act differently."*. For more details on the tutorial, see Appendix C.1.

4.2.3 Eye-tracking

The eye-tracker used was Tobii Pro X3-120[2]. The eye-tracker has a sampling rate of 120Hz, e.g., the eye-tracker registers 120 individual gaze points per second. We added an external processing unit provided by Tobii[3] so that the eye-tracking software runs on the external processing unit and relieves the PC's processor from the heavy workload of processing fixations.

[1] https://en.wikipedia.org/wiki/Pixel_2
[2] https://www.tobiipro.com/product-listing/tobii-pro-x3-120/
[3] https://www.tobiipro.com/product-listing/external-processing-unit/



(a) Desktop eye tracking.

(b) Mobile phone eye tracking.

Figure 4.1: Desktop and mobile phone eye-tracking setup.

The eye-tracker can be used in various setups by attaching it to monitors, such as in Figure 4.1a, or mounted in a custom Tobii Mobile Device Stand as shown in Figure 4.1b. Tobii Pro offers multiple software products for its eye-tracker hardware, such as the Tobii Studio[4] and the Tobii Pro Lab[5]. The software products allow calibration of the eye-tracker, real-time eye fixation tracking and recording of the participant screen, and analysis of the collected eye-tracking data. We used the Tobii Studio for the desktop setup of the experiment, and the Tobii Pro Lab for the mobile setup.

4.2.4 Search Tasks and Questions

We used the same questions as in our previous work in Chapter 3 except for one question. We replaced the question "*How many times did Michael Jordan play the NBA All-Star Games?*" with the following factoid question: "*What is the first studio album Rihanna has*

[4] https://www.tobiipro.com/product-listing/tobii-pro-studio/
[5] https://www.tobiipro.com/product-listing/tobii-pro-lab/

released?". The reason for the change is that many participants were unable to provide the correct answer for that particular question in our previous study.

Unlike the previous study where users were asked to submit their answer in a textbox, we instructed participants to stop once they were confident they had found an answer and to say the answer aloud to the study coordinator. The search task ends once a participant announces their answer out load to the researcher, regardless of whether or not the answer is correct. The researcher did not provide any feedback regarding their answer.

4.2.5 Search Interface

In this experiment, we included both desktop and mobile search users. Figure 4.2 shows the search interface for desktop and mobile. For both interfaces, the search task question is shown at the top of the page, and a search box is provided to allow users to query the search engine. The search box does not provide query suggestions. After a user submits a query, both interfaces show ten results with no pagination, i.e. users cannot click to view the second page of results. For the desktop interface, the page fold line is after the seventh SERP result, and for the mobile interface, the page fold is after the third result.

The web application with the search interface was built using Python's Django, Javascript and HTML. The web application was hosted locally and accessed via a browser. We used Internet Explorer browser for the desktop setup, and Google Chrome browser for the mobile setup. We choose Internet Explorer for desktop because the eye-tracking software does not support other browsers.

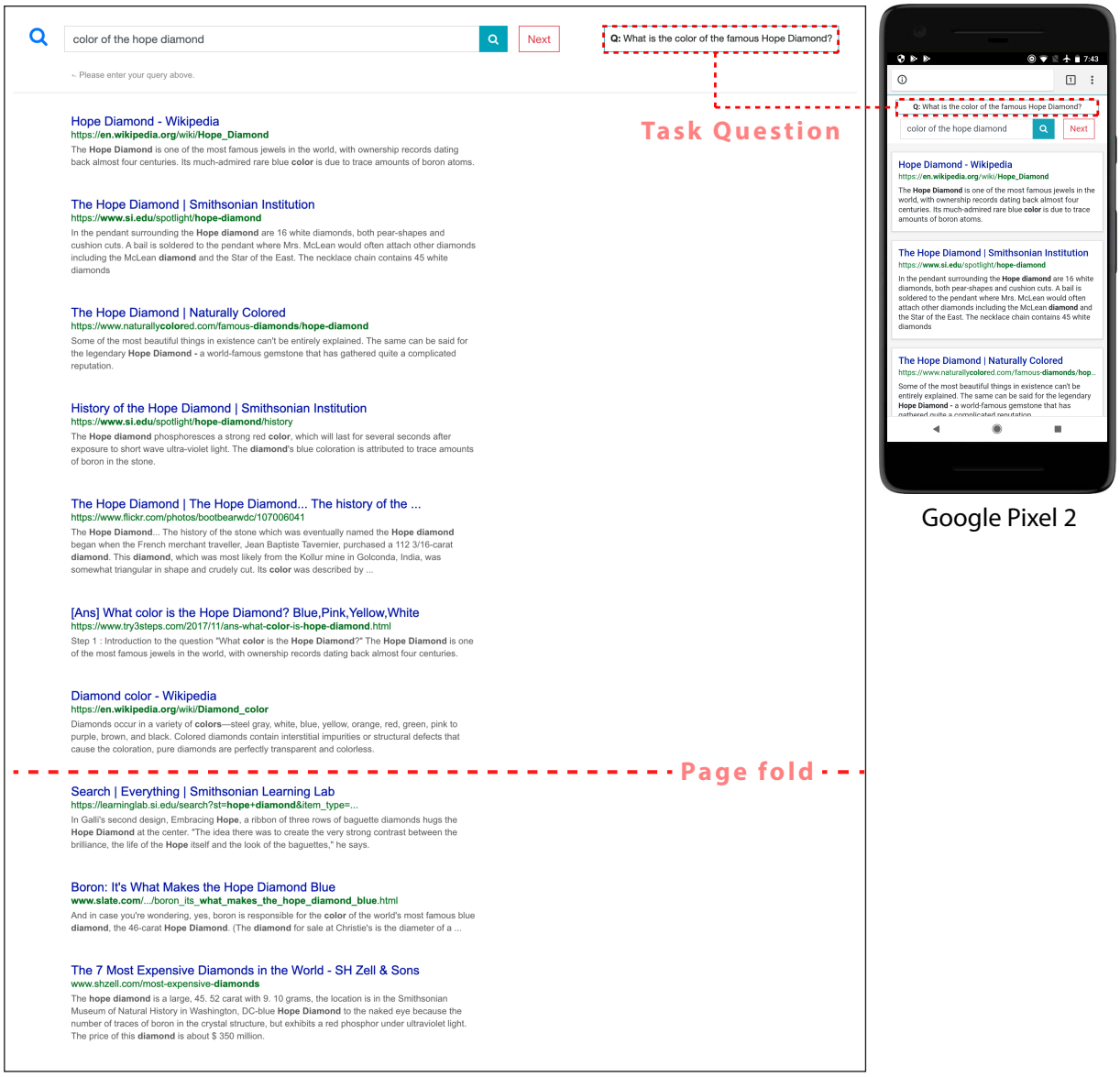


Figure 4.2: The search interface used for our web study on desktop and mobile. The interface was designed to look similar to commercial search engines, with a search box and submit button on the top of the page. We also show the current task question at the top of the page. The search interface fits seven results on the desktop monitor and three on our mobile device, a Google Pixel 2. The Google Pixel 2's actual size relative to the desktop's size is as shown.

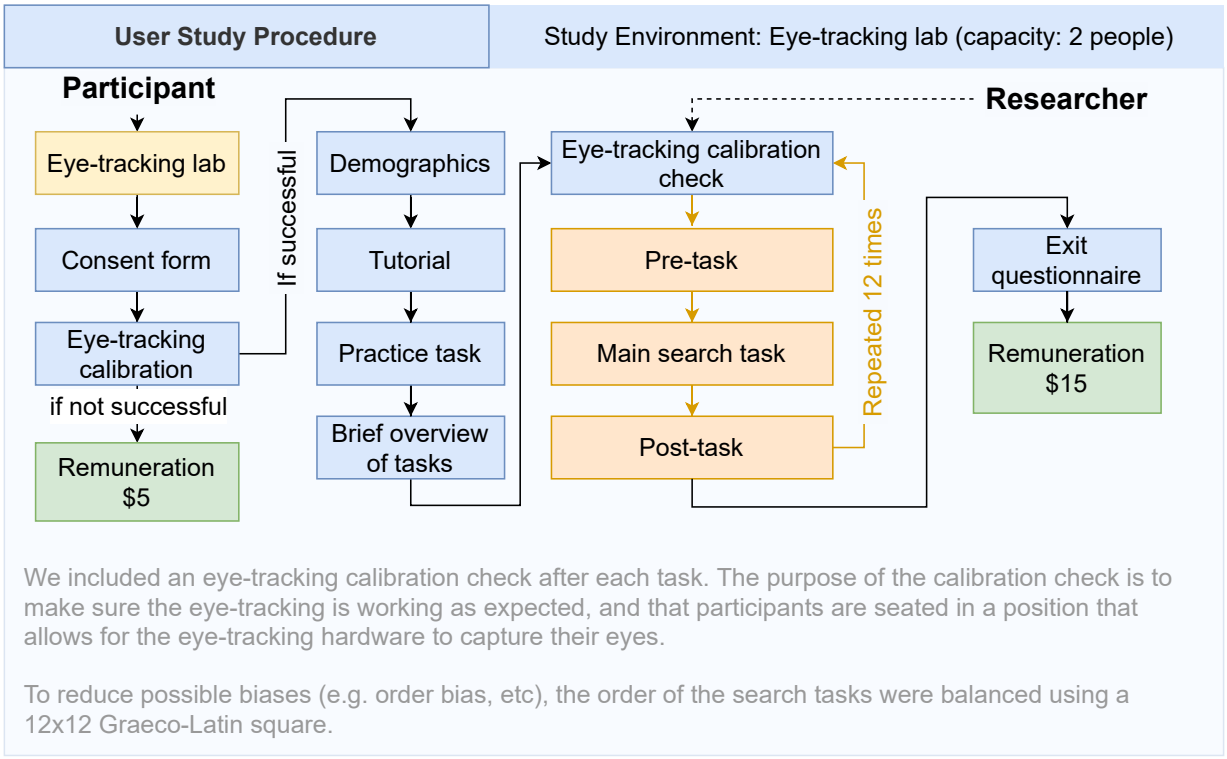


Figure 4.3: User study procedure.

4.2.6 Procedure

Figure 4.3 shows an overview of the user study procedure. Before the participants started the study, we asked them to sign an informed consent form. We then began by calibrating the eye-tracker. If the calibration was not successful for any reason (e.g., participant height, room lighting, etc), the participant was given $5 for their time. We started the study by collecting demographics and general information on participants' experience with search engines. We then provided participants with a tutorial on the study and how to use our search engine.

Each participant began with the practice question (question "P" in Table 3.1). The SERPs during the practice task are not manipulated. After the practice task, participants continue with the main study of twelve tasks. A study task is comprised of a pre-task questionnaire, a search task, and a post-task questionnaire. We provided the current search task question to the participant during the pre-task and asked the participant their perceived difficulty and familiarity with the question and whether they already know the

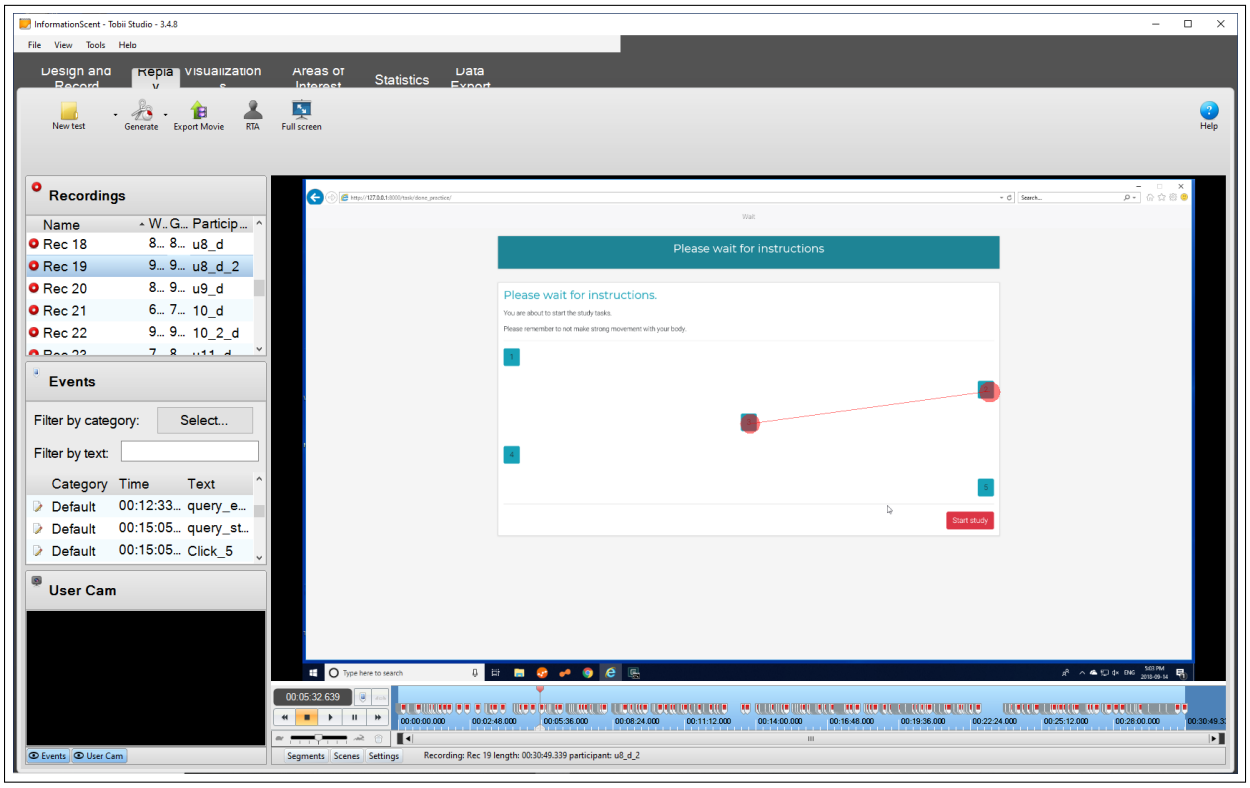


Figure 4.4: Screenshot of the eye-tracking calibration check step. The subject's eye fixation and saccades (gazing paths) are indicated by the red circles and arrows.

correct answer. We also showed the question to the participant in the search interface (Figure 4.2).

Each task begins with an eye-tracking calibration check. In this step, we showed the participant a calibration grid with 5 numbered points. We asked the participant to examine each point for a few seconds while the researcher confirms that the fixations corresponds to the correct grid point. Figure 4.4 shows a screen shot of the eye-tracking software used while this step is being conducted on desktop, with the subject's eye fixation and gazing paths indicated by the red circles and arrows. This critical step is to ensure that the eye-tracking data is accurate, and to remind participants to be seated at the correct distance from the eye-tracker. We repeated this process 12 times, each for a different search task question. After completing the twelve tasks, each participant was asked an exit questionnaire about their experience of the study.

4.2.7 Participants

After receiving ethics approval from our university's Office of Research Ethics, we recruited people through posters posted across our university. We began by collecting data for the mobile search setup. We recruited 22 people, 3 of whom were for pilot testing. We successfully calibrated and ran the study for 11 people on the mobile device. We did not use the remaining participants because of poor eye-tracker calibration for tall participants, participants with eyeglasses, or participants with some eye condition or disorder. After noticing these issues and to prevent such scenarios, we added extra requirements to participate in the study, including an overlooked requirement that participants should be fluent English speakers. To avoid calibration problems, we required future participants to not wear eyeglasses, have long eye lashes, wear mascara, and not have any eye condition or disorder.

For desktop search, we recruited 30 participants, but we used only 24 participants' data in our study. We were unable to calibrate the eye-tracker for 5 participants, and one participant was for pilot testing the setup.

Our participants were university students: 15 females, 19 males and 1 who prefers another term. 27 students were enrolled in an undergraduate program and 8 in a graduate program. Their average age is 20.48, with a minimum age of 17 and a maximum age of 30. Their majors are 3 in art, 1 in environment, and 31 in a STEM major.

We advertised that participants would be remunerated $10 for their time and $15 if they were able to answer 10 out of the 12 questions correctly. However, each participant was given $15 dollars regardless of how many correct answers they have provided. This was done to add some incentive to participants to engage more in the study. After analyzing the participants' data, 29 participants answered the 12 questions correctly and the lowest score was 10 correct answers.

4.2.8 Collected Measurements

Below are a description of each of the collected measurements for this user study.

Submitted queries: All queries submitted to the search engine by the participants during their 12 tasks.

Action: The action made by the user once they are shown the manipulated/Bing SERP. An action could be a requery, a document click, or a *snippet answer*. For document clicks during manipulated SERPs, we record whether or not the clicked search result was relevant.

A snippet answer indicates a participant has announced their answer to the question by reading the snippet of a search result without clicking on the result. The items in our manipulated SERPs do not contain the correct answer in their snippet, but the Bing search results can directly contain answers.

Time to action: Time to action is measured from the moment the result is shown to the user to the moment the action is triggered (e.g. clicking a document, clicking the search bar, or time of announcing the answer from a snippet). This was measured using the eye-tracking software, where we replayed each search session and tagged the timestamps of when the SERP was shown to the user. In a few cases, participants clicked the search bar, then started looking at SERP results. In these cases, the end time of the action is their first keystroke in the search bar.

The time period between time to action is important as it involves the decision making process by the user. Measurements described below are recorded within this time period.

Mouse moves: The number of mouse moves the participant has made. Two consecutive mouse moves include a 200 milliseconds or more idle period between each other.

Number of Fixations: The total number of fixations made by a participant during a task. Number of fixation is provided by Tobii's eye-tracking software.

Fixation Duration at SERP items: We created 10 areas of interest (AOI) using the Tobii software, one for each search result in the SERP. We record the total fixation duration a user looked at a search result using the fixation data.

Eye Fixation Sequence: The complete sequence of fixations at each search result. An example sequence is `1→2→1`, which indicates a user has looked at the first search result, the second result then back to the first result. We define *Unique Fixation Sequence* as the sequence of unique search results fixated by the user. For example, the unique fixation sequence for the example above is `1→2`.

4.2.9 Data Post-processing

After collecting eye-tracking data of participant while interacting with SERPs, we needed to determine whether a certain eye fixation corresponds to a search result. To do so, we created 10 areas of interest (AOIs) around each search result for each SERP shown to participants. The AOIs were created using the Tobii eye-tracking software. As people scroll down the page, the AOIs position are updated to reflect the changes in the page. The AOIs act as Boolean variables when the eye-tracking fixation data is exported, with a true value indicating that a particular fixation is within a particular AOI. Figure 4.5

shows example screenshot of the AOIs created for both desktop and mobile, with each AOI numbered from 1 to 10 to indicate the rank of the search result.

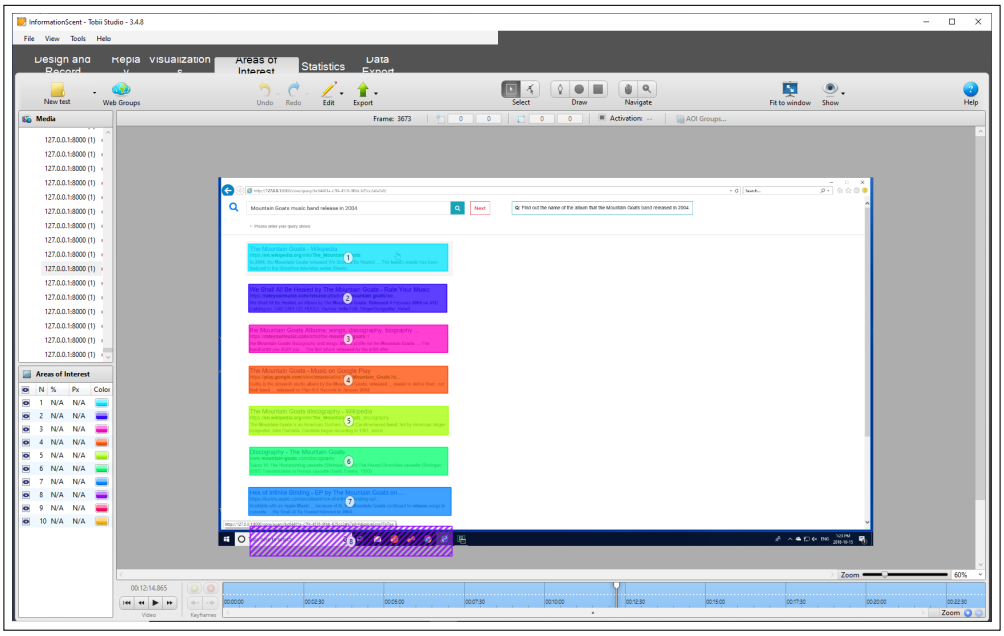
4.3 Result and Discussion

We have learned from our work in Chapter 3 that user type is important to understand the probability of abandonment. We also hypothesized that the study participants could possibly be *economic* and *exhaustive* users as described in previous eye-tracking studies (Aula et al., 2005; Dumais et al., 2010), but we lacked eye-tracking data to confirm their hypothesis in our study. While observing our users complete their tasks, we indeed noticed *economic* and *exhaustive* behavior described in previous eye-tracking studies Aula et al. (2005); Dumais et al. (2010). *Economic* users tend to examine fewer results and give up quicker than *exhaustive* users. This observation motivated us to classify users as either economic or exhaustive using our eye-tracking fixation data.

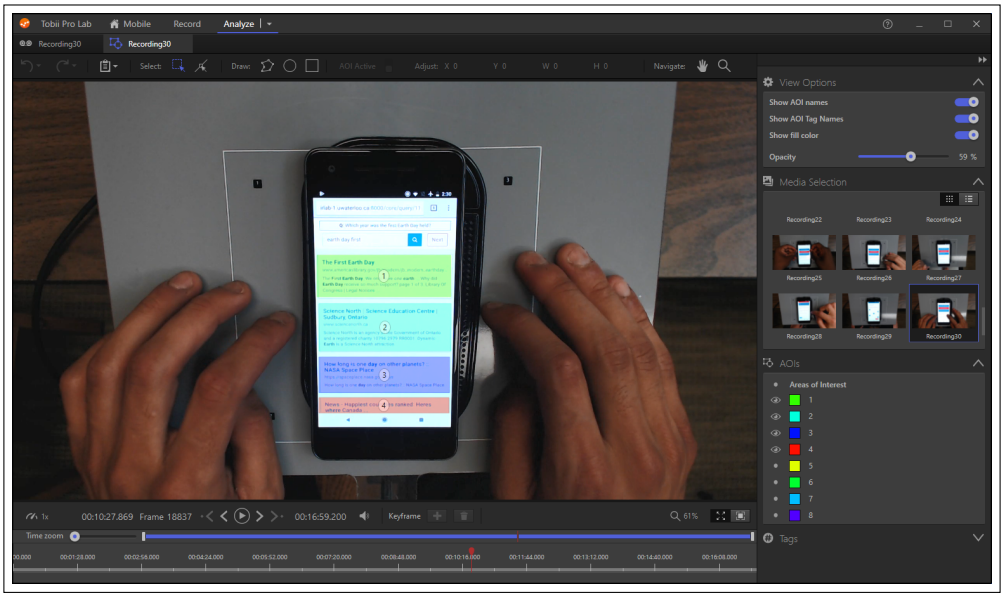
During our observation of users completing their tasks, we noticed a behavior users seem to follow. In many cases, we saw participants enter a query, examine few of the top results but not click on any, and then make a decision to requery. We hypothesize that a possible reason behind the requery decision is due to the ambiguity of their queries. To illustrate, in one case during question `Q4` (Publication date of Holes by Louis Sachar), a participant entered the query "holes novel" and examined, without clicking, the first three search results and then reformulated their query. We think that after the user examined a few search results, the user might have realized their query is under-specified and likely to fail and therefore decided to stop their SERP examination. The user's next reformulated query included the author's name. This motivated us to assess query quality in terms of specificity to the question and check whether it affects user behavior (see Section 4.3.3).

From our observations, we understood that user type, query quality, and rank of the topmost relevant result are factors that likely affect how people examine search results and their decision to click or requery. The question then arises: given a search result and our knowledge of users and queries, when and where do these factors start to matter?

We used decision trees to aid in understanding the overall behavior of users. Decision trees are known for capturing interactions between variables while providing a simple interpretation of the data (Breiman et al., 1984). We model users' first action, i.e. whether users click a search result or requery. Input to the decision tree consists of the rank of the topmost relevant result (1-10, No Correct), user type (economic or exhaustive, see Section 4.3.2), and query type (weak or strong, see Section 4.3.3). We built decision trees

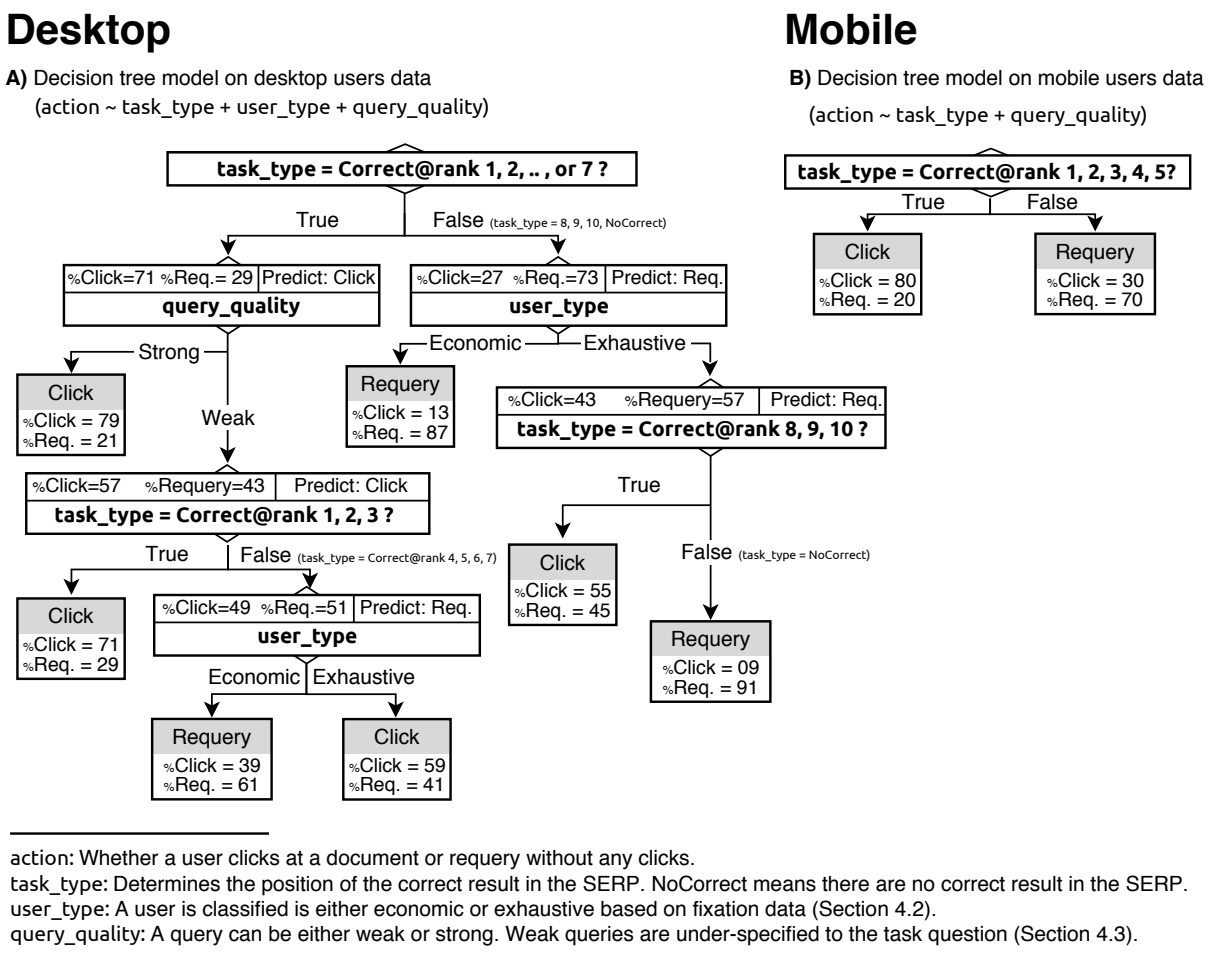


(a)



(b)

Figure 4.5: (a) Screenshot of Tobii Studio software with areas of interest (AOIs) generated for each search result. (b) Screenshot of the Tobii Pro Lab with AOIs generated for each result shown on the mobile page.



action: Whether a user clicks at a document or requery without any clicks.
task_type: Determines the position of the correct result in the SERP. NoCorrect means there are no correct result in the SERP.
user_type: A user is classified is either economic or exhaustive based on fixation data (Section 4.2).
query_quality: A query can be either weak or strong. Weak queries are under-specified to the task question (Section 4.3).

Figure 4.6: Decision tree models for desktop and mobile users.

using the recursive partitioning algorithm (Breiman et al., 1984) as implemented in the `rpart` (Therneau and Atkinson, 1997) package in `R`. The `rpart` algorithm works by recursively partitioning the data into multiple nodes and selecting splits based on node impurity. We used information gain as our splitting index. *Bing* treatments were excluded from the decision tree modeling for we do not control the quality of the SERP and thus do not know the rank of the topmost correct result.

Figure 4.6A shows the decision tree produced for desktop search. The model selects whether or not a relevant result is above or below the page fold (task type = Correct@1, . . . , or 7) as the root of the decision tree, which means this is the most important information to whether a user will click or requery. When the topmost relevant result is below the page

fold, economic users will requery 87% of the time, while exhaustive users are more likely to click on a result.

When the topmost relevant result is above the page fold, query quality becomes important to determining a click or requery. A strong query means that 79% of the time a user will click. In contrast, a user's behavior changes for a weak query based on whether or not the topmost relevant result is found in the top 3 ranks. With weak queries, if the topmost result is in the top 3 ranks, users are likely to click (71%). For weak queries and results at ranks 4-7, behavior again depends on user type. Economic users are more likely to requery (61%) than click on a result at ranks 4-7 if they issue a weak query, while exhaustive users are more like to click (59%) than requery.

Because query quality affects user behavior, we believe that researchers should consider whether it is appropriate to supply queries to users in studies as opposed to allowing users to interactively query the search engine. Many controlled user studies make use of a fixed set of results produced from a fixed "query", but such queries hide that users appear to have a sense of the quality of their query and thus modify their behavior appropriately.

Figure 4.6B shows the decision tree for mobile search. In mobile, the position of the correct SERP items is the only important factor determining a user's action, with rank 1-5 being most important. Our smaller amount of user data for mobile search may limit the usefulness of the decision tree for understanding mobile search behavior.

In the next sections, we look at requeries and examination, how we classified users and queries, and how users and queries influence examination.

4.3.1 Abandonments and Examination

Figure 4.7 and Table 4.1 show the requery probability across ranks and during *NoCorrect* tasks, where there is no correct item in the SERP, and *Bing* tasks, where the results are not manipulated. The probability of requery is high when we place the correct item at ranks 8 to 10 in desktop and in 6 to 10 in mobile. While increasing rank means an increasing probability of requerying rather than clicking, the question arises whether or not users are viewing results at higher ranks, viewing but ignoring documents at higher ranks, or some combination of the two. This question was raised by Guan and Cutrell (2007) who concluded that in some cases users fail to look at results and in some cases they discount lower ranked results. An important difference between Guan and Cutrell's work and ours is that we tightly controlled the search results to only have non-relevant and relevant results while they manipulated the rank of the "best" result while allowing the search engine to

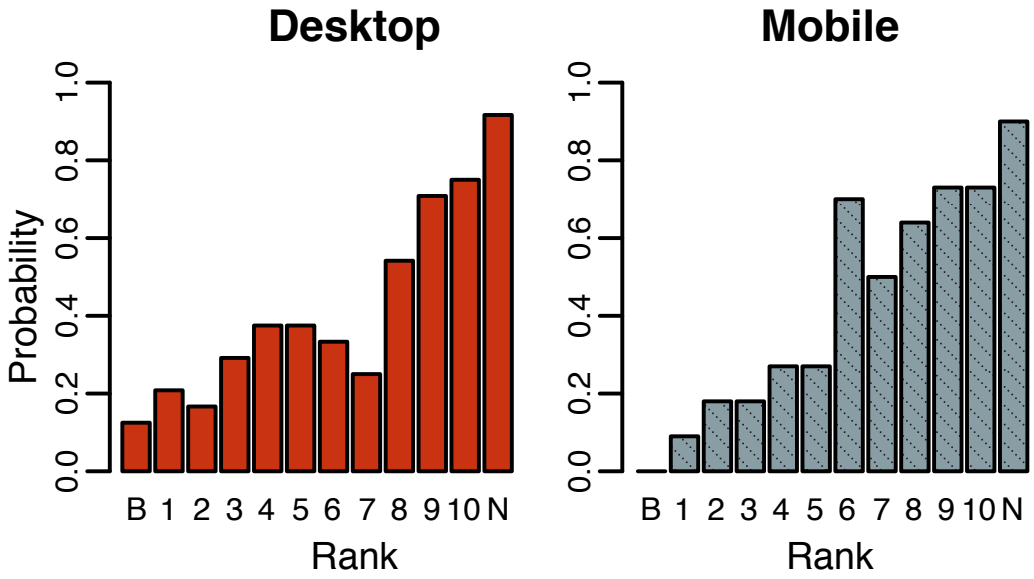


Figure 4.7: Probability of query abandonment in desktop and mobile. X-axis indicates rank of the relevant item in the manipulated SERP. `B` and `N` indicate Bing and **NoCorrect** tasks.

provide other results, which we presume may have also appeared somewhat relevant to the users.

To address this question, we measured how many times a user decided to requery when they have not seen the correct item. If it is the case that users requery often, this serves as a suggestive piece of evidence that an examination sequence, that does not include the correct item, causes users to requery. We used our eye-tracking data to determine how much time users spent fixating at correct items and their resulting action.

Table 4.2 shows the frequency of requeries, clicks, and snippet answers (answers provided by reading the snippet alone) grouped by the duration of fixation at the correct item. The table also shows the frequencies during *NoCorrect* tasks and *Bing* tasks. We first notice that only two people clicked on wrong documents when there is no relevant document in the SERP and that the majority of users decided to requery. We also notice that 85% of the time a user would requery if they have not seen the correct item or quickly glanced over it, and that 88% of the time, a user would click at the correct document if they have examined it for $\geq$1 seconds. In summary, if a user sees the correct item, they click it, if they have not seen it, they requery.

To what extent does rank matter when they see the correct item? Table 4.3 shows the

Table 4.1: Probabilities of query abandonment action, click on wrong/correct SERP items, and average time to query abandonment.

Correct Doc. Rank	Probability of Abandonment		Probability of Wrong Click		Probability of Correct Click		Avg. Seconds to Abandonment	
	Desktop	Mobile	Desktop	Mobile	Desktop	Mobile	Desktop	Mobile
Bing	0.12	0.00	–	–	–	–	4.73	–
1	0.21	0.09	0.00	0.00	0.79	0.91	9.44	7.28
2	0.17	0.18	0.04	0.00	0.79	0.82	6.81	7.58
3	0.29	0.18	0.08	0.00	0.62	0.82	6.32	3.21
4	0.38	0.27	0.00	0.00	0.62	0.73	9.65	8.23
5	0.38	0.27	0.00	0.09	0.62	0.64	4.68	4.37
6	0.33	0.70	0.08	0.00	0.58	0.30	4.37	8.13
7	0.25	0.50	0.12	0.00	0.62	0.50	6.90	5.24
8	0.54	0.64	0.08	0.00	0.38	0.36	5.63	6.66
9	0.71	0.73	0.00	0.09	0.29	0.18	6.16	7.96
10	0.75	0.73	0.04	0.09	0.21	0.18	6.73	5.16
NC	0.92	0.90	0.08	0.10	0.00	0.00	8.66	6.65

Table 4.2: Frequency table of actions grouped by duration of fixation at the correct item. Data is for desktop users.

Time fixating at correct document	Requery	Wrong Click	Correct Click	Snippet Answer
<200ms	64	10	1	0
≥200ms, <1sec	18	1	26	0
≥1sec	14	0	106	0
Frequencies in NoCorrect and Bing tasks				
NoCorrect	22	2	0	0
Bing	3	Total Clicks: 16		5

probability of a requery or clicking at a correct result when it has been seen (≥1 second), across the 10 ranks. The table shows that no matter where the correct item is, if the user sees it, they are more likely to click on it than requery. The results are also similar when we set the threshold to ≥200ms. Our results indicate that if a user sees the correct answer, the rank does not seem to have an influence on their clicking decision.

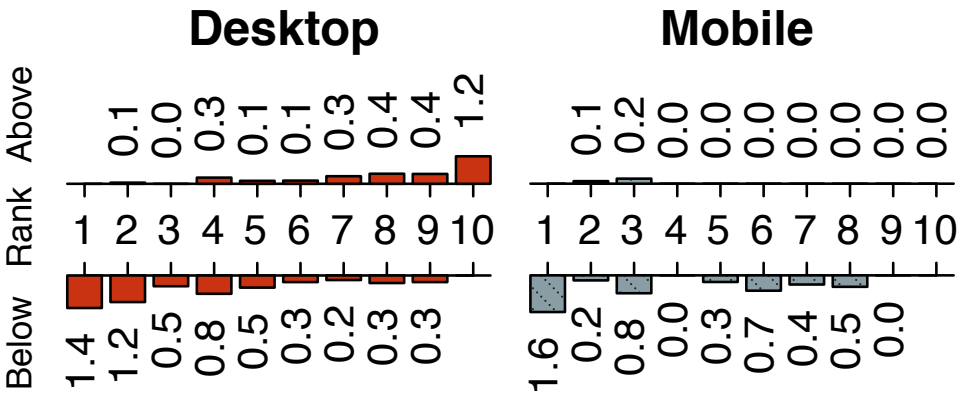


Figure 4.8: Mean number of unique SERP items looked at (fixated $\geq$ 200ms) after the user has seen the correct result and before it is clicked. Bars above/below the rank indicate the mean number of items whose rank is above/below the correct result's rank. Users tend to stop their examination after seeing the correct result, and in the first two ranks in desktop, users stop their examination after examining about 1 more non-relevant item.

Table 4.3: Probability of requery or click on a correct item when it has been seen ($\geq$1 sec). SE reported in brackets.

Desktop Users										
Probability of	Rank of Correct Item									
	1	2	3	4	5	6	7	8	9	10
Correct Click	.7[.2]	.9[.1]	.8[.2]	.9[.2]	1[0]	1[0]	.9[.1]	.9[.2]	.9[.2]	1[0]
Requery	.3[.2]	.1[.1]	.2[.2]	.1[.2]	0[0]	0[0]	.1[.1]	.1[.2]	.1[.2]	0[0]
Mobile Users										
Correct Click	.9[.3]	.9[.3]	1[0]	.8[.4]	1[0]	0.8[.4]	1[0]	1[0]	1[0]	1[0]
Requery	.1[.3]	.1[.3]	0[0]	.2[.4]	0[0]	0.2[.4]	0[0]	0[0]	0[0]	0[0]

Of note, we do not dispute previous research claiming that there exists a position bias or that searchers trust the ranking of search engines as Joachims et al. (2005) have shown. Hofmann et al. (2014) also has a similar finding from their eye-tracking study with query suggestions, where they show that users' strong bias towards examining top-ranking items is an effect due to examination bias. When there are multiple relevant documents, as in Joachims et al. study, the authors show that users trust and click higher ranking items irrespective of actual relevance. Users seem to trust the ranking presented by the search

engine and click on what the search engine has chosen to be higher in the list. When the number and position of relevant documents in the SERP are controlled, as in our study, users are most likely to stop their examination and click on the relevant document once they see it. Figure 4.8 shows the mean number of items a user examines after seeing the correct item, clearly showing how examination stops.

4.3.2 User Types

Motivated by our observations of users and previous research (Zhang et al., 2018; Aula et al., 2005; Klöckner et al., 2004b; Dumais et al., 2010), we investigated the possibility of different user types in our study. Previous research (Aula et al., 2005) has indicated that exhaustive users tend to be slower and spend more time analyzing search results than economic users. Using the fixation data we collected, we plot the distribution of the average total number of fixations during the search tasks. If there exist two types of users, we should be able to see a bimodal distribution of total fixations. While a strong bimodal distribution is missing, Figure 4.9 does show that the distribution of total fixations for desktop users has a large spread, and we selected those users with more than 650 fixations to be labeled exhaustive users while those with fewer than 650 fixations to be economic users. In total, we have 11 exhaustive users and 13 economic users. For mobile, we were unable to see a difference in the distribution (see figure 4.10), and we treat mobile users as equal in our analysis.

Given the two types of user in desktop search, how different are their examination patterns when presented with our controlled SERPs? Exhaustive users tend to spend more time analyzing the SERP, therefore, we expect that they are more likely to see the correct result in the SERP. Figure 4.11 shows the probability of seeing the correct result when placed at different ranks for the different user types and for mobile. Indeed, exhaustive

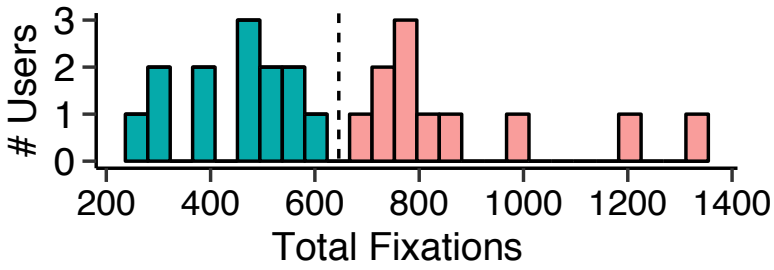


Figure 4.9: Total fixations histogram on desktop users.

users' probability of seeing the correct result is higher than that of economic users. Their probability stays high and decreases a little when the correct result is placed below the fold (ranks 8 to 10). Economic users, on the other hand, are less likely to see the correct result as its position decreases in the list. Previous research has shown that people tend to examine search results in a linear fashion and the time required to reach a specific rank increases linearly (Joachims et al., 2005; Cutrell and Guan, 2007). As economic users spend more and more of their time examining non-relevant items, they are more likely to stop their examination and thus not see the correct result. The result also explains why the page fold is an important factor: economic users rarely examine beyond the page fold but exhaustive users do. In mobile, it seems that people are willing to scroll beyond the page fold to look at more items, and the decrease of probability after the 5th result explains why the decision model picked *Correct@1..5* tasks as the decision criterion.

We investigated other differences of behavior in the two groups. Table 4.4:left shows the averages of different recorded measures and their statistical significance. Statistical significance was computed using generalized linear-mixed models as implemented in the lme4 package (Bates et al., 2015) for R (R Core Team, 2014), with study subjects and search questions treated as random effects. Total number of fixations, sequence length, and duration of fixation on top, middle, and bottom ranks are statistically significant. Exhaustive users tend to read 1.5 more unique results than economic users and produce 2.96 longer sequence lengths. The time spent fixating at different ranking areas is significantly less for economic users. The time to action is statistically significant with economic users showing a faster time to make a decision.

4.3.3 Query Analysis

As we explained earlier, we noticed that some users would examine only a few results and then requery with a more specific query. This motivated us to look into user queries and

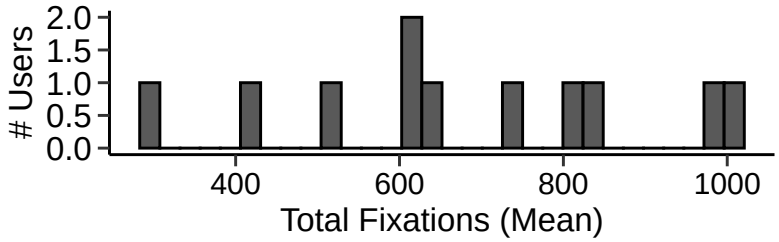


Figure 4.10: Total fixations histogram on mobile users.

Table 4.4: Left: Averages and significance testing of different measures between the two user groups in the desktop setup (*/**/*** indicates statistical significance at p<0.05/0.01/0.001). Right: Averages grouped by user type and under different query types.

	Economic (Eco.) Users	**Exhaustive (Exh.) Users**	***p*-val**	**All Queries**	**Weak Queries**			**Strong Queries**		
				All Users	Eco.	Exh.	All	Eco.	Exh.	All
# of fixations	448.9	869.7	***	641.7	391.5	774.3	552.0	479.3	909.7	684.6
Seq. length	5.2	8.2	***	6.6	4.2	7.0	5.4	5.8	8.7	7.2
Unique Seq. length$_{>0ms}$	4.3	5.7	***	4.9	3.5	5.5	4.3	4.7	5.7	5.2
Unique Seq. length$_{>=200ms}$	3.8	5.3	***	4.5	3.1	5.0	3.9	4.1	5.3	4.7
top_ranks_duration$_{1,2,3}$	1.7	2.9	***	2.2	1.6	2.4	2.0	1.7	3.1	2.4
mid_ranks_duration$_{4,5,6,7}$	1.2	2.5	***	1.4	0.7	2.3	1.4	1.4	2.5	2.0
bottom_ranks_duration$_{8,9,10}$	0.2	1.4	***	0.8	0.1	1.2	0.6	0.3	1.5	0.9
correct_item_duration$_{>0ms}$	1.3	2.2	***	1.8	1.1	2.2	1.7	1.4	2.3	1.8
Time to action	5.1	9.5	***	7.1	4.5	8.8	6.3	5.4	9.9	7.5
Time to requery	5.0	10.2	***	6.8	4.3	8.1	5.5	5.6	11.4	7.7
Time to click	5.1	9.2	***	7.3	4.8	9.2	7.1	5.3	9.3	7.4
# Mouse moves	2.3	4.7	**	3.4	2.1	4.4	3.1	2.4	4.8	3.6
# Query terms	3.8	4.6	**	–	–	–	–	–	–	–

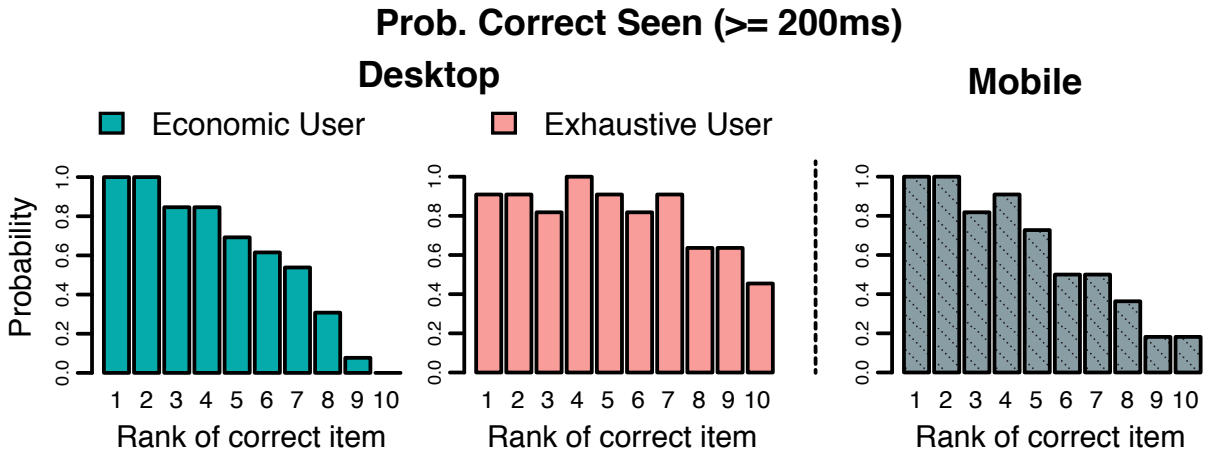


Figure 4.11: Probability of correct result seen.

evaluate them based on specificity to the search task questions. We looked at all users' queries and noticed that there exists a number of queries we consider to be under-specified and that could possibly return non-relevant results if entered in a real search engine. For example, in `Q4`, one participant entered "holes novel" as their first query without any mention of the author name and possibly assumed that the search engine would know. Similarly, "art of war chapters" on `Q8`, "mad cow" on `Q9`, "mountain goats" on `Q2`, or "UN world heritage sites", "rain man album" on `Q7`, "north carolina campus" on `Q10`, and "canada united nations 1999" on `Q11`. All of these queries are missing important terms such as author name or location. We do not consider these queries to be completely bad, but we believe that they are of weak quality in terms of specificity and can be improved and considered less ambiguous by including some of the most important or specific terms.

Based on preliminary analysis of queries, we decided to assess all queries based on their specificity/generality to the question. We consider a query to be either weak or strong. We wrote a short description and a tutorial of what we considered to be a weak or strong query and provided it to assessors that we hired for query assessment. In the description, we mentioned that a strong query *"includes important terms in the question and you consider specific enough to the question. The query is not ambiguous."* And a weak query *"is not specific enough, can lead to off topic results, can be considered ambiguous or contain any misspellings."* The tutorial included examples of different questions and what would be considered weak and strong. After the first assessor completed their judging, we hired a second assessor for verification and to test agreement. The two assessors were fluent in English and had good experience in using search engines. Both assessors had not

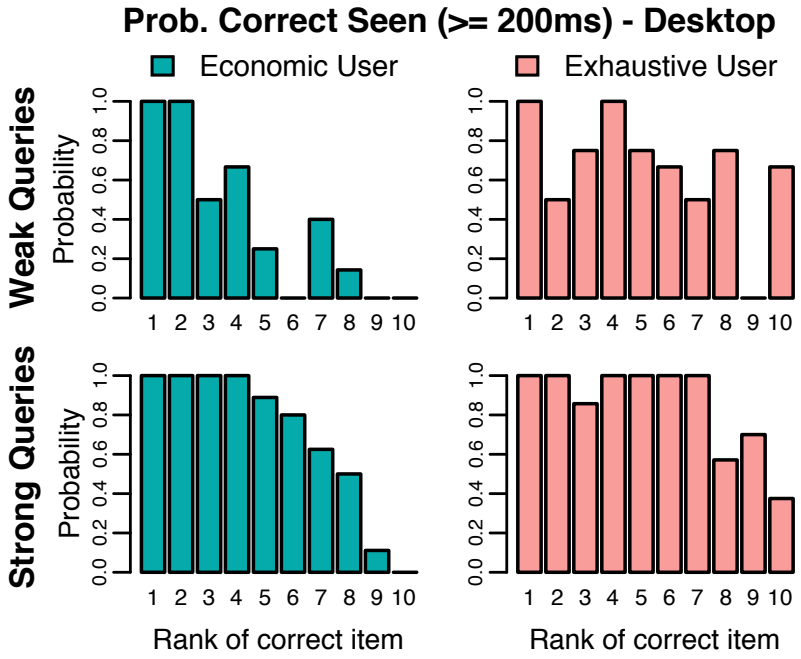


Figure 4.12: Probability of correct result seen under weak and strong queries for different user types.

participated in the study, nor were aware of the purpose of the study. We asked assessors to be careful assessing the queries and take as much time as needed. The assessors were given $20 for their time.

The total number of queries is 12 × 24 (tasks × participants) for desktop and 12 × 11 for mobile. The two assessors took about 1 hour to finish assessing all queries. The Krippendorff's alpha (Klaus, 2004) for inter-rater reliability of the two assessors is 0.80, which indicates substantial agreement (Klaus, 2004). Our two assessors judged 27% and 28% of the submitted queries as weak, accordingly. Since both assessors have substantial agreements, we use the first assessor's judgments as the final judgment for the query type.

The percentages of weak queries made by economic and exhaustive users are 52% and 48%, accordingly. Only 18% of the weak queries were queries with a misspelling. Table 4.5 shows the percentages of weak and strong queries among different groups. Exhaustive users are less likely to requery regardless of query type.

Figure 4.12 shows the probability of seeing the correct result under weak and strong query quality for both user types. Interestingly, query quality seems to have a stronger influence on economic users than exhaustive users, as we have shown in Figure 4.6. Under weak queries, if the correct result is placed at the top of the list, economic users' probability

Table 4.5: Percentages of query types for queries ended with a requery, and for diff. user types. Data is for desktop users.

Query Type %	In requeries where correct not seen and task type != {NC, Bing}	In user type	
		Eco.	Exh.
Weak	47% (80% are by Economic users)	35%	30%
Strong	53% (71% are by Economic users)	65%	70%

of seeing the correct result is high. Their probability of seeing the correct result quickly drops as the correct result is placed lower in the list. Exhaustive users, on the other hand, are willing to examine more results in weak queries. In strong quality queries, economic users are willing to go further in the list than if they have entered a weak query. If an economic user enters a weak query and the correct result is not in the top of the list, they stop their examination and requery.

Table 4.4:right shows the averages of our measures grouped by the type of query and user. We do not compute the number of query terms by query type as different questions may require different number of terms. Although query type is not a controlled variable, we notice that weak queries have fewer fixations and shorter time to action on average. The average time to requery for weak queries (excluding *NoCorrect* tasks) is 5.16 seconds and 7.42 for strong queries, and the result is statistically significant using unpaired two sided Student's t-test ($p = 0.017$).

4.3.4 Eye Fixation Sequence Analysis

Our original hypothesis is that query quality can affect user behavior. In the previous section, we saw that almost half of the requeries originated from weak quality queries. We investigated the examining behavior of users to understand the relationship between users, queries and search results. We are also interested in knowing whether the position of the relevant document matters in weak queries. To approach this, we computed the most common sequences of eye fixations that resulted in a requery. Table 4.6 shows our results. Sequences that are short or end before or at the page fold are often made by economic users. Shorter sequences also have a higher percentage of weak queries than longer sequences. In mobile, the most common sequence is 1→2→3. The next two sequences include 4→5 and 4→5→6. In order to view the 4th, 5th, or 6th result, users would need to scroll down as the page fold line is after the 3rd result. This explains why the 4th and 5th result were included in the mobile users' decision tree.

Table 4.6: Top 5 sequences that resulted in a requery action.

	Eye Fixation Sequence	Coverage	Weak Queries %	Strong Queries %	Users Econ. %
Desktop	1→2	0.11	61.54	38.46	100.00
	1	0.09	45.45	54.55	100.00
	1→2→3→4→5→6→7	0.09	27.27	72.73	90.91
	1→2→3	0.07	50.00	50.00	75.00
	1→2→3→4→5→6	0.04	40.00	60.00	80.00
Mobile	1→2→3	0.18	50.00	50.00	–
	1→2→3→4→5	0.11	16.67	83.33	–
	1→2→3→4→5→6	0.09	40.00	60.00	–
	1→2→3→4	0.07	25.00	75.00	–
	1→2	0.07	100.00	0.00	–

4.4 Summary and Conclusion

While it is well known that users are less likely to examine lower ranked search results, we show that regardless of rank, if a user sees a relevant result, the user will click it with high probability. We confirm our hypothesis in our prior work in Chapter 3 that the *exhaustive* and *economic* user types as characterized by Aula et al. (2005) play a significant role in understanding requeries without clicks. What drives a user's examination to end their search process at certain ranks in the search result? We find that certain ranks and display issues affect user examination patterns. However, most interestingly, we found that the quality of a user's query may be known to the user, and the user will modify their examination pattern based on query quality. This gives us an understanding of the likelihood of people to examine certain ranks under different types of query quality and can be seen as motivation to design effectiveness measures that include factors other than the relevance of search results.

In particular, in this work, we show that:

- Under our search tasks, users seem to enter queries that fall under two categories: queries that are considered ambiguous or under-specified (denoted as weak) or queries that are more specific to the task's question (denoted as strong).
- The first three search results are special. If a user issues a query unlikely to produce good results (i.e. weak queries), the user is more likely to reformulate after finding the top three results to be non-relevant than if the user had issued a query expected

to produce good results. If a relevant document is in the first three search results, the user will click on it.

- If the user is an exhaustive user, they are less influenced by the quality of their queries and are more persistent than economic users. Rank has much less effect on their likelihood of viewing a relevant result than it does for economic users.
- Economic users are unlikely to scroll and view search results off of the page, and thus are likely to requery when the topmost relevant result is below the page fold.
- We provide a view of the search process and abandonment encompassing three important parts, users, queries, and search results, and show the influence of users and queries to each other at specific ranks in the search result.

Chapter 5

Visualizing Searcher Gaze Patterns

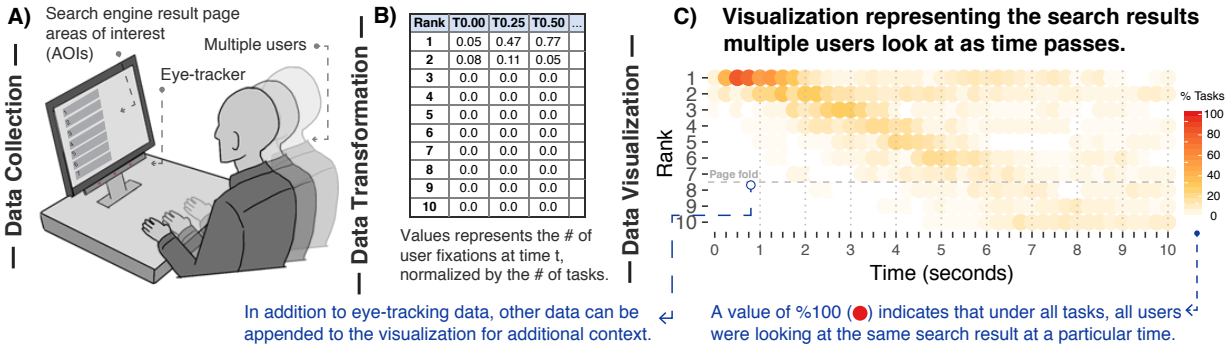


Figure 5.1: Overview and an example of the method used to visualize eye-tracking data of search engines result pages.

In the previous chapter, we used an eye-tracker to understand the process of query abandonment in more details. In this chapter, we discuss a method for visualizing eye-tracking data that is suitable for the typical SERP interface where search results are placed in a linear order. As opposed to the typical eye-tracking heatmaps visualization, this method allows integration of timing data into the visualization, as well as other useful information. We use our eye-tracking collected as part of our study in Chapter 4 to display some visualizations and show its usefulness in communicating different search behavior. Due to color variation in the visualizations, this chapter is best viewed in full color on paper or on a high definition screen.

5.1 Introduction

Eye-tracking is an important tool for understanding and analyzing searcher behavior (Goldberg et al., 2002; MacFarlane et al., 2017; Dumais et al., 2010; Buscher et al., 2009; Liu et al., 2015). Eye-trackers report to the researcher the gaze location of a computer user, called a "fixation". Even short experimental sessions generate a large stream of data, e.g. number of fixations, fixation duration, fixation location, etc. Data visualization allows for the quick summarization of these complex data streams and facilitates exploratory data analysis (Blascheck et al., 2017), which is important to generating new hypotheses about user behavior and decision-making. For example, using eye-tracking attention heatmaps, researchers were able to determine the F-shaped reading pattern, or Golden Triangle, that we described in Chapter 2.1.3 (See Figure 2.2).

The spatio-temporal structure of gazing data allows for different and unique visualization techniques. In this chapter, we review some of the existing visualization techniques and then show a visualization for temporal and Area of Interest (AOI)-based gaze data that is suitable for the typical "10 blue links" search engine interface. Our visualization is suitable for scenarios where AOIs are built in a linear (or somewhat linear) ordering and where the fixation locations within the AOI are not needed. Unlike some existing techniques, the visualization we propose allows us to combine data from multiple searchers and include timing information, while avoiding unnecessary visual clutter. We compare the method of visualization with eye-tracking attention heatmaps, and show the value of the proposed method of visualization in understanding user behavior and in communicating different patterns of searchers' gaze behavior.

Like Räihä et al. (2005), the visualization method shown in this chapter is designed to either combine eye-tracking data from many people or to show individual sessions, while incorporating time. In the following sections, we explain the process of generating the visualization and provide examples of the visualization demonstrating different patterns of gazing behavior from a previous search engine user study. We base these visualizations on data from our previous user study in Chapter 4. While the statistical analysis reported in that user study confirmed the differences between the searcher and query types illustrated by our visualizations, the visualizations themselves provide additional insights into the scope and nature of the differences.

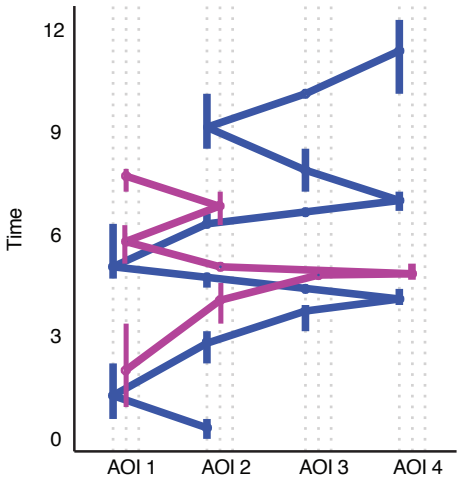


Figure 5.2: Example of two users AOIs examining behaviors. Y-axis indicates time. Based on Raschke et al. (2012)

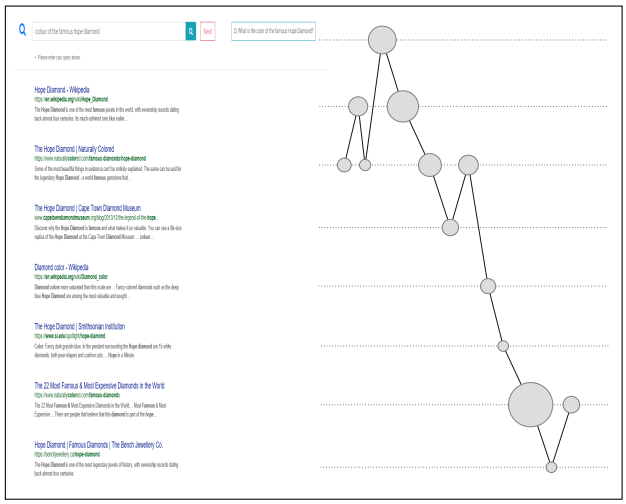


Figure 5.3: An example visualization of single user examining behaviour on a SERP. X-axis indicates time. Based on Räihä et al. (2005)

5.2 Visualization Method Overview

A typical data collection process in an IR eye-tracking study involves using an eye-tracker to track users' eye fixations within a monitor screen while they interact with a search engine interface and complete some search task. For the visualization, we assume a typical search bar and ranked list of results presented linearly. While today's SERPs contain a variety of components such as ads, verticals, and knowledge graphs that may not be linearly ordered, many user studies still employ the typical "10 blue links" search engine interface to study different aspect of the search process. We believe this visualization method would still be applicable and useful for researchers.

As an abstraction method, we built AOIs around each search result to determine when and if a user examined a search result at a specific rank (See Figure 4.5). Figure 5.1A shows an example of this eye-tracking data collection process. Using the eye-tracking data, we abstract the data into multiple interaction periods, each starting from the moment a SERP is presented to the user, to the time the user makes their first action, e.g., a click or an abandonment of the search result. With time being a key variable, we can determine how many users were looking at a specific rank at a particular point of time during their interaction period.

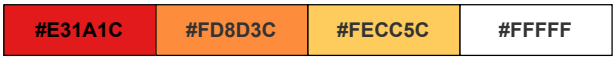


Figure 5.4: Color encoding used in our visualization.

We normalize the values based on the number of search tasks in the data, such that a value of 100% would indicate that all users under all tasks in the group were looking at the same rank during a particular time period. Figure 5.1B shows an example of the transformed data resulted from the abstraction steps.

The visual encoding design is based on the values in the transformed data. The x-axis indicates time, and the Y-axis indicates the rank of the search result. The color encoding of data points (Figure 5.4) was chosen to indicate intensity while adhering to perceptual ordering, an important element in the color theory of information visualization (Munzner, 2014, Chapter 10.3.2).

We then use `R`'s `ggplot2` library to implement the visual encoding and execute the visualization technique. We provide the `R` code for researchers to experiment with and generate visualizations from their own eye-tracking data in Appendix E.1. Figure 5.1C shows an example of the final output of the visual encoding.

The visualization can be useful in understanding gaze patterns of searchers, communicating how far down the ranking people examine and how quickly they examine the results. It also allows for the inclusion of other relevant data, such as the location of the "page fold" where the searcher was required to scroll. Such information embedded in the visualization can provide additional context that can be useful to increase our understanding of the data.

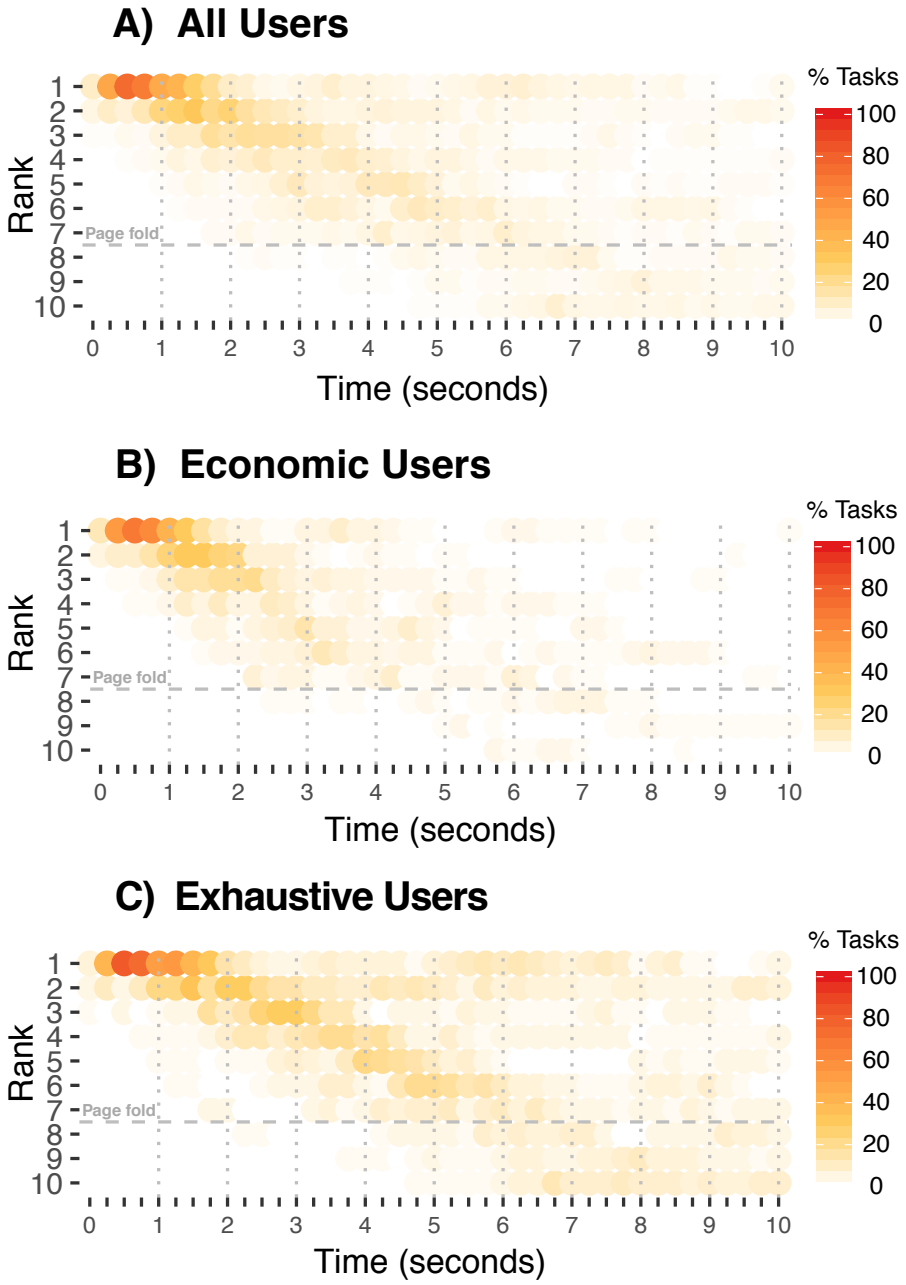


Figure 5.5: Example of our visualization using eye-tracking data from the user study in Chapter 4 for search tasks where the only relevant document is below the fold (rank 8, 9, or 10), or when there is no relevant documents in the list.

5.3 Study Data

We base our visualizations on the data from the user study described in Chapter 4. In that work, eye-tracking data was collected to study query abandonment in web search — the behavior of abandoning search results without any clicks — while controlling different qualities of SERPs. In each task in their study, participants were asked to use a search engine interface to find an answer to a simple factoid question (e.g. "How many chapters are in the art of war book by Sun Tzu?"). The interface was designed to appear similar to commercial web search engines and returned 10 search results per query with no pagination. The page fold in the interface occurs after the 7th search result. This is the location where searchers would need to scroll down the page to view the rest of the search results.

In total, 24 users completed 12 tasks in a balanced order. In 11 of the tasks, the results of the searcher's first query were manipulated to include either one relevant result, containing the answer to the question, placed at ranks 1 to 10, or no relevant results at all. The 12th task returned the non-manipulated search results from the Bing API as a control. The set of non-relevant and relevant documents for each factoid question were chosen prior to the study. This data was then analyzed to determine factors affecting the examination and abandonment of search results.

This eye-tracking data includes information on the task type (i.e., where the relevant result is ranked or if it exists in the search results), the time the user issued their query, the time and duration of each fixation, and 10 Boolean variables indicating whether the fixation is within one of AOI representing the 10 search results.

In addition to eye-tracking data, the data also includes information on the user type, i.e. whether they are considered an *economic* or *exhaustive* user, and the type of each query (*weak* or *strong*).

In our user study described in Chapter 4, the user type was determined from the distribution of the average total number of fixations by users during their search tasks. This definition of economic vs. exhaustive users follows prior literature, where economic users are those that typically make their decisions (e.g. to click or to requery) "faster and based on less information than exhaustive" users (Aula et al., 2005).

To label queries by type, two assessors were hired to judge the queries submitted by the searchers for each question to indicate whether they should be considered as under-specified queries (which were labeled as "weak") or queries that are more specific to the question (labeled as "strong"). Details of the assessment is described in Chapter 4.

5.4 Results and Discussion

One question to investigate is the willingness of people to scroll beyond the page fold to view more search results. Using only data from tasks where the relevant result is placed below the fold or when there are no relevant results in the list, we plotted the visualization to see if and when people examine these low ranking search results.

Figure 5.5A shows the visualization of all searchers under such tasks. We notice that people start to examine results below the fold (area below the horizontal dashed gray line) after about 5 seconds. Prior literature (Aula et al., 2005) indicates that economic searchers tend to process results and make their actions faster than exhaustive searchers. Figure 5.5B&C show the visualization under the two types of searchers. Here, we see examination of low ranking results is mostly done by exhaustive searchers, whereas economic searchers take their next action without examining results below the fold.

We also explored the gaze patterns of searchers under different query types. *Strong queries* are those that are more specific, unambiguously defining the searcher's information need. *Weak queries* are less specific and more ambiguous.

Figure 5.6 shows the visualizations under each group. For comparison, Figure 5.8 provides the same visualizations using traditional heatmaps. From Figure 5.6, we notice how gazing behavior changes under the four groups. When economic searchers submit a weak query, they examine fewer search results than if they issued a stronger query, as shown in the top-left part of the visualization. In contrast, exhaustive searchers keep examining results and even scroll below the fold. Economic searchers stop once they reach the fold. Another apparent difference in the behaviour between economic and exhaustive searchers can be seen where exhaustive searchers fill the upper diagonal of the visualization more than economic users and appear to spend more time examining each result. For example, when comparing Figure 5.6A and 5.6B, we see that exhaustive searchers spend more time scanning down the ranked list than economic searchers. While heatmaps (e.g. Figure 5.8) can be useful, such timing information cannot be deduced from the heatmap figures alone. For example, in many of the figures of our visualization, the most "red" point is not at the very left. This is because of the differences of the time to examine the first results, and because some users started their first examination on the second item. This is one example where the method of visualization can be useful for researchers to visualize variability in examination, and can compliment traditional heatmaps.

5.5 Conclusion

We presented a method for visualizing searcher's gaze patterns from eye-tracking data. The visualization can be useful in communicating differences between searchers' gaze patterns. For example, Figures 5.5 and 5.6 enable us to quickly and easily visualize gaze patterns and investigate differences in gazing behavior between different types of searchers and queries. We believe the visualization is useful in other experiments as well, such as gaze patterns while searching for answers to factoid vs. complex questions, or gaze patterns when including other SERP components such as knowledge boxes or images.

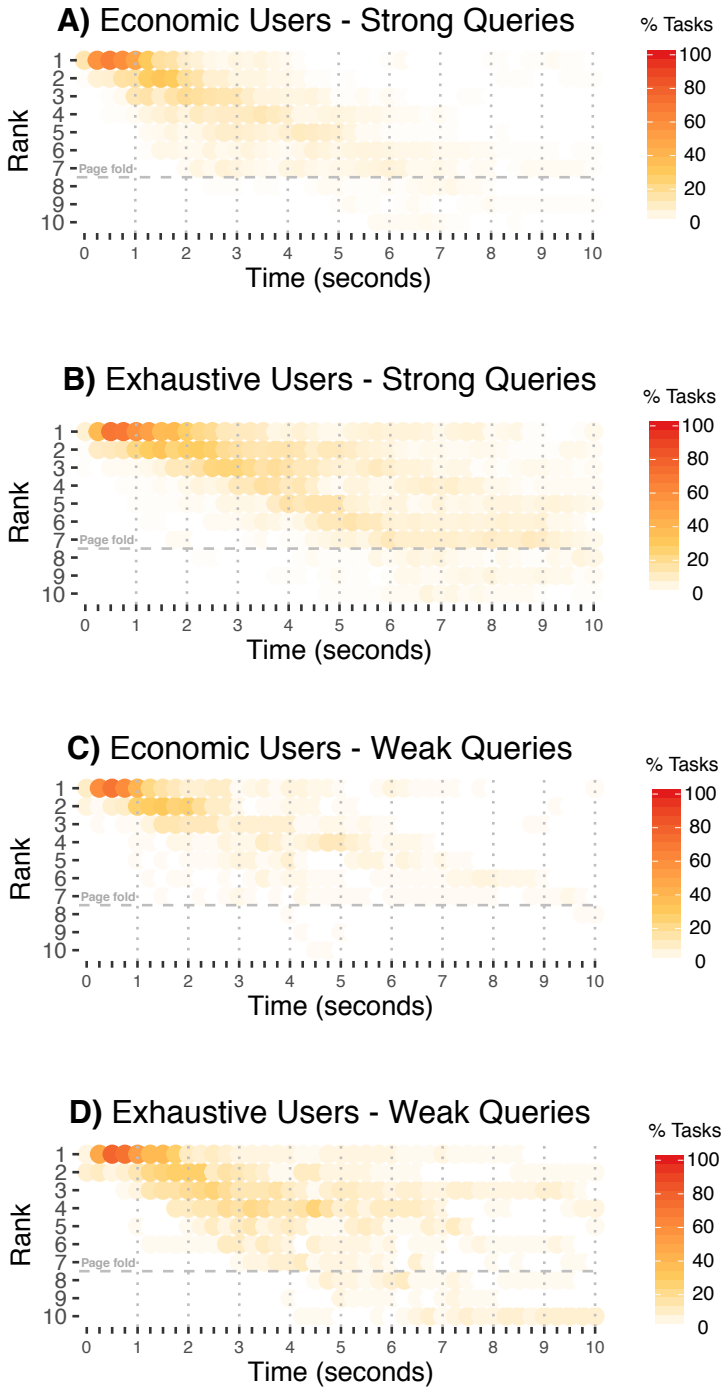
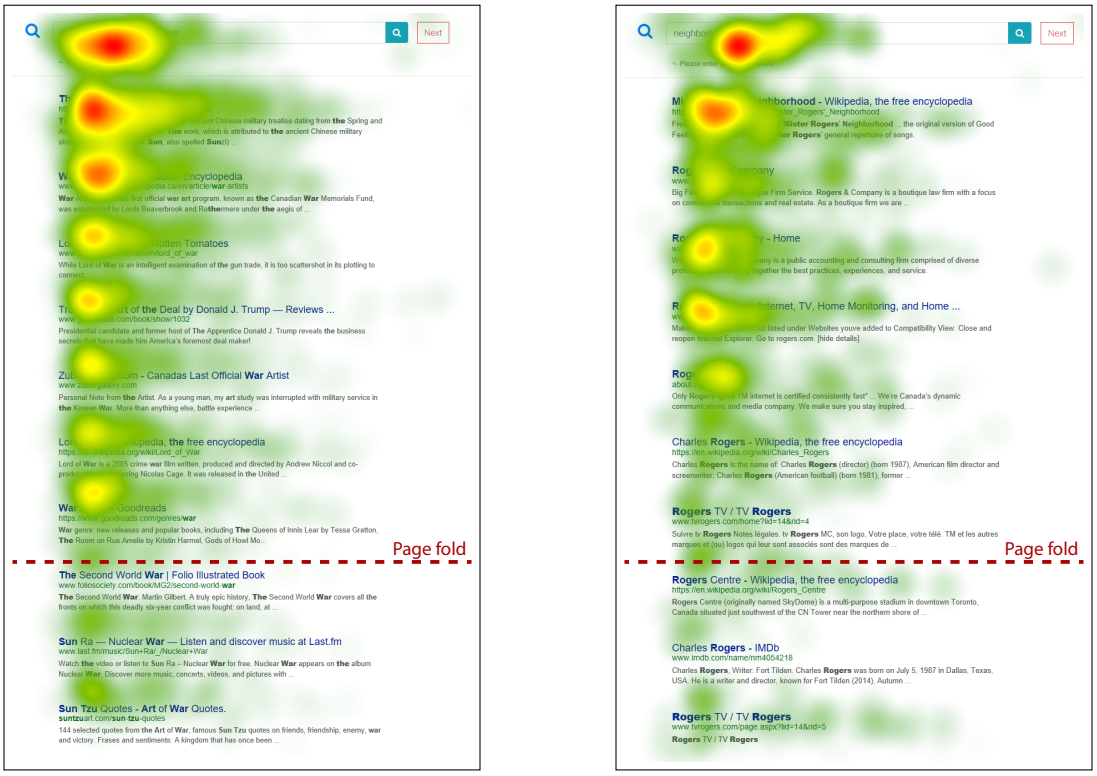


Figure 5.6: Our visualizations for different types of users under different quality of queries.

(a) Economic users on strong queries

(b) Economic users on weak queries

Figure 5.7: Visualizations using relative duration attention heatmaps generated by the eye-tracking software using default settings. The above figures are for economic users. These may be compared with the corresponding visualization in Figure 5.6.



(a) Exhaustive users on strong queries

(b) Exhaustive users on weak queries

Figure 5.8: Heatmaps for exhaustive users. These may be compared with the corresponding visualization in Figure 5.6.

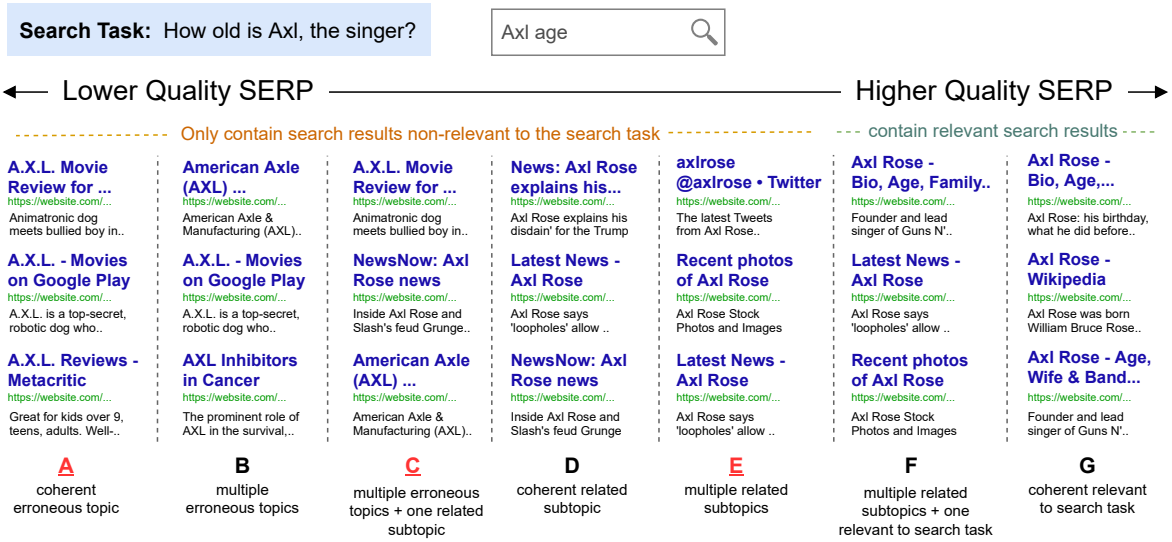
Chapter 6

The Effect of Non-Relevant Results on Mobile Search Behavior

6.1 Introduction

After a user submits a query to a web search engine, the user interacts with the search engine results page (SERP). The user's interaction with the SERP is influenced by the content of the SERP and how the user perceives it. From our past research, we know that users are more willing to examine down the ranked list if they cannot find relevant results. If they do find relevant results, they are less likely to continue examining down the ranked list. This simplified notion of examination, of course, has limits. For example, many users are unlikely to even scroll to see more search results if the visible results are all non-relevant. Abualsaud and Smucker (2019), and Azzopardi et al. (2020) have shown that the degree of document relevance influences the extent to which people continue to examine search results or not.

Moffat and Wicaksono (2018) have proposed that the likelihood of users continuing to examine search results is modulated by not just the presence or absence of relevant results, but by the nature of the non-relevant results. They propose that as users encounter *egregiously non-relevant* results, they are less likely to continue examining search results. In their paper, they called for a user study to investigate the actual behavior of users as they encounter different degrees of non-relevant results. Not only is this chapter an answer to Moffat and Wicaksono's call for a user study, but we go further by examining how different types of non-relevant documents, and different types of SERPs, can affect user behavior.



We adopt Moffat and Wicaksono's proposal to broaden the notion of what it means for a document to be non-relevant. We consider a search result to be *egregiously non-relevant* if it is considered far off from the search topic, and *within-topic non-relevant* if it is non-relevant to the search tasks but is related to the search topic in some sense.

For example, let us imagine a user who is interested in knowing the age of Axl Rose, the celebrity singer. In this search task, a relevant search result should contain information about the singer's age (e.g., the singer's Wikipedia page). It is reasonable to say that search results regarding the singer's online photo albums or websites on the singer's latest breaking news are both subtopics related to Axl Rose, but not relevant to the user's search task (i.e., finding the singer's age). We consider these to be within-topic non-relevant documents. Other search results that contain the term "axl", such as the A.X.L movie or the AXL gene, are search results we consider as egregiously non-relevant, as they have nothing in common with the singer Axl Rose.

In this work, we turn our attention to understanding user behavior when the SERP has no relevant documents. We first define a spectrum of different possible ways a SERP can be designed and constructed (see Figure 6.1). The left end of the spectrum represents the worst possible SERP. Here, the SERP contains coherent search results on a single egregious topic (e.g., Figure 6.1A). As we move from left to right along the spectrum, the SERP quality improves with different mixes of non-relevant and relevant documents (e.g., Figure 6.1B-F). The far-right is the best of the spectrum, and here the SERP is of high quality and contains highly relevant documents (e.g., Figure 6.1G).

To understand how different SERP qualities influence users' interaction, we conducted a user study where we asked participants to search for answers to simple factoid and informational questions. In each search task, when a participant submitted their first query that contained relevant terms to the search task question, we showed them manipulated results representing one of three SERPs in our spectrum: search results coherent on an egregiously non-relevant topic (Figure 6.1A), search results on multiple egregiously non-relevant topics and one within-topic non-relevant result (Figure 6.1C), or search results on multiple topics that are within the topic of the search but not relevant to the task (Figure 6.1E). These SERPs contained only non-relevant search results but differed in their coherence and the types of non-relevant results included. We logged user behavior while participants interacted with the SERPs so that we could analyze differences in user interactions.

From our user study, we show that:

- Users' interactions are influenced differently by the type and quality of the SERP presented to them. While every manipulated SERP contained only non-relevant documents, when users were shown egregiously non-relevant results, the fraction of users requesting to view more results at least once is a low 0.28. The fraction jumped to 0.41 when we included one subtopic-related result among other egregiously non-relevant results, and further increased to 0.56 when users were shown a SERP containing within-topic non-relevant search results.

- The time it takes users to abandon the SERP is different depending on the quality of the SERP presented to them. Users spent a median of 5.4 seconds when the SERP was the lowest quality in our spectrum, i.e., when it only contained egregious search results. The time increased as the quality of the SERP improved. When users were shown SERPs containing multiple within-topic non-relevant search results, users took about 7.11 seconds before abandoning the results.

- When users were presented with a SERP containing search results coherent on a single egregious topic, users would abandon the search result with a high probability (0.95). The probability decreased to 0.87 when the SERP contained a lesser amount of egregiously non-relevant results, and down to 0.79 when the SERP had no egregious results and only contained subtopic related non-relevant search results. While all the SERP results contain no relevant information to the search task, this result indicates that users may incorrectly click on non-relevant documents when the results seem encouraging.

This experiment shows how examination behavior changes depending on the quality of the SERP and whether it seems encouraging or discouraging towards users' information needs. We consider these results to have important implications on the design and evaluation of search systems and how relevance labels are collected in IR. We discuss these implications further in Section 6.5.

6.2 Different Search Results Scenarios

The utility of a SERP can differ depending on what information it can provide to the searcher. A search result page that contains no useful information has less value than a search result page that leads the user to find relevant information, even if it does not contain anything relevant to the user's information need. There are different possible ways search results can be designed. In Figure 6.1, we show possible scenarios a search engine might return for the search task "Axl age". We explain what these SERPs can represent and why we believe the utility of the search results are better from scenario **A** to **G** in Figure 6.1.

- **A (coherent egregious topic)**: The search results are coherent with each other and are all related to a single egregious topic. For example, in Figure 6.1, all three search results are on A.X.L the movie, which is not related to the actual search topic (Axl, the singer). This scenario represents a search engine that completely fails to understand the user's information need and focuses on returning results on a single egregious topic.
- **B (multiple egregious topics)**: The search results are related to different egregious topics. In the figure, all three search results are not related to the actual search topic nor to each other. We consider these search results to be better than the previous one because it represents a search engine that attempts to diversify the search results but was not successful in returning anything related to the user's information need.
- **C (multiple egregious topics and one within-topic non-relevant)**: This scenario is similar to **B** except it contains a search result (news on Axl Rose) that is related to the search topic but not relevant to the actual search task (age of Axl Rose). While the results do not help the user find what they are looking for, we consider the results better than the previous scenarios.
- **D (coherent within-topic non-relevant)**: In this scenario, the search results are coherent with each other and are on a single topic (Axl Rose News) that is related

to the search topic but not relevant to the search task. This scenario can represent a search engine that understood the searcher's topic but did not return anything relevant to the search task. Instead of diversifying the search results in the search topic to hopefully include relevant results, the search engine focuses on a single topic. We believe this scenario is better than the previous ones as it excludes all egregious search results.

- **E (multiple within-topic non-relevant)**: This scenario is similar to **D**, except the search results are diversified within the search topic. For example, in Figure 6.1E, the search results contain three subtopics: Axl Rose social media, Axl Rose photos, and Axl Rose news. While the search results do not contain anything relevant to the search task, it attempts to diversify the search results in the search topic.
- **F (multiple within-topic non-relevant, including one relevant to search task)**: This is similar to the previous scenario **E**, except the search results include one item that is relevant to search task and contains relevant information that satisfies the user's information need.
- **G (coherent relevant to search task)**: All search results are coherent and contain relevant information to the search task.

6.3 Hypotheses

While the three scenarios in our user study (**A**, **C** and **E**) do not contain any search results that are relevant to the search task question nor contain the correct information, we suspect that user search behavior under these conditions will differ. In particular, we hypothesize that:

- **H1:** The fraction of users clicking the "view more" button is the lowest when users are shown results coherent on a single egregious topic (**A**). In other words, when search results seem to be moving away from leading the searcher to the correct information, the searcher will be less inclined to view more items within the search page.
- **H2:** When users are shown search results coherent on a single egregious topic (**A**), they will abandon their search results faster than users shown search results that somewhat seem promising yet do not contain a correct result (e.g. **C**, and **E**). In other words, as the overall representation of the search results seems to be moving away from leading the searcher to the correct information, the searcher will abandon the search results faster.

6.4 User Study

We created a user study to address our hypotheses. In our study, we focus on comparing user behavior under the three scenarios: **A (coherent egregious topic)**, **C (multiple egregious topics and one within-topic non-relevant)** and **E (multiple within-topic non-relevant)**. We choose these three scenarios because their search results do not contain any relevant item to the search task, and have various degrees of irrelevance. In this work, we want to understand search behavior where users fail to find anything relevant to our search task questions. ’

Our original experiment plans called for us to use our lab’s eye-tracker with participants on both mobile and desktop search interfaces. Since the start of the COVID-19 pandemic (March 2020), and to today (May 2021) our university has disallowed in-person studies such as ours. To enable us to conduct this experiment remotely via Zoom, and be able to monitor where in a results list a participant was viewing, we modified the experiment to work on mobile devices only, which have a small viewport. Section 6.4.4 discusses the user interface in more detail.

6.4.1 Experimental Protocol

Figure 6.2 shows an overview of the experiment protocol. The consent form is shown in Appendix F. Participants were given access to our search engine and were asked to use our search engine to find an answer to their search task question (e.g., “How old is Axl, the singer?”). Participants needed to complete the study while sharing their screen via Zoom. We started the study by collecting demographic information from the participants. Participants were then redirected to a tutorial page where we explained the user study task and expectations. We provided a practice task to allow participants to familiarize themselves with the search interface and the process of completing a task. During the practice task, we used the Bing API to return search results for submitted queries. Once they completed the practice task, participants proceeded to the main study tasks. Each study task consisted of a pre-questionnaire, the actual searching task, and a post-questionnaire. The questionnaires helped familiarize the participants with the question, collect information on topic familiarity, and collect any feedback participants wanted to provide. A search task is completed when the participant announces their answer to the researcher. Participants completed 12 of these tasks in a balanced order. After finishing all the tasks, participants filled an exit questionnaire on their experience and answered some questions from the researcher regarding their search behavior.

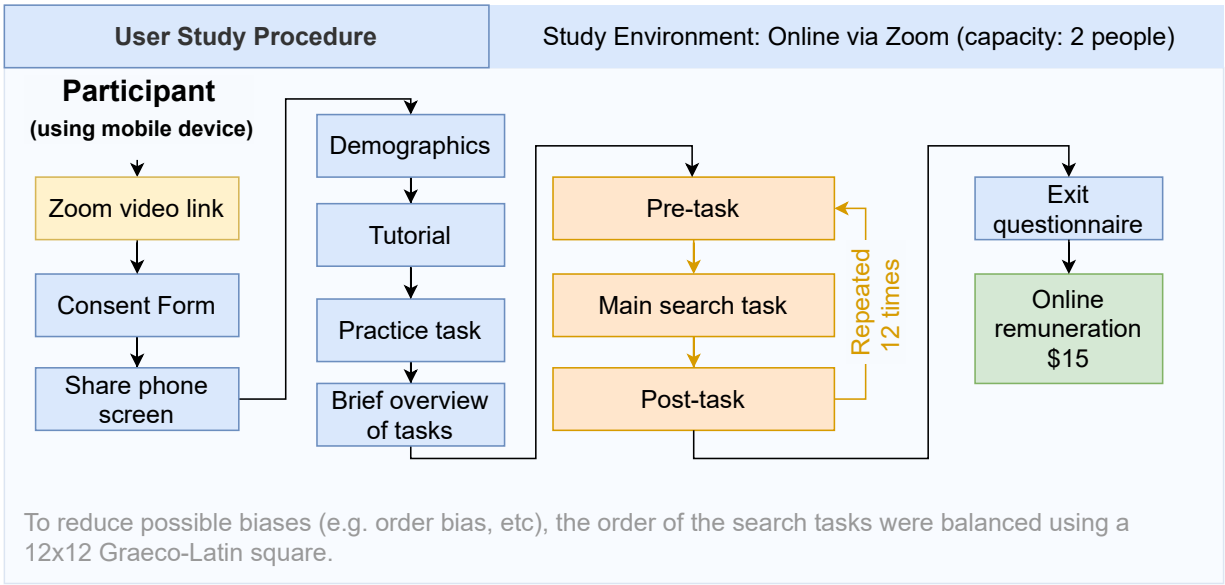


Figure 6.2: User study procedure.

6.4.2 Tutorial

The tutorial in this user study was the same as in the study in Chapter 4 (see Appendix C.1). The only differences is that we show screenshots of the search interface on a mobile device. Participants were also told *"You can view more results by clicking the view more button at the bottom of page"*.

6.4.3 Search Tasks

We asked participants to complete 12 search tasks. In each task, the participant needed to use our search engine to find an answer to their search task question. The questions require no prior knowledge and are likely to be familiar to participants. The list of search tasks are shown in Table 6.1. In six of these tasks, the search engine results were manipulated to show results of varying qualities. The other six tasks had search results returned from the Bing API. These tasks are added to ensure participants do not notice irregularities in the quality of search results presented to them. We instructed participants to stop the search process once they were confident about their answer and to say the answer out loud to the researcher. We used topics from our previous work in Chapters 3 and 4, and the TREC Web 2014 and 2012 Tracks. The search tasks were designed to be simple as not to confuse

Table 6.1: A list of the search tasks used in our user study. The related subtopics and erroneous topics are used to construct the search engine result page for our tasks scenarios. The cells shaded in blue in the erroneous topics column refers to the topic chosen for tasks under scenario (A). The cells shaded in orange in the related subtopics column refers to the subtopic shown in the search results for tasks under scenario (C).

#	(Topic) / Search Task	Related subtopics (not relevant to search task)	Erroneous topics
1	(Holes novel by Louis Sachar) **What is the publication date of holes by Louis Sachar?**	Holes Movie (based on novel) Holes Soundtrack and Music Classroom activities related to "Holes" novel	Golf holes-in-one Black holes
2	(UN world heritage sites) **What site was selected as Canada's United Nation world heritage sites in 2016?**	America/Europe countries world heritage sites Heritage sites selection criteria UNESCO's activities	Toursim Canada trips Canadian history and heritage
3	(Art of War book by Sun Tzu) **How many chapters are in the Art of War by Sun Tzu?**	Quotes from Sun Tzu Comparisons of Sun Tzu and Machiavelli Sun Tzu's Art of War applied to business	Art and exhibitions related to war The War of Art by Steven Pressfield (Book)
4	(Mister Rogers' Neighborhood tv show) **What is the opening theme song for "Mister Rogers' Neighborhood" tv show?**	Biographical information for Fred Rogers Quotes from Mister Rogers Characters in Mister Rogers' Neighborhood	Neighborhood festival Rogers network
5	(Mountain Goats music band) **How many members are in Mountain Goats band?**	Mountain Goats tickets Mountain Goats album reviews Mountain goats band social media	Mountain goats (animal) Goat Mountain trail
6	(Axl Rose singer) **How old is Axl, the singer?**	Axl Rose latest news Axl Rose photos Axl Rose twitter and social media	A.X.L. Movie American Axle & Manufacturing H (AXL)
7	(Doom video game) **Find information about Doom, the video game.**	Doom movie Doom Soundtrack and music Doom novel series	Doom Mountain Doom (Japanese band)
8	(Learning Golf) **Find information on how to choose a good golf school.**	Golf instructional videos Online instructions/tips for putting Golf tournaments latest news	Golf online video games Volkswagen Golf Back to School
9	(Figs fruit) **Find information on nutritional or health benefits of figs.**	Recipes that use figs The different varieties of figs Growing figs	Figs & Olives Toronto (resturant) Fig Tree Cave
10	(Fidel Castro) **Find some quotes from Fidel Castro, the Cuban prime minster.**	Health of Fidel Castro news Ozzie Guillen and Fidel Castro relationship Fidel Castro's family members	The Castro (neighbourhood in CA) Castro (clothing)
11	(Yoga exercise) **Find information on yoga for seniors.**	Yoga poses tips and lessons Yoga during pregnancy Benefits of yoga for kids	Yoga (Hindu astrology) Yoga pyrops (fish)
12	(Barcelona FC) **Find information on the history of Barcelona, the football club.**	Barcelona FC tickets Barcalona FC transfer news Barcelona FC gear store	City guide of Barcelona Barcelona demographics

people and to reduce confounding variables that may arise.

6.4.4 Search Interface

Figure 6.3 shows an example of the search interface on an iPhone X. The search task question is shown at the top of the page. A search box is provided to allow users to enter their search queries. Query suggestion was not provided in the interface. Once a user submits a query, three search results were shown to the user by default. We decided to show three results because most phones can fit three search results within the page fold. The page fold line is the line between the part of the page you can see without scrolling and the part of the page you can see when you scroll down the page. To view more results, users would need to click on the "More results" button at the bottom of the page. When a user requests more results, another set of search results will be added to the end of the SERP. In our interface, we add three search results each time a user requests more results. The interface allows up to 15 search results to be shown in the page.

In our study, we wanted to obtain a good way of recording how many results people examine. While an eye-tracker would suffice, unfortunately, we are not able to conduct eye-tracking user studies due to the pandemic. By using the "More results" button, we were able to record the depth of examination, even if it may have reduced people's willingness to go further.

6.4.5 Constructing Search Results Pages

For each topic in Table 6.1, we created three related subtopics and two egregious topics that share some of the topic's keywords. For example, for topic #6 "Axl Rose" where the search task is to find the age of Axl Rose, we consider subtopics such as "Axl Rose social media" or "Axl Rose latest news" to be subtopics related to the singer "Axl Rose" but not relevant to search task on finding the singer age. These subtopics are what we consider as within-topic non-relevant. Topics such "A.X.L movie" or "American Axle & Manufacturing (AXL)" are egregious topics that are not related to the singer. Some of these related subtopics and egregious topics were directly copied from TREC web tracks topics, while others were created by ourselves.

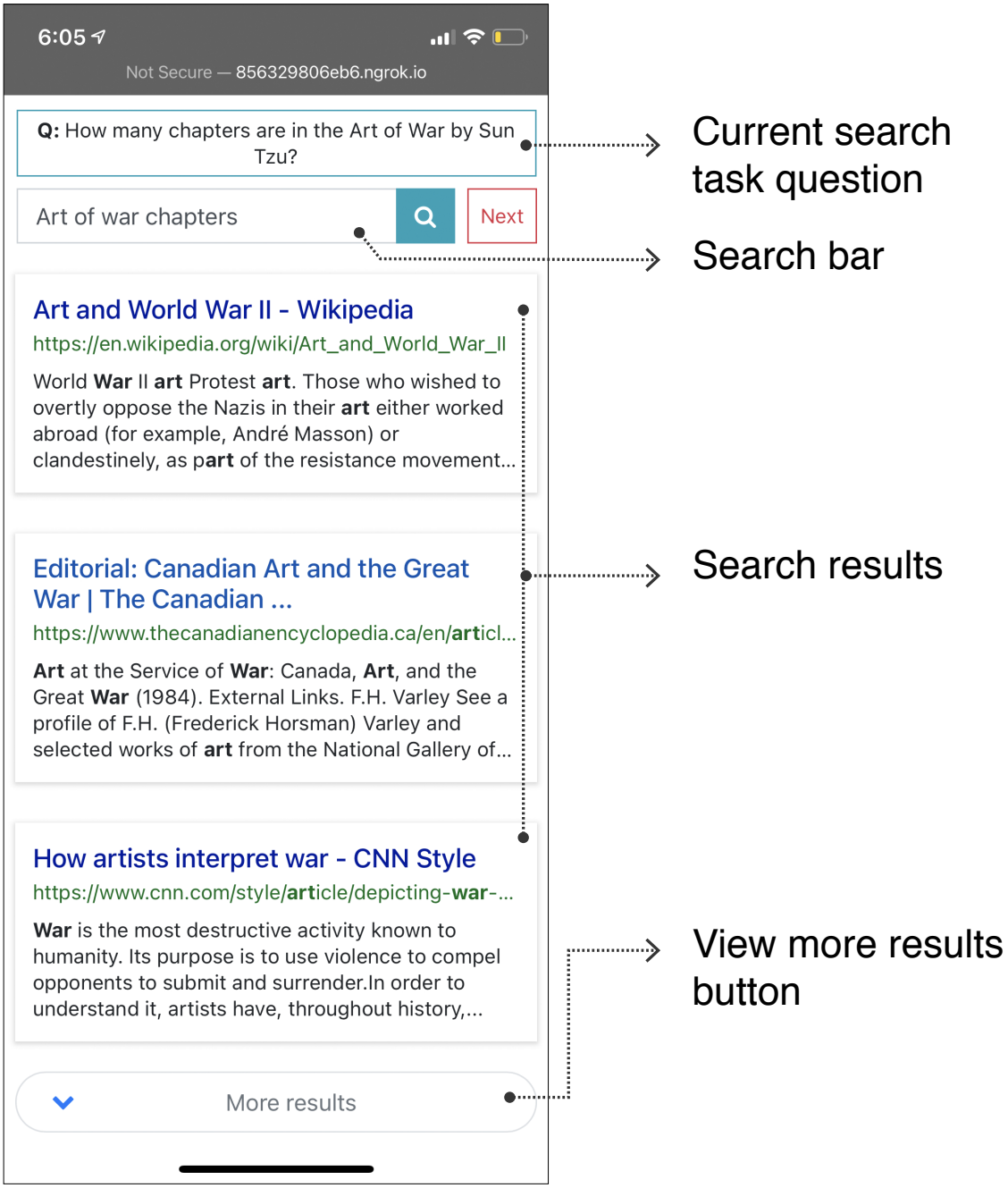


Figure 6.3: Screenshot of the search interface. The interface shows three search results by default. Clicking on more results shows an extra three results, up to 15 results for each query.

Related Subtopics and Egregious Topics

To construct coherent search results on an egregious topic for scenario (**A**), we selected a single egregious topic from Table 6.1 for that task (shaded in blue). For scenario **C**, we selected the two egregious topics from Table 6.1 for that task and one related subtopic (shaded in orange) for our within-topic non-relevant. For this scenario, every three search results in the SERP contain one of each three topics in random order. This guarantees that the user was shown the egregious and the within-topic non-relevant search results within the page, without needing to scroll down or click on the "view more results" button. Finally, for scenario **E**, we selected all three related subtopics from Table 6.1 for that task. Again, every three search results in the SERP contained one of each three subtopics in random order.

Finding Search Results for Each Scenario

For each scenario, we used the Bing API to query the subtopic or the egregious topic to find related search results. We selected 15 search results from the Bing API to show users for that scenario. Actual examples of search results are shown in Figure 6.1 for task #6 and in Figure 6.3 for task #3 under the single egregious topic (**A**) scenario.

When are Search Results Shown?

Our manipulated search results for conditions **A**, **C** and **E** are shown once a user submits a query containing any relevant keywords for the task, similar to the technique described in Chapter 3. For example, for the task on Axl Rose, any query that contains the term "Axl" will trigger our manipulated search results to be shown to the user. All subsequent queries during the task will use the Bing API to fetch the result.

6.4.6 Study Design and Procedure

Balanced Design

Each participant completed 12 search tasks. Out of the 12 tasks, there were two tasks for each of our three treatments. The remaining six tasks were tasks where participants received results from the Bing API. The purpose of these Bing tasks is to make sure people do not notice our manipulations and to induce normal behavior when the manipulated

SERPs are shown. To mitigate topic or order biases, we used a 12×12 Graeco-Latin square to balance the search topics and treatments across task order. The square forms a single block where each row represents the order of tasks that a participant completes.

Implementation

The web application in Figure 6.3 was built using the Django web framework and JavaScript (JS). JS was used for client-time tracking of user behavior, such as clicks, keystrokes, and dwell time. The server was hosted locally and participants accessed the server using ngrok[1].

6.4.7 Participants

After receiving ethics approval from our university's Office of Research Ethics, we advertised the user study in a mailing list for graduate students, the university's graduate studies affairs website, and two Reddit groups: the group associated with the university, and the group associated with the city where the university is located.

We recruited 26 participants in total. Two of these were used as pilot users to verify that our study procedure and our system work as expected. Four were removed from the analysis due to technical issues (e.g. slow internet connection or phone application crashing). In total, 20 users data were included in the analysis. Out of the 20 users, 9 are female, and 11 are male. Our participants included university students (16 undergraduate and 2 graduate students) and 2 professionals. Students were enrolled in STEM programs, art, environment, and social work. The average age of participants was 21.75, with a minimum age of 18 and maximum age of 33.

We provided a remuneration of $15 online payment to each participant as an appreciation of their time.

6.4.8 Collected Measures

We collected the following data from each participant:

Submitted queries: All queries submitted to our search engine during their tasks.

Action: The action the user has made once they are presented with the search results. An action could be a document click, or a good or bad query abandonment. Good abandonment only occurs during control tasks, where answers to task questions may appear in

[1] https://ngrok.com/

some of the search snippets. Bad abandonment is when a user leaves the search results because they are unsatisfied, without clicking on any document.

Time to action: The time users take to make their action, starting from the moment the search results are presented to the user to the moment the action is triggered. For abandonment, the action ends when the user clicks on the search bar.

Requests for more results: For each query, the number of times a user has requested to see additional search results in the SERP.

Time to more results requests: The time a user took to click on the "More results" button.

6.5 Results & Discussion

In our study, participants used our search engine to find answers to 12 search tasks shown in Table 6.1. For six of these tasks, we directly retrieved query results from the Bing API. For the remaining six tasks, there were two for each condition **A**, **C**, and **E** described in Section 6.2. In these tasks, we show users manipulated SERPs constructed prior to the study, representing different qualities. These SERPs are shown once a user enters a query containing any relevant term to the task. All subsequent queries results were fetched from the Bing API.

Figures 6.4, 6.5, 6.6, and 6.7 show our main results. In Figure 6.4 (left), we show how likely it is that users request to view more results when the SERP represents our different conditions. When search results are coherent on a single egregious topic (**A**), the fraction requesting more results is the lowest (0.28) compared to **C** (0.41), and **E** (0.56). Using a Chi-square test, Table 6.2 reports statistical significance on whether the condition type has an effect on whether users will request more search results. In the two conditions **A** and **E** where the SERPs are in the extreme opposite side of the spectrum, the difference is statistically significant ($\chi^2 = 6.35$, $p = .01$). In other words, the result shows that when the SERP contains promising non-relevant results, users are more likely to request to view more results compared to when they are shown the lowest quality SERP. The time users spend before requesting more documents is shown in Figure 6.5 and appears to be similar across the conditions, with condition **E** slightly lower than the other conditions. Using one-way ANOVA, we did not find any statistically significant difference in the time to first "more result" click ($F(2, 46) = 0.414$, $p = 0.66$).

Using the fractions of requesting more results, we calculated examination probabilities for different ranks. Since we display three additional search results each time a user requests

Table 6.2: Result of chi-square test of independence between experimental conditions and requests for more search results. Star symbol indicates statistically significance ($p < 0.05$).

	A	**C**
C	$\chi^2 = 1.41$, $p = .23$	
E	$\chi^2 = 6.35$, $p = .01^*$	$\chi^2 = 1.85$, $p = .17$

more, the ranks are grouped into sets of three. Figure 6.4 (right) shows the result, which indicates how lower-ranking search results are less likely to be examined on lower quality SERPs.

We also computed the probability of users making an abandonment or clicking at a document. Figure 6.6 shows the probability of the first action in each condition. Table 6.3 shows statistical significance analysis using Chi-square test. Between the two conditions **A** and **E**, the difference is statistically significant ($\chi^2 = 4.13$, $p = .04$). The probability is the highest when users are shown results coherent on a single egregious topic (**A**). As we move from condition **A** to **E**, the probability of bad query abandonment decreases, with the lowest probability when users are shown search results diversified to include multiple within-topic non-relevant (**E**). This result is interesting as it indicates users are more likely to click on wrong documents when the SERP as a whole appears encouraging, even when those documents do not contain any information related to the search task. Users who clicked on a wrong document returned back to the SERP and reformulated their query.

Figure 6.7 shows the time users take to abandon their queries. In other words, the plot shows how long before users decide that the search results are not worthy of examining and decide to reformulate their queries. When the search results are not relevant to the search task but include multiple related subtopics (**E**), users spend a median time of 7.11 seconds before deciding to abandon their query. The time is decreased to 5.8 seconds when the search results have egregious search results but contain documents about a single related subtopic (**C**). It is further decreased to 5.4 seconds when all the search results are coherent on a single egregious topic. We tested these differences using a one-way ANOVA. The analysis of variance showed that the effect of the type of condition on the time to query abandonment was not significant ($F(2, 99) = 1.739$, $p = 0.18$).

Figure 6.4 and Table 6.2 directly address and confirm our hypothesis **H1** on whether users would examine more results when the quality of search results worsen. The figure shows that users are less likely to view more when the results seem discouraging. Figure 6.7 addresses **H2**, where we hypothesized that users would abandon their search results the fastest when shown the most discouraging results. We could not confirm our **H2** as we did

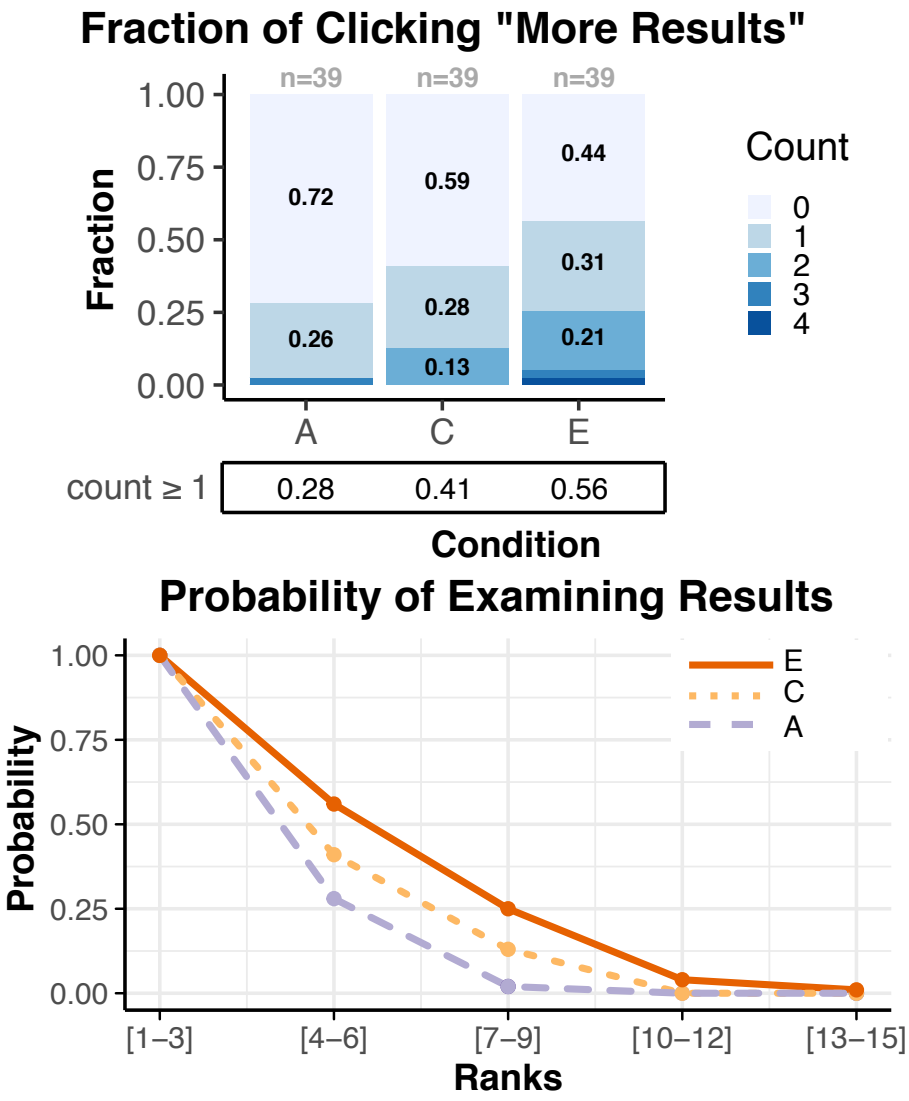


Figure 6.4: Top: The fraction of clicking at "More results" button under each condition. The count indicates the number of times a user requested to view more results. Among our three conditions, users' lowest fraction of viewing more results is when they are presented with results coherent on a single egregious topic (A). Users' highest fraction of viewing more results is when they are presented with search results containing multiple within-topic non-relevant results. Bottom: Based on the left figure, we calculated the probability of examination at different ranks. The ranks are grouped into three elements because we show three results in the SERP each time a user clicks the "More results" button.

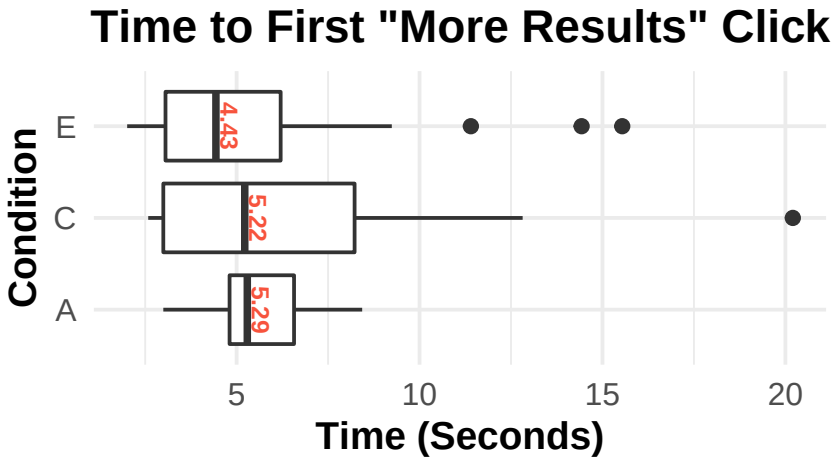


Figure 6.5: Time to first click on "More results" button under each condition.

Table 6.3: Result of chi-square test of independence between experimental conditions and bad abandonment. Star symbol indicates statistically significance ($p < 0.05$).

	A	**C**
C	$\chi^2 = 1.41$, $p = .23$	
E	$\chi^2 = 4.13$, $p = .04^*$	$\chi^2 = 0.83$, $p = .36$

not find any statistically significant difference between the conditions. While the results in Figure 6.7 are not statistically significant, we see a trend that reflects what we hypothesized.

Overall, this experiment shows that human effort and user behavior are adaptive to the search results' quality, and that examination behavior changes depending on whether the search results are encouraging or discouraging towards the information need. Indeed, when we asked one of the participants on why they repeatedly requested to view more results when the search results are on diverse multiple topics, but never when the results are on a coherent egregious topic, the participant mentioned "*because I could tell the search engine didn't understand what I was trying to communicate*". Another participant noted that after they are presented with search results, they "*immediately get a sense of if I might be going around the right or wrong direction*". This has important implications on the procedure of collecting relevance judgments and evaluation in information retrieval. Historically, relevance was collected using a binary scale, e.g., a document can be considered either relevant or not relevant, and more recently using a multi-level scale but with non-relevant

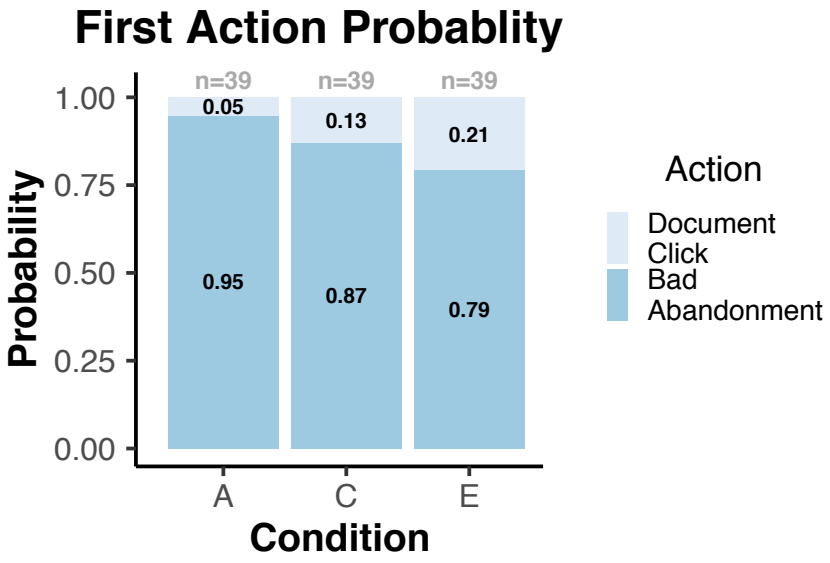


Figure 6.6: The probability of the first action users would make after being presented with the search results.

documents often having a single category. If we would like to better measure and understand user's experience of the search system, instead of focusing mainly on how to label relevant documents, we should also broaden our notion of what it means for a document to not be relevant (i.e., would users find this document encouraging or discouraging?) and include graded-relevance for both relevant and non-relevant documents. Previous research has also shown that assessors assign different scores to non-relevant documents (Figure 1 in Turpin et al. (2015)). Having a better relevance labeling of documents that accounts for user behavior can help us move forward in building metrics that reflect the overall user search experience (e.g., their rate of query abandonment and the number of search results examined on different types of search results).

The results of the experiment also raise the question of which type of SERP should a search engine aim to return. Ideally, a search engine should provide its users their information need at the lowest cost in terms of user effort. This requires a good understanding of users' queries and intentions. In cases where the search engine has limited knowledge of what the user is trying to achieve with their query, our results in Figure 6.7 indicate that it may be best to return a set of search results coherent on a single topic if diversifying the search results fails to include a relevant document. This forces the users to reformulate their query while saving themselves the additional cost of further examining the search

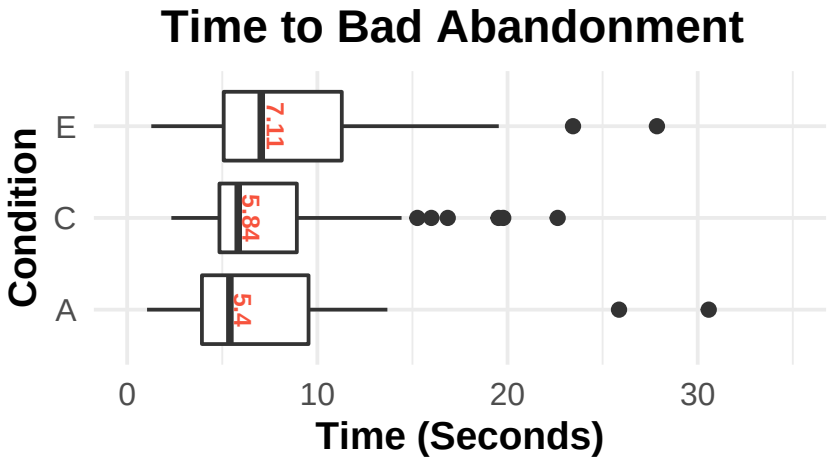


Figure 6.7: Time to bad query abandonment under each condition.

results, or mistakenly clicking on a wrong document. As one participant noted, "*If I did a search and I wasn't getting the results I wanted right away, I would reword the query*"

6.6 Limitation

A limitation of this work is that we only investigated search behavior on questions that require no prior knowledge, and are most likely to be familiar to participants. We purposely selected these questions as we wanted to capture search interaction without introducing compounding variables. More complex questions, such as those that require more understanding or memory recall are shown to require different search behavior (Moffat et al., 2017, 2013).

6.7 Conclusion

Many previous web search user studies have focused on understanding how relevant documents influence examination. In this work, we investigated how the nature of non-relevant documents in a SERP can influence users' interaction with search results. In a web search user study, we carefully controlled the search results shown in the SERP, based on a spectrum of SERP quality that we defined (Figure 6.1), and asked participants to use our

search engine to complete question-based search tasks. When participants interacted with our search engine, we showed them controlled SERPs that only contained non-relevant documents, but differed in the coherence and type of non-relevant documents included in the SERP. The controlled SERPs contained either 1) results coherent on a single off topic, 2) results on multiple off-topics and one related to the topic of the search task fails to have any relevant information to the search task, or 3) diversified results on multiple related subtopics, but also do not have any relevant information to the search task. While all of the search results in the controlled SERP are considered non-relevant, we found that users are likely to examine more results when presented with diversified search results on multiple subtopics. As the SERP contains more off-topic non-relevant results, users are less likely to examine more than the first three search results, and spend less time before abandoning the results and reformulating their query.

The results of our experiment illustrate that people change their behavior based on the nature of non-relevant documents in search results. The results are important to both search engine designers and the design of effectiveness measures for the accurate evaluation of search quality. Our findings suggest that in order to better reflect the overall user search experience, we need to rethink the importance of non-relevant documents in our current research on evaluation and relevance assessment, and particularly, extend graded relevance to non-relevant documents as well.

Chapter 7

Conclusion and Future Work

Through the research presented in this dissertation, we studied bad query abandonment in web search in more detail. We conducted three user studies to investigate how people behave under different types of queries and how that affects users decision to abandon their queries. In this final chapter, we summarise our work and conclude by outlining additional directions for future research.

7.1 Summary

Our first work in studying query abandonment was completed by conducting a lab-based user study. In particular, the study, described in Chapter 3, investigated the effect of SERPs of different quality on the rate and time to abandon search results. Our focus in this user study is to determine the rate of query abandonment under different SERP qualities and the time it takes users to make their decision. In our study, we asked participants to find answers to a set of questions. Each participant completed a total of 12 search tasks, with each task including 1 factoid question. We manipulated the search results for 11 out of the 12 tasks. In those tasks, whenever the participant enters a query with relevant terms, our hand-crafted manipulated SERP will be shown to the user. In 10 search tasks, we included 1 relevant document placed at rank 1 to 10, and in 1 task, all the results are not relevant. Our experiment led to the following results: users make their decisions to abandon or click quickly, and the median time from query to abandonment was 7.7 seconds. The probability of an abandonment increases as the user has to search further down the ranked list to find a relevant document. In particular, the probability of a query abandonment approximately doubles when the topmost relevant document is at rank 2

rather than at rank 1. The time it takes users to make a decision of whether to abandon or not appears to be independent of search results quality. After further analysis of the data, there appears to be two classes of users that behave differently, and possibly corresponds to what the IR literature describes as economic and exhaustive users. One class of users seems to be focusing on higher ranked items on whether or not to abandon the search results. The other group seems to be more exhaustive in the their examination behavior (e.g., they appear more likely to examine the whole ranked list). The group of users who abandon quicker are able to find the answer and complete the search task faster than the other group.

Our first experiment left us with some questions, in particular, what search results people examine before making their decision to abandon, and how far down the ranked list people examine. We conducted another lab-based user study that included eye-tracking to study query abandonment in more details. In this study, described in Chapter 4, we focus on understanding how far in the search result list people examine before making their decision to abandon their queries, and what factors might influence their decision. The study was conducted in a private lab with eye-tracking setup. The design of the experiment was similar to our first work, but with the addition of using both desktop and mobile devices. After conducting the experiment and analyzing the data, we found that the first three search results are important. If a user issues a query unlikely to produce good results, which we denote as weak, the user is more likely to abandon after finding the top three results to be non-relevant than if the user had issued a stronger query that is expected to produce good results. If a relevant document is in the first three search results, the user are highly likely to click on it. Not all users are the same, however. If the user is an exhaustive user, they are less influenced by the quality of their queries (i.e. strong v.s. weak) and are more persistent than economic users. Rank has much less effect on their likelihood of viewing a relevant result than it does for economic users. Economic users, are unlikely to scroll and view search results off of the page, and thus are more likely to abandon the query when the topmost relevant result is below the page fold. In addition to these findings, we also show that for mobile search, users are likely to scroll to view the first five results, but if a relevant result is not seen, they will then abandon their query. We also show a decision tree model that uses the factors of rank, user type, and query quality and demonstrates the importance of these factors to understanding a user's decision to click or requery as their first action. The decision tree provides a holistic view of users' interactions with search engines.

We use eye-tracking data collected from our previous experiment to visualize searchers' gazing pattern in web search. We show how time-series heat maps can be useful in understanding gaze patterns of searchers, communicating how far down the ranking people

examine and how quickly they examine the results. These visualization are suitable for the typical "10 blue links" where search results are ordered linearly from the top to the bottom of the page. Examples of the visualization on real users completing their searches are included in Chapter 5.

Our two previous experiments primarily focused on query abandonment when there is at most one relevant document placed at different ranks. Instead of focusing on the presence or absence of relevant results, in our next work described in Chapter 6, we investigated how the nature of non-relevant documents can affect query abandonment and search result examination. We examined how different types of non-relevant documents and different types of SERPs can affect user behavior. We conducted a mobile-based user study with 12 search tasks. In six of these tasks, the search engine results were manipulated to show results of three types of SERP qualities. The three SERPs qualities differ in their coherence and the type of non-relevant results included in the SERP (e.g., off-topic non-relevant or related-subtopic non-relevant). The other six tasks had search results returned from the Bing API. After analyzing the data, we found that users' interactions are influenced differently by the type and quality of the SERP presented to them. While every manipulated SERP contained only non-relevant documents, when users were shown erroneous non-relevant results, the probability of users requesting to view more results at least once is a low 0.28. The probability jumped to 0.41 when we included one subtopic-related result among other erroneous non-relevant results, and further increased to 0.56 when users were shown a SERP containing within-topic non-relevant search results. The time it takes users to abandon the SERP is different depending on the quality of the SERP presented to them. Users spent a median of 5.4 seconds when the SERP was the lowest quality in our spectrum, i.e., when it only contained erroneous search results. The time increased as the quality of the SERP improved. When users were shown SERPs containing multiple within-topic non-relevant search results, users took about 7 seconds before abandoning the results, a difference of 1.6 seconds from our lowest quality SERP. When users were presented with a SERP containing search results coherent on a single erroneous topic, users would abandon the search result with a high probability (0.95). The probability decreased to 0.87 when the SERP contained a lesser amount of erroneous non-relevant results, and down to 0.79 when the SERP had no erroneous results and only contained subtopic related non-relevant search results. While all the SERP results contain no relevant information to the search task, this result indicates that users may incorrectly click on non-relevant documents when the results seem encouraging. Our findings suggest that in order to better reflect the overall user search experience, we need to judge non-relevant documents using multiple levels, similar to the typical levels used when judging relevant documents.

7.2 Future Work

Below, we briefly describe a number of potential avenues for future work.

Beyond First Query

In this dissertation, we focused on a user's action once they are presented with the SERP of their first query for the search task. Clearly, the search process can expand multiple queries before a user's information need is satisfied.

Our first consideration for future work revolves around query abandonment beyond users' first submitted query. We can imagine different scenarios where the nature of the search task and users' perceived difficulty of the search task may play a role in users decision of abandonment. For example, as a user clicks and processes search results, other search results with repeated information may be of less interest to a user who is trying to accumulate new knowledge. Subsequent queries that contain the same previously viewed search results, or search results with already known information, might prompt the user to abandon and reformulate their query, hoping for better results with new information.

Users' perceived difficulty of the search task may possibly influence users' decision to query abandonment. As the user starts their search process, they may start with a perceived level of difficulty in finding the information that satisfies their information need. If the SERPs of their first few queries do not contribute anything towards their information need, the questions then arises whether there is an increase in perceived difficult of the search task, and therefore increasing efforts by examining more search results by means of desperation.

Understanding how the rate of query abandonment changes at different stages of the search process is an interesting area for future work.

Query Quality and User Expectation

Our work in Chapter 4 indicates users examine fewer items when their query is somewhat ambiguous to their search topic. This raises interesting questions in how people formulate their queries, their perceived quality of query and what their expectation is of the quality the search results.

Incorporating Abandonment in Search Engines Evaluation

The findings from our work provides motivation for future work that takes into consideration how people abandon search results, and how the type of query affects examination. In IR, evaluation is typically done by aggregating scores for different queries representing different topics. However, as represented in our work, people enter different types of queries for the same topic (or information need), and some of these queries may or may not be considered ambiguous. To evaluate search engines and better reflect how people use them, we should also consider different types of queries.

References

Abualsaud, M. and M. D. Smucker (2019). Patterns of search result examination: Query to first action. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, New York, NY, USA, pp. 1833–1842. Association for Computing Machinery.

Ai, Q., S. T. Dumais, N. Craswell, and D. Liebling (2017). Characterizing email search using large-scale behavioral logs and surveys. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, Republic and Canton of Geneva, CHE, pp. 1511–1520. International World Wide Web Conferences Steering Committee.

Arguello, J. and R. Capra (2012). The effect of aggregated search coherence on search behavior. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, New York, NY, USA, pp. 1293–1302. Association for Computing Machinery.

Aula, A., P. Majaranta, and K.-J. Räihä (2005). Eye-tracking reveals the personal styles for search result evaluation. In *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction*, INTERACT'05, Berlin, Heidelberg, pp. 1058–1061. Springer-Verlag.

Aula, A., P. Majaranta, and K.-J. Räihä (2005). Eye-tracking reveals the personal styles for search result evaluation. In M. F. Costabile and F. Paternò (Eds.), *Human-Computer Interaction - INTERACT 2005*, Berlin, Heidelberg, pp. 1058–1061. Springer Berlin Heidelberg.

Azzopardi, L. (2011). The economics in interactive information retrieval. In W. Ma, J. Nie, R. Baeza-Yates, T. Chua, and W. B. Croft (Eds.), *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pp. 15–24. ACM.

Azzopardi, L., R. W. White, P. Thomas, and N. Craswell (2020). Data-driven evaluation metrics for heterogeneous search engine result pages. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 213–222.

Balatsoukas, P. and I. Ruthven (2012). An eye-tracking approach to the analysis of relevance judgments on the web: The case of google search engine. *Journal of the American Society for Information Science and Technology 63*(9), 1728–1746.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles 67*(1), 1–48.

Bhattacharya, N., S. Rakshit, J. Gwizdka, and P. Kogut (2020). Relevance prediction from eye-movements using semi-interpretable convolutional neural networks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, New York, NY, USA, pp. 223–233. Association for Computing Machinery.

Blascheck, T., K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl (2017). Visualization of eye tracking data: A taxonomy and survey. *Computer Graphics Forum 36*(8), 260–284.

Bota, H., K. Zhou, and J. M. Jose (2016). Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, New York, NY, USA, pp. 131–140. Association for Computing Machinery.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth.

Broder, A. (2002, September). A taxonomy of web search. *SIGIR Forum 36*(2), 3–10.

Brückner, L., I. Arapakis, and L. A. Leiva (2020). Query abandonment prediction with recurrent neural models of mouse cursor movements. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, New York, NY, USA, pp. 1969–1972. Association for Computing Machinery.

Buscher, G., E. Cutrell, and M. R. Morris (2009). What do you see when you're surfing? using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, New York, NY, USA, pp. 21–30. Association for Computing Machinery.

Buscher, G., R. W. White, S. Dumais, and J. Huang (2012). Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, New York, NY, USA, pp. 373–382. Association for Computing Machinery.

Clark, M., I. Ruthven, P. O. Holt, and D. Song (2012). Looking for genre: The use of structural features during search tasks with wikipedia. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, New York, NY, USA, pp. 145–154. Association for Computing Machinery.

Cutrell, E. and Z. 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*, CHI '07, New York, NY, USA, pp. 407–416. ACM.

Das Sarma, A., S. Gollapudi, and S. Ieong (2008). Bypass rates: Reducing query abandonment using negative inferences. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, New York, NY, USA, pp. 177–185. Association for Computing Machinery.

Diriye, A., R. White, G. Buscher, and S. Dumais (2012). Leaving so soon?: Understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, New York, NY, USA, pp. 1025–1034. ACM.

Dumais, S. T., G. Buscher, and E. Cutrell (2010). Individual differences in gaze patterns for web search. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, New York, NY, USA, pp. 185–194. ACM.

Eickhoff, C., S. Dungs, and V. Tran (2015). An eye-tracking study of query reformulation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, New York, NY, USA, pp. 13–22. ACM.

Goldberg, J. H., M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky (2002). Eye tracking in web search tasks: Design implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research Applications*, pp. 51–58.

Granka, L., M. Feusner, and L. Lorigo (2008). *Eye Monitoring in Online Search*, pp. 347–372. Berlin, Heidelberg: Springer Berlin Heidelberg.

Granka, L. A., T. Joachims, and G. 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*, SIGIR '04, New York, NY, USA, pp. 478–479. ACM.

Guan, Z. and E. Cutrell (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, New York, NY, USA, pp. 417–420. ACM.

Hearst, M. A. (2009). *Search User Interfaces* (1st ed.). New York, NY, USA: Cambridge University Press.

Hofmann, K., B. Mitra, F. Radlinski, and M. Shokouhi (2014). An eye-tracking study of user interactions with query auto completion. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, New York, NY, USA, pp. 549–558. Association for Computing Machinery.

Jansen, B. J. and U. Pooch (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology 52*(3), 235–246.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, and G. Gay (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, New York, NY, USA, pp. 154–161. ACM.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay (2007, April). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst. 25*(2), 7–es.

Joachims, T. and F. Radlinski (2007, August). Search engines that learn from implicit feedback. *Computer 40*(8), 34–40.

Khabsa, M., A. Crook, A. H. Awadallah, I. Zitouni, T. Anastasakos, and K. Williams (2016). Learning to account for good abandonment in search success metrics. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, New York, NY, USA, pp. 1893–1896. Association for Computing Machinery.

Kim, J., P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon (2017). What snippet size is needed in mobile web search? In *Proceedings of the 2017 Conference on*

Conference Human Information Interaction and Retrieval, CHIIR '17, New York, NY, USA, pp. 97–106. Association for Computing Machinery.

Klaus, K. (2004). Content analysis: An introduction to its methodology.

Klöckner, K., N. Wirschum, and A. Jameson (2004a). Depth- and breadth-first processing of search result lists. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, New York, NY, USA, pp. 1539–1539. ACM.

Klöckner, K., N. Wirschum, and A. Jameson (2004b). Depth- and breadth-first processing of search result lists. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, New York, NY, USA, pp. 1539. Association for Computing Machinery.

Li, J., S. Huffman, and A. Tokuda (2009a). Good abandonment in mobile and pc internet search. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, New York, NY, USA, pp. 43–50. ACM.

Li, J., S. Huffman, and A. Tokuda (2009b). Good abandonment in mobile and pc internet search. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, New York, NY, USA, pp. 43–50. ACM.

Liu, Y., C. Wang, K. Zhou, J. Nie, M. Zhang, and S. 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, New York, NY, USA, pp. 849–858. ACM.

Liu, Z., Y. Liu, K. Zhou, M. Zhang, and S. 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, New York, NY, USA, pp. 193–202. Association for Computing Machinery.

Lorigo, L., M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology 59*(7), 1041–1052.

MacFarlane, A., G. Buchanan, A. Al-Wabil, G. Andrienko, and N. Andrienko (2017). Visual analysis of dyslexia on search. In *Proceedings of the 2017 Conference on Conference*

Human Information Interaction and Retrieval, CHIIR '17, New York, NY, USA, pp. 285–288. Association for Computing Machinery.

Maxwell, D. and L. Azzopardi (2018). Information scent, searching and stopping. In G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury (Eds.), *Advances in Information Retrieval*, Cham, pp. 210–222. Springer International Publishing.

Moffat, A., P. Bailey, F. Scholer, and P. Thomas (2017, June). Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Syst. 35*(3).

Moffat, A., P. Thomas, and F. Scholer (2013). Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM International Conference on Information Knowledge Management*, CIKM '13, New York, NY, USA, pp. 659–668. Association for Computing Machinery.

Moffat, A. and A. F. Wicaksono (2018). Users, adaptivity, and bad abandonment. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '18, New York, NY, USA, pp. 897–900. Association for Computing Machinery.

Munzner, T. (2014). Visualization analysis and design. In *A.K. Peters visualization series*.

Navalpakkam, V., L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, New York, NY, USA, pp. 953–964. Association for Computing Machinery.

Ong, K., K. Järvelin, M. Sanderson, and F. Scholer (2017). Using information scent to understand mobile and desktop web search behavior. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, New York, NY, USA, pp. 295–304. ACM.

Palani, S., A. Fourney, S. Williams, K. Larson, I. Spiridonova, and M. R. Morris (2020). An eye tracking study of web search by people with and without dyslexia. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, New York, NY, USA, pp. 729–738. Association for Computing Machinery.

Papoutsaki, A., J. Laskey, and J. Huang (2017). Searchgazer: Webcam eye tracking for remote studies of web search. In *Proceedings of the 2017 Conference on Conference*

Human Information Interaction and Retrieval, CHIIR '17, New York, NY, USA, pp. 17–26. Association for Computing Machinery.

Pirolli, P. and S. Card (1999). Information foraging. *Psychological review 106*(4), 643.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radlinski, F., M. Kurup, and T. Joachims (2008). How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, New York, NY, USA, pp. 43–52. ACM.

Räihä, K.-J., A. Aula, P. Majaranta, H. Rantala, and K. Koivunen (2005). Static visualization of temporal eye-tracking data. In *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction*, INTERACT'05, Berlin, Heidelberg, pp. 946–949. Springer-Verlag.

Raschke, M., X. Chen, and T. Ertl (2012). Parallel scan-path visualization. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, New York, NY, USA, pp. 165–168. ACM.

Silverstein, C., H. Marais, M. Henzinger, and M. Moricz (1999, September). Analysis of a very large web search engine query log. *SIGIR Forum 33*(1), 6–12.

Song, Y., X. Shi, R. White, and A. H. Awadallah (2014). Context-aware web search abandonment prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '14, New York, NY, USA, pp. 93–102. Association for Computing Machinery.

Spina, D., M. Maistro, Y. Ren, S. Sadeghi, W. Wong, T. Baldwin, L. Cavedon, A. Moffat, M. Sanderson, F. Scholer, and J. Zobel (2017). Understanding user behavior in job and talent search: An initial investigation. In J. Degenhardt, S. Kallumadi, M. de Rijke, L. Si, A. Trotman, and Y. Xu (Eds.), *Proceedings of the SIGIR 2017 Workshop On eCommerce co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, eCOM@SIGIR 2017, Tokyo, Japan, August 11, 2017*, Volume 2311 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Stamou, S. and E. N. Efthimiadis (2009). Queries without clicks: Successful or failed searches. In *SIGIR 2009 Workshop on the Future of IR Evaluation*, pp. 13–14.

Stamou, S. and E. N. Efthimiadis (2010). Interpreting user inactivity on search results. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. van Rijsbergen (Eds.), *Advances in Information Retrieval*, Berlin, Heidelberg, pp. 100–113. Springer Berlin Heidelberg.

Therneau, T. and E. Atkinson (1997, 01). An introduction to recursive partitioning using the rpart routines. *Mayo Clinic 61*.

Turpin, A., F. Scholer, S. Mizzaro, and E. Maddalena (2015). The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, New York, NY, USA, pp. 565–574. Association for Computing Machinery.

Wang, C., Y. Liu, and S. Ma (2016). Building a click model: From idea to practice. *CAAI Transactions on Intelligence Technology 1*(4), 313 – 322.

Wang, Y., D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang, and Q. Mei (2016). Beyond ranking: Optimizing whole-page presentation. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, New York, NY, USA, pp. 103–112. Association for Computing Machinery.

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

White, R. W. and D. Morris (2007). Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, New York, NY, USA, pp. 255–262. Association for Computing Machinery.

Williams, K., J. Kiseleva, A. C. Crook, I. Zitouni, A. H. Awadallah, and M. Khabsa (2016). Is this your final answer? evaluating the effect of answers on good abandonment in mobile search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, New York, NY, USA, pp. 889–892. Association for Computing Machinery.

Wu, W.-C. and D. Kelly (2014). Online search stopping behaviors: An investigation of query abandonment and task stopping. *Proceedings of the American Society for Information Science and Technology 51*(1), 1–10.

Wu, W.-C., D. Kelly, and A. Sud (2014). Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, New York, NY, USA, pp. 557–566. ACM.

Zhang, H., M. Abualsaud, and M. D. Smucker (2018). A study of immediate requery behavior in search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, New York, NY, USA, pp. 181–190. ACM.

APPENDICES

Appendix A

Datasets

Data collected as part of the user experiments are available to researchers on request. Please see https://github.com/ammsa/query-abandonment-data for more details.

Appendix B

Search Results Quality and Query Abandonment

ORE OFFICE USE ONLY

ORE #_____

APPLICATION FOR ETHICS REVIEW OF RESEARCH INVOLVING HUMAN PARTICIPANTS

Please remember to PRINT AND SIGN the form and forward with all attachments to the Office of Research Ethics, EC5, 3rd floor.

A. GENERAL INFORMATION

1. Title of Project: Question Answering Performance of Search Engines

2. a) Principal and Co-Investigator(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:
Mark D. Smucker (faculty)	Management Sciences	38620	mark.smucker@uwaterloo.ca

2. b) Collaborator(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:

3. Faculty Supervisor(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:

4. Student Investigator(s)

Name	Department	Ext:	e-mail:	Local Phone #:
Haotian Zhang	Computer Science, School of		h435zhan@uwaterloo.ca	5197226812
Mustafa Abualsaud	Computer Science, School of		m2abuals@uwaterloo.ca	

5. Level of Project: Faculty Research **Specify Course:**

Research Project/Course Status: New Project\Course

6. Funding Status (If Industry funded and a clinical trial involving a drug or natural product or is medical device testing, then Appendix B is to be completed):

Is this project currently funded? Yes

- If Yes, provide Name of Sponsor and include the title of the grant/contract: NSERC : CRD, User Behavior Models for Information Access Evaluation, 2400-119733

- If No, is funding being sought OR if Yes, is additional funding being sought? No
- Period of Funding: September 2015 to August 2018

7. Does this research involve another institution or site? No
If Yes, what other institutions or sites are involved:

8. Has this proposal, or a version of it, been submitted to any other Research Ethics Board/Institutional Review Board? No

9. For Undergraduate and Graduate Research:

Has this proposal received approval of a Department Committee? Not Dept. Req.

10. a) Indicate the anticipated commencement date for this project: 12/1/2016

b) Indicate the anticipated completion date for this project: 12/1/2017

11. Conflict of interest: Appendix B is attached to the application if there are any potential, perceived, or actual financial or non-financial conflicts of interest by members of the research team in undertaking the proposed research.

B. SUMMARY OF PROPOSED RESEARCH

1. Purpose and Rationale for Proposed Research

a. Describe the purpose (objectives) and rationale of the proposed project and include any hypothesis(es)/research questions to be investigated. For a non-clinical study summarize the proposed research using the headings: Purpose, Aim or Hypothesis, and Justification for the Study. For a clinical trial/medical device testing summarize the research proposal using the following headings: Purpose, Hypothesis, Justification, and Objectives.
Where available, provide a copy of a research proposal. For a clinical trial/medical device testing a research proposal is required:
Purpose: To collect information on how people use web search for the purpose of answering fact-based questions.

Hypothesis: We hypothesize that as the quality of the search results decreases, people will be more likely to issue a new search query and the longer users will take to find a good quality page that helps them answer the question.

Justification for the Study: While the IR field offers good analysis on how user search for documents, it is lacking information on how much users, given a set of documents, are motivated to find answers, and when do they decide to give up on the list of documents they're provided with and decide to generate a new list.

Objectives: This project will collect data allowing us to estimate a conditional probability of user changing search query and searching for new results given the search result lists with varying ranking quality. With this data, we can construct a model of human performance and compare its predictions to actual human performance also measured as part of this project.

b. In lay language, provide a one paragraph (approximately 100 words) summary of the project including purpose, the anticipated potential benefits, and basic procedures used.
In this study, we will ask participants to use our specifically designed web-search engine for answering various fact-based questions. Participants will be allowed to enter their own queries

and look for web documents that they feel will help them find answers. After collecting the data, we will analyze it to see how and when do users decide to give up on a list of documents, and generate a new list.

C. DETAILS OF STUDY

1. Methodology/Procedures

a. Indicate all of the procedures that will be used. Append to form 101 a copy of all materials to be used in this study.

Computer-administered task(s) or survey(s) None are standardized.
Unobtrusive observations
Logging of computer usage.

b. Provide a detailed, sequential description of the procedures to be used in this study. For studies involving multiple procedures or sessions, provide a flow chart. Where applicable, this section also should give the research design (e.g., cross-over design, repeated measures design).
This study will have one phase with 10 different tasks. This protocol uses with

slight modifications the protocol given by Toms et al. "WiIRE: the Web interactive information retrieval experimentation system prototype," Information Processing and Management, 40, 2004, pp. 655-675.

Protocol

1. Introduction

2. Consent Form

3. Demographic and Background Questionnaire

4. Overview of Experiment

5. Tutorial

6. Practice Interface

7. Pre Task Questionnaire (once for each task)

8. Task (once for each task)

9. Post Task Questionnaire (once for each task)

10. Thank You

Phase will involve the participants being presented with web pages formatted similar to popular web search engines such as Google, Yahoo, or Microsoft's Bing. Participants will be asked to use this interface to view document summaries, view the underlying documents and change search queries. Participants will be asked to search for correct answers to our pre-defined questions. We will collect timing information and associated computer usage data unobtrusively during both phases of the study.

c. Will this study involve the administration/use of any drug, medical device, biologic, or natural health product? No

d. Will you be using, processing and/or storing any biological materials of human origin such as blood, tissue,

cells or bodily fluids?
No

2. Participants Involved in the Study

a. Indicate who will be recruited as potential participants in this study.
UW Participants:
Undergraduate students
Graduate students
Faculty and/or Staff

b. Describe the potential participants in this study including group affiliation, gender, age range and any other special characteristics. Describe distinct or common characteristics of the potential participants or a group (e.g., a group with a particular health condition) that are relevant to recruitment and/or procedures. Provide justification for exclusion based on culture, language, gender, race, ethnicity, age or disability. For example, if a gender or sub-group (i.e., pregnant and/or breastfeeding women) is to be excluded, provide a justification for the exclusion.

Adults fluent in English, familiar with web search (e.g. Google, Yahoo, Bing), and capable of unassisted use of a computer with keyboard, mouse, and LCD monitor.

c. How many participants are expected to be involved in this study? For a clinical trial, medical device testing, or study with procedures that pose greater than minimal risk, sample size determination information is to be provided.

48 to 60 plus a couple of participants during the pilot phase. The study will involve 12 questions. We know that human performance in text retrieval varies across both humans and the search topics. We use a 12x12 block design (12 participants and 12 topics forming a 12x12 Latin square), and thus will have 48 or 60 participants depending on if we have 4 or 5 blocks. This will be a convenience sample of students and other adults of the University of Waterloo community.

3. Recruitment Process and Study Location

a. From what source(s) will the potential participants be recruited?
Other UW sources: Posters across campus.

b. Describe how and by whom the potential participants will be recruited. Provide a copy of any materials to be used for recruitment (e.g. posters(s), flyers, cards, advertisement(s), letter(s), telephone, email, and other verbal scripts).
We will post posters around the University of Waterloo campus.

c. Where will the study take place? On campus: On campus: CPH 4335

4. Remuneration for Participants
Will participants receive remuneration (financial, in-kind, or otherwise) for participation? Yes
If Yes, provide details:
We will pay all participants $15 with an advertised payment of $10 for participation and a $5 bonus for getting at least 10 out of 12 questions answered correctly. Regardless of participant performance, we will pay all participants the full $15. This payment structure is designed to motivate good performance while not harming any person who may be unable to achieve the actual desired performance. Should participants need to leave before completing all tasks, they will be paid on a pro-rated basis for the number of tasks completed rounded to the nearest $5. It is expected that the participants will need to spend around 1 hour to complete the study.

5. Feedback to Participants

Describe the plans for provision of study feedback and attach a copy of the feedback letter to be used. Wherever possible, written feedback should be provided to study participants including a statement of appreciation, details about the purpose and predictions of the study, restatement of the provisions for

confidentiality and security of data, an indication of when a study report will be available and how to obtain a copy, contact information for the researchers, and the ethics review and clearance statement.
Participants will be advised that if they are interested in the outcomes of the study, they may contact the principal investigator at a later time to learn about any resulting publications.

D. POTENTIAL BENEFITS FROM THE STUDY

1. Identify and describe any known or anticipated direct benefits to the participants from their involvement in the project.
There are no known direct benefits to the participants from their involvement in the project.

2.Identify and describe any known or anticipated benefits to the scientific community/society from the conduct of this study.
Information retrieval (text search) has become part of daily life for many Canadians, as well as people around the world. This study has the long term potential to allow researchers to better evaluate retrieval systems. With better evaluation tools that allow for faster and more accurate evaluations, the rate at which retrieval systems improve should increase. With better retrieval systems, people are able to find information previously hidden and the better relevant information sorted, the better decisions they are able to make.

E. POTENTIAL RISKS TO PARTICIPANTS FROM THE STUDY

1. For each procedure used in this study, describe any known or anticipated risks/stressors to the participants. Consider physiological, psychological, emotional, social, economic risks/stressors. A study–specific current health status form must be included when physiological assessments are used and the associated risk(s) to participants is minimal or greater.

Minimal risks anticipated.
Participants will be asked to use a computer with keyboard, mouse, and LCD monitor to answer brief questionnaires as well as to read and answer questions according to given result lists. These activities are common to everyday life and pose no greater risk. The search questions that will be utilized are those that might be posed by a lay person and all of them deal with matters people can encounter in normal life. All documents come from popular search engines.

2. Describe the procedures or safeguards in place to protect the physical and psychological health of the participants in light of the risks/stressors identified in E1.
As the study involves only minimal risk, no explicit procedures or safeguards will be in place other than to provide a safe, usable computer system in a university computing lab commonly used by students.

F. INFORMED CONSENT PROCESS

1. What process will be used to inform the potential participants about the study details and to obtain their consent for participation?
Information letter with written consent form

2. If written consent cannot be obtained from the potential participants, provide a justification for this.

3. Does this study involve persons who cannot give their own consent (e.g. minors)? No

G. ANONYMITY OF PARTICIPANTS AND CONFIDENTIALITY OF DATA

1. Provide a detailed explanation of the procedures to be used to ensure anonymity of participants and confidentiality of data both during the research and in the release of the findings.
All participants will be issued an anonymous identifier (ID). The mapping from a participant's name to the ID will be maintained for the length of the study in case the participant forgets the

ID. This mapping will be kept in a locked cabinet in a secure location during the study and will be destroyed at the completion of the study. After the study concludes, there will be no way to identify a participant to the data. All computer usage will be with computers in a University of Waterloo computer lab and not with personally identifiable computers, i.e. participants will not use their own computer. All data collected will be retained indefinitely and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to "participant 1" or some other similar anonymous name. Participants' names will never appear in any publication that results from this study.

2. Describe the procedures for securing written records, video/audio tapes, questionnaires and recordings. Identify (i) whether the data collected will be linked with any other dataset and identify the linking dataset and (ii) whether the data will be sent outside of the institution where it is collected or if data will be received from other sites. For the latter, are the data de-identified, anonymized, or anonymous?
Each of the questions that we provide to the participants are defined by us. The search result list for each of of the questions will be from the Bing search engine. The initial result lists are manipulated and fixed due to the control of quality. We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participant's data will remain identifiable as coming from an individual, i.e. "participant 1", "participant 2", etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

3. Indicate how long the data will be securely stored and the method to be used for final disposition of the data.
Paper Records
Confidential shredding after a minimum of 7 year(s).
Electronic Data
Erasing of electronic data after a minimum of 7 year(s).
Location: Principal investigator's office (paper) and on secure computers.

4. Are there conditions under which anonymity of participants or confidentiality of data cannot be guaranteed?
No

H. PARTIAL DISCLOSURE AND DECEPTION

1. Will this study involve the use of partial disclosure or deception? Partial disclosure involves withholding or omitting information about the specific purpose or objectives of the research study or other aspects of the research. Deception occurs when an investigator gives false information or intentionally misleads participants about one or more aspects of the research study. Yes

If Yes,
(i) explicitly state if it is partial disclosure and/or deception,
(ii) if applicable, describe the partial disclosure, that is what information is being withheld or omitted concerning the purpose/objectives or procedures,
(iii) if applicable, describe all of the deception(s) to be used in this study, AND
(iv) provide a justification for each use of partial disclosure and deception.

We will use a minor form of deception. We will tell participants that there is a bonus payment of $5 for answering 10 or more of 12 questions correctly. In all cases, we will award the $5 to the participants regardless of their performance. We are using this form of deception to motivate participants to perform at a high level of quality, for otherwise we fear that participants may simply answer false answers to quickly finish the study.

If Yes, **outline the process to be used to debrief participants**.
Provide a hard copy of the written debriefing sheet for participants, the researcher's verbal debriefing script (if applicable), and for deception, the materials used to obtain consent following debriefing

As part of our thank you / debriefing letter, we will explain that they were awarded the bonus regardless of their performance.

Researchers must ensure that all supporting materials/documentation for their applications are submitted with the signed, hard copies of the ORE form 101/101A. Note, materials shown below in bold are normally required as part of the ORE application package. The inclusion of other materials depends on the specific type of projects.

Protocol Involves a Drug, Medical Device, Biologic, or Natural Health Product

If the study procedures include administering or using a drug, medical device, biologic, or natural health product that has been or has not been approved for marketing in Canada then the researcher is to complete Appendix A. Appendix A is to be attached to each of the one copy of the application that are submitted to the ORE. Information concerning studies involving a drug, biologic, natural health product, or medical devices can be found on the ORE website.

Please **check** below all appendices that are attached as part of your application package:

- Recruitment Materials: A copy of any poster(s), flyer(s), advertisement(s), letter(s), telephone or other verbal script(s) used to recruit/gain access to participants.
- Information Letter and Consent Form(s)*. Used in studies involving interaction with participants (e.g. interviews, testing, etc.)
- Data Collection Materials: A copy of all survey(s), questionnaire(s), interview questions, interview themes/sample questions for open-ended interviews, focus group questions, or any standardized tests.
- Feedback letter *

* Refer to sample letters.

NOTE: The submission of incomplete application packages will increase the duration of the ethics review process.

To avoid common errors/omissions, and to minimize the potential for required revisions, applicants should ensure that their application and attachments are consistent with the *Checklist For Ethics Review of Human Research Application*

Please note the submission of incomplete packages may result in delays in receiving full ethics clearance. We suggest reviewing your application with the Checklist For Ethics Review of Human Research Applications to minimize any required revisions and avoid common errors/omissions.

INVESTIGATORS' AGREEMENT

I have read the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, 2nd Edition (TCPS2) and agree to comply with the principles and articles outlined in the TCPS2. In the case of student research, as Faculty Supervisor, my signature indicates that I have read and approved this application and the thesis proposal, deem the project to be valid and worthwhile, and agree to provide the necessary supervision of the student.

NEW As of May 1, 2013, all UW faculty and staff listed as investigators must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. Each investigator is to indicate they have completed the TCPS2 tutorial. If there are more than two investigators, please attach a page with the names of each additional investigator along with their TCPS2 tutorial completion information.

_____ _____
Print and Signature of Principal Investigator/Supervisor **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

_____ _____
Print and Signature of Principal Investigator/Supervisor **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

Each student investigator is to indicate if they have completed the Tri-Council Policy Statement, 2nd Edition Tutorial (http://pre.ethics.gc.ca/eng/education/tutorial-didacticiel/). If there are more than two student investigators, please attach a page with the names of each additional student investigator along with their TCPS2 tutorial completion information.

_____ _____
Signature of Student Investigator **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

_____ _____
Signature of Student Investigator **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

FOR OFFICE OF RESEARCH ETHICS USE ONLY:

_____ _____
Jannet Ann Leggett, JD **Date**
Chief Ethics Officer
OR
Julie Joza, MPH
Senior Manager, Research Ethics
OR
Sacha Geer, PhD
Manager, Research Ethics
OR
Nick Caric, MDiv
Research Ethics Advisor

ORE 101
Revised September 2016

Title of Project: Question Answering Performance of Search Engines

Principal Investigator

Mark D. Smucker (msmucker@uwaterloo.ca)

Student Investigators

- Haotian Zhang, +1-519-722-6812, h435zhan@uwaterloo.ca
- Mustafa Abualsaud, m2abuals@uwaterloo.ca ,

Summary of the Project:

In this study, we will ask participants to use our specifically designed web-search engine for answering various fact-based questions. Participants will be allowed to enter their own queries and look for web documents that they feel will help them find answers. We will measure participants' behaviour and performance. With this data, we plan to build better models of human performance.

Procedure:

Your participation in this study is voluntary. Participation involves using a search engine to answer questions. One example question is that what is the height of the CN tower?

You will be asked to complete several brief questionnaires and to search for and save answers towards given search questions for 12 topics using a search engine. The questionnaires that you will be asked to complete consist of a demographic questionnaire and a questionnaire concerning the search topic before each search topic task and a questionnaire about the task after each search topic task.

To participate, you must be a fluent speaker of English and require no assistance with using a computer with a keyboard, mouse, and LCD monitor.

The study will take approximately 1 hour.

We will record both your answers and your interaction with the computer. We may also make note of and record anything we observe, including what you say, while you are participating in the study.

You may decline to answer any question that you prefer not to answer. You may stop participating in the study at any point and withdraw your consent without penalty.

Expectations for your Participation:

Some participants may finish before other participants. Please focus on your own work and continue to work at your own pace. Please work on a given task form start to finish. If you need to take a break, please do so between tasks. Once you have answered a question, do not attempt to go back and change your answer. All answers are final.

This scientific research study requires your full attention. If you are unable to give this research your full attention, please excuse yourself from the study. In particular:

You may not use your mobile phones during the study [Please keep your phones on silent mode]

You may not listen to music during the study

You may not discuss the answers or talk to other participants during the study

You may not use the computer for checking email, viewing web pages, or other activities during the study

If you use the computer for non-study related activities or use a mobile phone, we will end your participation and ask you to leave

Confidentiality and Data Security:

You will be issued an anonymous identifier (ID) as a participant in this study. The mapping from your name to the ID will be maintained for the length of the study in case you forget the ID. This mapping will be kept in a locked cabinet in a secure location during the study and will be destroyed at the completion of the study. After the study concludes, there will be no way to identify you to the data. All computer usage will be with computers in a University of Waterloo computer lab and not with personally identifiable computers, i.e. you will not use your own computer. All data collected will be retained for a minimum of 7 years and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to "participant 1" or some other similar anonymous name. Your name will never appear in any publication that results from this study.

We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participant's data will remain identifiable as coming from an individual, i.e. "participant 1", "participant 2", etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

Remuneration for Your Participation:

You will be paid $10. If you answer at least 10 of the 12 questions correctly, we will pay you a bonus of $5 for a total of $15. Should you stop before completing the study, you will be paid on a pro-rated basis based on the number of questions answered and rounded up to the nearest $5. The amount received is taxable. It is your responsibility to report this amount for income tax purposes. If you participation is ended early due to your actions (e.g. using the computer for non-study related activities), you will be paid on the pro-rated basis detailed above.

Risks and Benefits:

There is minimal risk to you from participation in this study. Computer use and searching for relevant documents are common everyday activities and pose no anticipated risk greater than that encountered in everyday activities. The search topics that will be utilized are those that might be posed by a lay person in regular every day use of a search engine. All documents come from web sites.

There are no direct benefits to you from participation. However, we hope the study will provide results that can lead to advances in the evaluation and development of advanced text retrieval systems that will benefit society at large.

Research Ethics Clearance:

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics at the University of Waterloo. However, the final decision about participation is yours. Should you have comments or concerns resulting from your participation in this study, please contact Dr. Julie Joza in the Office of Research Ethics at 519-888-4567, Ext. 38535 or jajoza@uwaterloo.ca.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#21930). If you have questions for the Committee contact the Chief Ethics Officer, Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

For all other questions please contact the researchers at the email/phone number above.

Thank you for your assistance in this project.

B.1 Tutorial quiz questions

Below are the tutorial quiz questions, with bolded choices indicating correct answers.

1. How many search tasks do you need to complete in this study?
   (a) 1 Question
   (b) 2 Questions
   (c) ...
   (d) **12 Questions**
2. How long can you spend one question?
   (a) 1 minute
   (b) 2 minutes
   (c) ...
   (d) **I can spend as much time as I need to answer the question**.
3. How many times can you type new queries?
   (a) 1 time
   (b) 2 times
   (c) ...
   (d) **I can enter as many queries as I need to answer the question.**
4. Can you use Google, Bing or Yahoo or find answers?
   (a) Yes
   (b) **No**
5. After I make a judgement:
   (a) **I need to proceed and finish the next task. All answers are final.**
   (b) I may use the web browser to go back and change my answer.
6. During the study:

(a) I should give all my full attention to the study.

(b) I may only use the computer for the study and not use it for email, web browsing, or other activities.

(c) I need to turn off my mobile phone and not use it.

(d) **All of above.**

Appendix C

Patterns of Search Result Examination

ORE OFFICE USE ONLY

ORE #_____

APPLICATION FOR ETHICS REVIEW OF RESEARCH INVOLVING HUMAN PARTICIPANTS

Please remember to PRINT AND SIGN the form and forward with all attachments to the Office of Research Ethics, EC5, 3rd floor.

A. GENERAL INFORMATION

1. Title of Project: Question Answering Performance of Search Engines

2. a) Principal and Co-Investigator(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:
Mark D. Smucker (faculty)	Management Sciences	38620	mark.smucker@uwaterloo.ca

2. b) Collaborator(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:

3. Faculty Supervisor(s)

NEW As of May 1, 2013, all UW faculty and staff listed as investigation must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. The tutorial takes at least three hours; it has start and stop features.

Name	Department	Ext:	e-mail:

4. Student Investigator(s)

Name	Department	Ext:	e-mail:	Local Phone #:
Mustafa Abualsaud	Computer Science, School of		m2abuals@uwaterloo.ca	2508849647
Haotian Zhang	Computer Science, School of		h435zhan@uwaterloo.ca	519722812

5. Level of Project: Faculty Research **Specify Course:**

Research Project/Course Status: New Project\Course. (Extending and modfication of existing project)

6. Funding Status (If Industry funded and a clinical trial involving a drug or natural product or is medical device testing, then Appendix B is to be completed):

Is this project currently funded? Yes

- If Yes, provide Name of Sponsor and include the title of the grant/contract: NSERC : NSERC : CRD, User Behavior Models for Information Access Evaluation, 2400-119733

- If No, is funding being sought OR if Yes, is additional funding being sought? No
- Period of Funding: September 2015 to August 2018

7. Does this research involve another institution or site? No
If Yes, what other institutions or sites are involved:

8. Has this proposal, or a version of it, been submitted to any other Research Ethics Board/Institutional Review Board? No (Extending and modfication of existing project)

9. For Undergraduate and Graduate Research:

Has this proposal received approval of a Department Committee? Not Dept. Req.

10. a) Indicate the anticipated commencement date for this project: 1/15/2018

b) Indicate the anticipated completion date for this project: 1/30/2019

11. Conflict of interest: Appendix B is attached to the application if there are any potential, perceived, or actual financial or non-financial conflicts of interest by members of the research team in undertaking the proposed research.

B. SUMMARY OF PROPOSED RESEARCH

1. Purpose and Rationale for Proposed Research

a. Describe the purpose (objectives) and rationale of the proposed project and include any hypothesis(es)/research questions to be investigated. For a non-clinical study summarize the proposed research using the headings: Purpose, Aim or Hypothesis, and Justification for the Study. For a clinical trial/medical device testing summarize the research proposal using the following headings: Purpose, Hypothesis, Justification, and Objectives.
Where available, provide a copy of a research proposal. For a clinical trial/medical device testing a research proposal is required:
Purpose: To collect information on how people use web search for the purpose of answering fact-based questions.

Hypothesis: We hypothesize that as the quality of the search results decreases, people will be more likely to issue a new search query and the longer users will take to find a good quality page that helps them answer the question.

Justification for the Study: While the IR field offers good analysis on how user search for documents, it is lacking information on how much users, given a set of documents, are motivated to find answers, and when do they decide to give up on the list of documents they are
provided with and decide to generate a new list.

Objectives: This project will collect data allowing us to estimate a conditional probability of user changing search query and searching for new results given the search result lists with varying ranking quality. With this data, we can construct a model of human performance and compare its predictions to actual human performance also measured as part of this project.

b. In lay language, provide a one paragraph (approximately 100 words) summary of the project including

purpose, the anticipated potential benefits, and basic procedures used.
In this study, we will ask participants to use our specifically designed web-search engine for answering various fact-based questions. Participants will be allowed to enter their own queries and look for web documents that they feel will help them find answers. After collecting the data, we will analyze it to see how and when do users decide to give up on a list of documents, and generate a new list.

C. DETAILS OF STUDY

1. Methodology/Procedures

a. Indicate all of the procedures that will be used. Append to form 101 a copy of all materials to be used in this study.

Computer-administered task(s) or survey(s) None are standardized.
Audio-recording
Video-recording
Unobtrusive observations
Phone-administered task(s) or survey(s). Logging of phone/computer usage. Logging of eye tracking data. Video/audio recording of phone usage.

b. Provide a detailed, sequential description of the procedures to be used in this study. For studies involving multiple procedures or sessions, provide a flow chart. Where applicable, this section also should give the research design (e.g., cross-over design, repeated measures design).
This study will have one phase with 10 different tasks. This protocol uses with slight modifications the protocol given by Toms et al. "WiIRE: the Web interactive information retrieval experimentation system prototype," Information Processing and Management, 40, 2004, pp. 655-675.
1. Introduction
2. Consent Form
3. Demographic and Background Questionnaire
4. Overview of Experiment
5. Tutorial
6. Practice Interface
7. Pre Task Questionnaire (once for each task)
8. Task (once for each task)
9. Post Task Questionnaire (once for each task)
10. Thank You

Phase will involve the participants being presented with web pages formatted similar to popular web search engines such as Google, Yahoo, or Microsoft's Bing. Participants will be asked to use this interface to view document summaries, view the underlying documents and change search queries. Participants will be asked to search for correct answers to our pre-defined questions. An example questions is: "How long is the Las Vegas monorail in miles?"
We will collect timing information and associated computer usage data unobtrusively during both phases of the study.

Eye tracking data will be collected. To get accurate gazing data, calibration is needed by following a set of steps. These steps involve participants gazing at different part of the screen for few seconds, multiple times. We anticipate the total time in the calibration task should take a maximum of few minutes to be completed. If a participants eye gazing data cannot be calibrated, the participants will receive $5 for their time.

Video recording of the phone device while completing the study will be used to reply user sessions and to understand user behaviour. We will not record participants faces or any part that can be used to identify the participant. The camera will be setup to shoot an aerial view of the phone device and thus only participants hands while interacting with the phone will be recorded.

c. Will this study involve the administration/use of any drug, medical device, biologic, or natural health product? No

d. Will you be using, processing and/or storing any biological materials of human origin such as blood, tissue, cells or bodily fluids?
No

2. Participants Involved in the Study

a. Indicate who will be recruited as potential participants in this study.
UW Participants:
Undergraduate students
Graduate students
Faculty and/or Staff

b. Describe the potential participants in this study including group affiliation, gender, age range and any other special characteristics. Describe distinct or common characteristics of the potential participants or a group (e.g., a group with a particular health condition) that are relevant to recruitment and/or procedures. Provide justification for exclusion based on culture, language, gender, race, ethnicity, age or disability. For example, if a gender or sub-group (i.e., pregnant and/or breastfeeding women) is to be excluded, provide a justification for the exclusion.

Adults fluent in English, familiar with web search (e.g. Google, Yahoo, Bing), and capable of unassisted use of phones or a computer with keyboard, mouse, and LCD monitor.

c. How many participants are expected to be involved in this study? For a clinical trial, medical device testing, or study with procedures that pose greater than minimal risk, sample size determination information is to be provided.

48 to 60 plus a couple of participants during the pilot phase. The study will involve 12 questions. We know that human performance in text retrieval varies across both humans and the search topics. We use a 12x12 block design (12 participants and 12 topics forming a 12x12 Latin square), and thus will have 48 or 60 participants depending on if we have 4 or 5 blocks. This will be a convenience sample of students and other adults of the University of Waterloo community.

3. Recruitment Process and Study Location

a. From what source(s) will the potential participants be recruited?
Other UW sources: Posters across campus.
CS Graduate mailing list.

b. Describe how and by whom the potential participants will be recruited. Provide a copy of any materials to be used for recruitment (e.g. posters(s), flyers, cards, advertisement(s), letter(s), telephone, email, and other verbal scripts).
We will post posters around the University of Waterloo campus and/or use the CS graduate mailing list to inform people of the study.

c. Where will the study take place? On campus: CPH 4363

4. Remuneration for Participants
Will participants receive remuneration (financial, in-kind, or otherwise) for participation? Yes
If Yes, provide details:
We will pay all participants $15 with an advertised payment of $10 for participation and a $5 bonus for getting at least 10 out of 12 questions answered correctly. Regardless of participant performance, we will pay all participants the full $15. This payment structure is designed to motivate good performance while not harming any person who may be unable to achieve the actual desired performance. Should participants need to leave before completing all tasks, they will be paid on a pro-rated basis for the number of tasks completed rounded to the nearest $5. It is expected that the participants will need to spend around 1 hour to complete the study. Eye

tracking will be part of the study. If a participants eye gazing data cannot be calibrated, the participants will receive $5 for their time.

5. Feedback to Participants

Describe the plans for provision of study feedback and attach a copy of the feedback letter to be used. Wherever possible, written feedback should be provided to study participants including a statement of appreciation, details about the purpose and predictions of the study, restatement of the provisions for confidentiality and security of data, an indication of when a study report will be available and how to obtain a copy, contact information for the researchers, and the ethics review and clearance statement.
Participants will be advised that if they are interested in the outcomes of the study, they may contact the principal investigator at a later time to learn about any resulting publications.

D. POTENTIAL BENEFITS FROM THE STUDY

1. Identify and describe any known or anticipated direct benefits to the participants from their involvement in the project.
There are no known direct benefits to the participants from their involvement in the project.

2.Identify and describe any known or anticipated benefits to the scientific community/society from the conduct of this study.
Information retrieval (text search) has become part of daily life for many Canadians, as well as people around the world. This study has the long term potential to allow researchers to better evaluate retrieval systems. With better evaluation tools that allow for faster and more accurate evaluations, the rate at which retrieval systems improve should increase. With better retrieval systems, people are able to find information previously hidden and the better relevant information sorted, the better decisions they are able to make.

E. POTENTIAL RISKS TO PARTICIPANTS FROM THE STUDY

1. For each procedure used in this study, describe any known or anticipated risks/stressors to the participants. Consider physiological, psychological, emotional, social, economic risks/stressors. A study–specific current health status form must be included when physiological assessments are used and the associated risk(s) to participants is minimal or greater.

Minimal risks anticipated.

Participants will be asked to use either a phone device or a computer with keyboard, mouse, and LCD monitor to answer brief questionnaires as well as to read and answer questions according to given result lists. These activities are common to everyday life and pose no greater risk. The search questions that will be utilized are those that might be posed by a lay person and all of them deal with matters people can encounter in normal life.
All documents come from popular search engines.

2. Describe the procedures or safeguards in place to protect the physical and psychological health of the participants in light of the risks/stressors identified in E1.
As the study involves only minimal risk, no explicit procedures or safeguards will be in place other than to provide a safe, usable phone or computer system in a university.

F. INFORMED CONSENT PROCESS

1. What process will be used to inform the potential participants about the study details and to obtain their consent for participation?
Information letter with written consent form

2. If written consent cannot be obtained from the potential participants, provide a justification for this.

3. Does this study involve persons who cannot give their own consent (e.g. minors)? No

G. ANONYMITY OF PARTICIPANTS AND CONFIDENTIALITY OF DATA

1. Provide a detailed explanation of the procedures to be used to ensure anonymity of participants and confidentiality of data both during the research and in the release of the findings.
All participants will be issued an anonymous identifier (ID). The mapping from a participant's name to the ID will be maintained for the length of the study in case the participant forgets the ID. This mapping will be kept in a locked cabinet in a secure location during the study and will be destroyed at the completion of the study. After the study concludes, there will be no way to identify a participant to the data. All computer usage will be with computers in a University of Waterloo computer lab and not with personally identifiable computers, i.e. participants will not use their own computer. All data collected will be retained indefinitely and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to "participant 1" or some other similar anonymous name. Participants' names will never appear in any publication that results from this study.

2. Describe the procedures for securing written records, video/audio tapes, questionnaires and recordings. Identify (i) whether the data collected will be linked with any other dataset and identify the linking dataset and (ii) whether the data will be sent outside of the institution where it is collected or if data will be received from other sites. For the latter, are the data de-identified, anonymized, or anonymous?
Each of the questions that we provide to the participants are defined by us. The search result list for each of of the questions will be from the Bing search engine. The initial result lists are manipulated and fixed due to the control of quality. We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participant's data will remain identifiable as coming from an individual, i.e. "participant 1", "participant 2", etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

3. Indicate how long the data will be securely stored and the method to be used for final disposition of the data.
Paper Records
Confidential shredding after a minimum of 7 year(s).
Audio/Video Recordings
Erasing of audio/video recordings after a minimum of 7 year(s).
Electronic Data
Erasing of electronic data after a minimum of 7 year(s).
Location: Principal investigator's office (paper) and on secure computers.

4. Are there conditions under which anonymity of participants or confidentiality of data cannot be guaranteed?
No

H. PARTIAL DISCLOSURE AND DECEPTION

1. Will this study involve the use of partial disclosure or deception? Partial disclosure involves withholding or omitting information about the specific purpose or objectives of the research study or other aspects of the research. Deception occurs when an investigator gives false information or intentionally misleads participants about one or more aspects of the research study. Yes

If Yes,
(i) explicitly state if it is partial disclosure and/or deception,
(ii) if applicable, describe the partial disclosure, that is what information is being withheld or omitted concerning the purpose/objectives or procedures,
(iii) if applicable, describe all of the deception(s) to be used in this study, AND
(iv) provide a justification for each use of partial disclosure and deception.

We will use a minor form of deception. We will tell participants that there is a bonus payment of $5 for answering 10 or more of 12 **questio**ns correctly. In all cases, we will award the $5 to the participants regardless of their performance. We are using this form of deception to motivate participants to perform at a high level of quality, for otherwise we fear that participants may simply answer false answers to quickly finish the study.

If Yes, **outline the process to be used to debrief participants**.
Provide a hard copy of the written debriefing sheet for participants, the researcher's verbal debriefing script (if applicable), and for deception, the materials used to obtain consent following debriefing

As part of our thank you / debriefing letter, we will explain that they were awarded the bonus regardless of their performance.

Researchers must ensure that all supporting materials/documentation for their applications are submitted with the signed, hard copies of the ORE form 101/101A. Note, materials shown below in bold are normally required as part of the ORE application package. The inclusion of other materials depends on the specific type of projects.

Protocol Involves a Drug, Medical Device, Biologic, or Natural Health Product

If the study procedures include administering or using a drug, medical device, biologic, or natural health product that has been or has not been approved for marketing in Canada then the researcher is to complete Appendix A. Appendix A is to be attached to each of the one copy of the application that are submitted to the ORE. Information concerning studies involving a drug, biologic, natural health product, or medical devices can be found on the ORE website.

Please **check** below all appendices that are attached as part of your application package:

- Recruitment Materials: A copy of any poster(s), flyer(s), advertisement(s), letter(s), telephone or other verbal script(s) used to recruit/gain access to participants.
- Information Letter and Consent Form(s)*. Used in studies involving interaction with participants (e.g. interviews, testing, etc.)
- Information/Cover Letter(s)*. Used in studies involving surveys or **questio**nnaires.
- Data Collection Materials: A copy of all survey(s), **questio**nnaire(s), interview **questio**ns, interview themes/sample **questio**ns for open-ended interviews, focus group **questio**ns, or any standardized tests.
- Feedback letter *

* Refer to sample letters.

NOTE: The submission of incomplete application packages will increase the duration of the ethics review process.

To avoid common errors/omissions, and to minimize the potential for required revisions, applicants should ensure that their application and attachments are consistent with the *Checklist For Ethics Review of Human Research Application*

Please note the submission of incomplete packages may result in delays in receiving full ethics clearance. We suggest reviewing your application with the Checklist For Ethics Review of Human Research Applications to minimize any required revisions and avoid common errors/omissions.

INVESTIGATORS' AGREEMENT

I have read the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, 2nd

Edition (TCPS2) and agree to comply with the principles and articles outlined in the TCPS2. In the case of student research, as Faculty Supervisor, my signature indicates that I have read and approved this application and the thesis proposal, deem the project to be valid and worthwhile, and agree to provide the necessary supervision of the student.

NEW As of May 1, 2013, all UW faculty and staff listed as investigators must complete the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans Tutorial, 2nd Ed. (TCPS2) prior to submitting an ethics application. Each investigator is to indicate they have completed the TCPS2 tutorial. If there are more than two investigators, please attach a page with the names of each additional investigator along with their TCPS2 tutorial completion information.

_____ _____
Print and Signature of Principal Investigator/Supervisor **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

_____ _____
Print and Signature of Principal Investigator/Supervisor **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

Each student investigator is to indicate if they have completed the Tri-Council Policy Statement, 2nd Edition Tutorial (http://pre.ethics.gc.ca/eng/education/tutorial-didacticiel/). If there are more than two student investigators, please attach a page with the names of each additional student investigator along with their TCPS2 tutorial completion information.

_____ _____
Signature of Student Investigator **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

_____ _____
Signature of Student Investigator **Date**

Completed TCPS2 tutorial:
___YES ___NO ___ In progress

FOR OFFICE OF RESEARCH ETHICS USE ONLY:

_____ _____
Julie Joza, Acting Chief Ethics Officer **Date**

Heather Root, Senior Manager

Karen Pieters, Manager

Joanna Eidse, Research Ethics Advisor

Laura Strathdee, Research Ethics Advisor

Erin Van Der Meulen, Research Ethics Advisor

ORE 101
Revised September 2016

Title of Project: Question Answering Performance of Search Engines

Principal Investigator

Mark D. Smucker (msmucker@uwaterloo.ca) 519-888-4567 x38620

Student Investigators

Mustafa Abualsaud, +1 250-884-9647, m2abuals@uwaterloo.ca,
Haotian Zhang, +1-519-722-6812, h435zhan@uwaterloo.ca

Summary of the Project:

In this study, we will ask participants to use our specifically designed web-search engine for answering various fact-based questions. Participants will be using either a mobile device or a desktop computer to complete the study. Participants will be allowed to enter their own queries and look for web documents that they feel will help them find answers. We will measure participants' behaviour and performance. Eye tracking data will be collected using an eye tracker device. With this data, we plan to build better models of human performance.

Study Eligibility:

In order to participate in the study, you must:

1- Be a fluent speaker of English
2- Require no assistance with using a phone device or a computer with a keyboard, mouse, and LCD monitor.
3- Able to see without corrective lenses (eyeglasses). Eyeglasses can be obstructive and block the view of the participant's eyes resulting in inaccurate and unreliable data.
    a. if you wear contact lenses, be sure to wear your contact lenses to the study session.
4- Not wearing long lashes or mascara. Long eyelashes can be obstructive when the participant's eyes are less open, especially if the participant is wearing mascara. In some cases, eyelashes may completely block the view of the participant's pupils, making eye tracking impossible.
5- Do not have eye conditions/disorder, such as droopy eyelid (ptosis/blepharoptosis) or eye squinting (strabismus). Droopy eyelids or otherwise obstructive eyelids can block the view of the participant's pupils. In some cases, such eyelids may completely block the view of the participant's pupils, making eye tracking impossible

Procedure:

Your participation in this study is voluntary. Participation involves using a search engine to answer questions. One example question is that what is the height of the CN tower?

You will be asked to complete several brief questionnaires and to search for and save answers towards given search questions for 12 topics using a search engine. The questionnaires that you will be asked to complete consist of a demographic questionnaire and a questionnaire concerning the search topic before each search topic task and a questionnaire about the task after each search topic task.

The study will take approximately 1 hour.

We will record both your answers and your interaction with the phone used to complete the study. We may also make note of and record anything we observe, including what you say, while you are participating in the study.

Eye tracking data will be recorded using an eye tracker. To get accurate gazing data, calibration is needed by following a set of steps. These steps involve participants gazing at different part of the screen for few seconds, multiple times. We anticipate the total time in the calibration task should take a maximum of few minutes to be completed. If your eye gazing data cannot be calibrated, you will receive **$5** for your time.

Video recording of the phone device while completing the study will be used to reply user sessions and to understand user behaviour. We will not record your face or any part that can be used to identify you. The camera will be setup to shoot an aerial view of the phone device and thus only your hands while interacting with the phone will be recorded.

You may decline to answer any question that you prefer not to answer.

You may stop participating in the study at any point and withdraw your consent without penalty.

Expectations for your Participation:

Please focus on your own work and continue to work at your own pace. Please work on a given task form start to finish. If you need to take a

break, please do so between tasks. Once you have answered a question, do not attempt to go back and change your answer. All answers are final.

This scientific research study requires your full attention. If you are unable to give this research your full attention, please excuse yourself from the study. In particular:

You may not use your mobile phones during the study [Please keep your phones on silent mode]

You may not listen to music during the study

You may not discuss the answers or talk to other participants during the study

You may not use the phone device/computer for checking email, viewing web pages, or other activities during the study

If you use the phone device/computer for non-study related activities or use a mobile phone, we will end your participation and ask you to leave

Confidentiality and Data Security:

You will be issued an anonymous identifier (ID) as a participant in this study. The mapping from your name to the ID will be maintained for the length of the study in case you forget the ID. This mapping will be kept in a locked cabinet in a secure location during the study and will be destroyed at the completion of the study. After the study concludes, there will be no way to identify you to the data. All computer usage will be with computers in a University of Waterloo computer lab and not with personally identifiable computers, i.e. you will not use your own computer. All data collected will be retained for a minimum of 7 years and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to "participant 1" or some other similar anonymous name. Your name will never appear in any publication that results from this study.

We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participant's data will remain identifiable as coming from an individual, i.e. "participant 1", "participant 2", etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

Remuneration for Your Participation:

You will be paid $10. If you answer at least 10 of the 12 questions correctly, we will pay you a bonus of $5 for a total of $15. Should you stop before completing the study, you will be paid on a pro-rated basis based on the number of questions answered and rounded up to the nearest $5. The amount received is taxable. It is your responsibility to report this amount for income tax purposes. If your participation is ended early due to your actions (e.g. using the device for non-study related activities), you will be paid on the pro-rated basis detailed above. If your eye gazing data cannot be calibrated, you will receive $5 for your time.

Risks and Benefits:

There is minimal risk to you from participation in this study. Phone and computer use and searching for relevant documents are common everyday activities and pose no anticipated risk greater than that encountered in everyday activities. The search topics that will be utilized are those that might be posed by a lay person in regular everyday use of a search engine. All documents come from web sites. The eye tracking device is completely safe and has been tested and approved by certified labs.

There are no direct benefits to you from participation. However, we hope the study will provide results that can lead to advances in the evaluation and development of advanced text retrieval systems that will benefit society at large.

Research Ethics Clearance:

I would like to assure you that this study has been reviewed and received ethics clearance through the Office of Research Ethics at the University of Waterloo. However, the final decision about participation is yours.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#21930). If you have questions for the Committee contact the Chief Ethics Officer, Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

For all other questions please contact the researchers at the email/phone number provided on the first page of this letter. Thank you for your assistance in this project.

CONSENT FORM

I agree to participate in a study being conducted by Mustafa Abualsaud, a PhD student in the University of Waterloo's Department of Computer Science. I have made this decision based on the information I have received in the information letter. I have had the opportunity to ask questions and request any additional details I wanted about this study.

If I participate in this study, I will be asked to complete several brief questionnaires and to search for and save answers towards given questions with a text retrieval system.

As a participant in this study, I am aware that I may decline to answer any question that I prefer not to answer and that I may stop participating in the study at any point and withdraw my consent. I'm also aware that eye tracking data will be collected during my study. Should I stop before completing the study, I will be paid on a pro-rated basis for the number of questions answered and rounded up to the nearest $5. If I answer all of the questions and complete the study, I will be paid $10. If I answer at least 10 of the 12 questions correctly, I will be paid a bonus of $5 for a total of $15. If my eye gazing data cannot be calibrated, I will receive $5 for my time.

I am aware that any identifying information I provide will be kept confidential, and that any data presented, published or shared will be anonymized.

I agree to participate in this study [Question Answering Performance of Search Engines (approximately 60 minutes)]

[ ] YES [ ] NO **(Please check your choice)**

Participant Name: _____

(Please print) Participant Signature: _____

Date: _____

Witness Name: _____

Witness Signature: _____

Department of Computer Science
University of Waterloo

Participants Needed for Research in Web Search

- We are looking for volunteers to take part in a study of **web search** (e.g. Google, Bing, Yahoo).
- As a participant in this study, you would be asked to complete demographic and task-related questionnaires, search for answers toward given questions using a text retrieval system.
- Your participation would take approximately 1 hour.
- In appreciation for your time, you will receive $10 and another $5 for answering at least 10 of 12 questions correctly. All answers can be found using the search engine provided as part of the study.
- Eye tracking data will be recorderd.

- In order to participate in the study, you must:
  1- Be a fluent speaker of English
  2- Require no assistance with using a phone device or a computer with a keyboard/mouse/LCD monitor.
  3- Able to see **without corrective lenses (No eyeglasses)**.
  a. if you have contact lenses, be sure to wear your contact lenses to the study session.
  4- Not wearing long lashes or mascara. Long eyelashes may completely block eye tracking.
  5- Do not have eye conditions/disorder, such as droopy eyelid (ptosis/blepharoptosis) or eye squinting (strabismus).

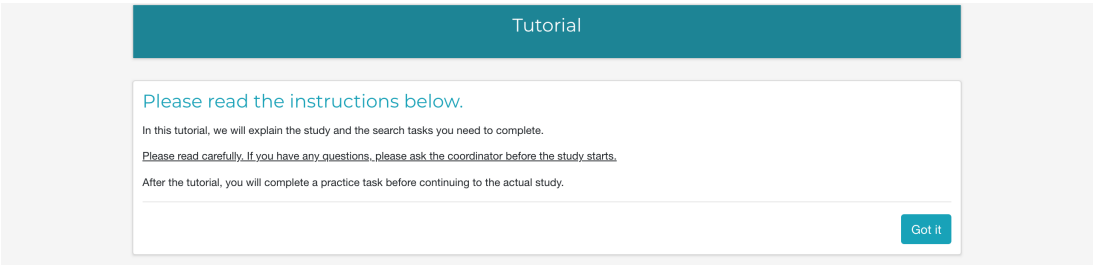
- For more information about this study, please contact:

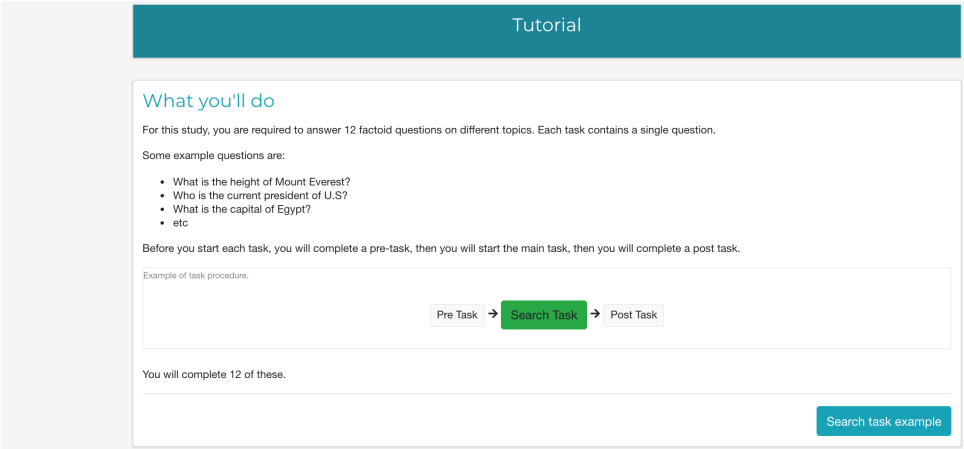
Mustafa Abualsaud
Department of Computer Science
at University of Waterloo
Email: m2abuals@uwaterloo.ca

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#21930).
If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

Please contact: m2abuals@uwaterloo.ca

C.1 Tutorial Screenshots


Tutorial
Please read the instructions below.
In this tutorial, we will explain the study and the search tasks you need to complete.
Please read carefully. If you have any questions, please ask the coordinator before the study starts.
After the tutorial, you will complete a practice task before continuing to the actual study.
Got it


Tutorial
What you'll do
For this study, you are required to answer 12 factoid questions on different topics. Each task contains a single question.
Some example questions are:
What is the height of Mount Everest?
Who is the current president of U.S?
What is the capital of Egypt?
etc
Before you start each task, you will complete a pre-task, then you will start the main task, then you will complete a post task.
Example of task procedure.
Pre Task
Search Task
Post Task
You will complete 12 of these.
Search task example

Tutorial

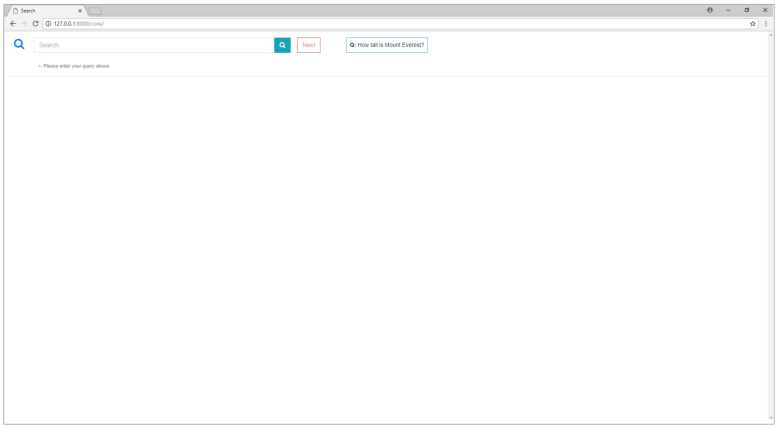
Search task example.

You will be given a factoid question. Your task is to search for the answer using our search engine.

The interface will show the question you need to answer, a search bar to enter your search queries, and a next button to proceed to the next question.

Below is an example screenshot of our search engine for one question: What is the height of Mount Everest.

Example screenshot search task interface.



You should use the search bar to enter a query and look for an answer.

Once you are confident that you found the answer, please say it out loud immediately. For example: I found the answer! it is ....

How to search

Tutorial

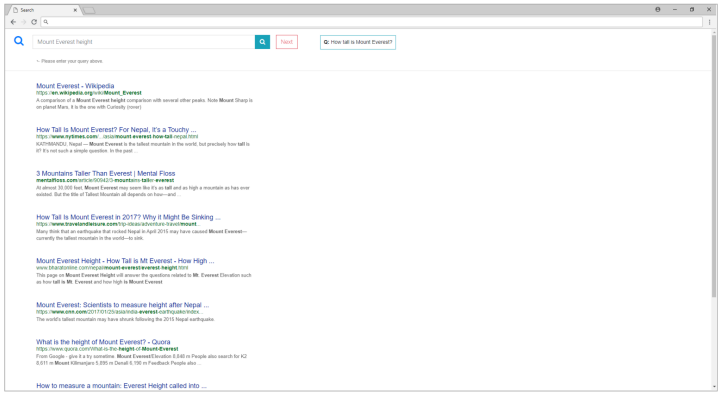
Search task example.

You can submit whatever query you think is good, for example:

- **Mount Everest height**
- **What is the height of Mount Everest in miles?**
- **How tall is Mount Everest**
- etc.

Let's submit the query: **Mount Everest height**.

Example on how to search for an answer.



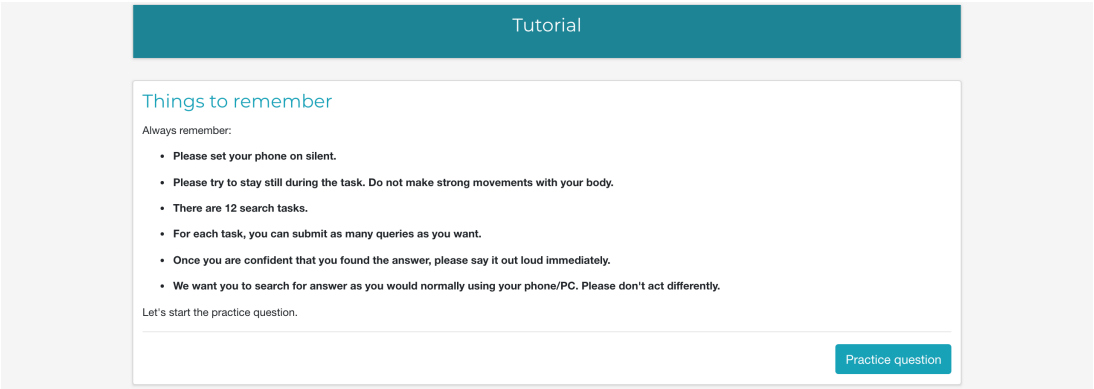
You will get a set of documents just like a regular search engine.

You can click on documents to open them.

Once you are confident that you found the answer, please say it out loud immediately. For example: I found the answer! it is ....

After you find the answer, return to the search page and click on the next button.

Things to remember


Tutorial
Things to remember
Always remember:
Please set your phone on silent.
Please try to stay still during the task. Do not make strong movements with your body.
There are 12 search tasks.
For each task, you can submit as many queries as you want.
Once you are confident that you found the answer, please say it out loud immediately.
We want you to search for answer as you would normally using your phone/PC. Please don't act differently.
Let's start the practice question.
Practice question

Appendix D

Eyetracker Standard Operating Procedures

UNIVERSITY OF WATERLOO

Protocol for Using an Eye Tracker Device for Research Study Participants

SOP created on: December 27, 2017 and **Ethics Clearance Received on:** [pending]
Revised on: December 27, 2017 and **Ethics Clearance Received on**: [pending]
SOP created by: Mustafa Abualsaud, PhD student, Department of Mathematics.
Signature: *MAbualsaud*

Date: December 27, 2017

□ I acknowledge that as the principal investigator/faculty supervisor I am responsible for updating this SOP and notifying the ORE through a modification form (Form 104) if any of the procedures as outlined above change or require revision.

A. PURPOSE AND BACKGROUND

Standard operating procedures are required for any research ethics application involving devices that collect bio-metric data. This SOP describes the procedure for researchers to conduct an information retrieval study with participants eye gazing tracked using an eye tracker device.

B. PROCEDURES/STUDY PROTOCOL

Are there any controlled act(s) to be performed: □ Yes ✓ No

1. Before the arrival of a participant, the researcher mounts the eye tracker in its correct position. The placement of the eye tracker depends on whether the study is conducted using a computer device or phone device (both devices are to be provided by the researcher).

    a. If the participant is to complete the using a computer device, the eye tracker is mounted on the monitor (see figure 2).
    b. If the participant is to complete the using a phone device, the eye tracker is mounted in a mobile device stand (purchased from the same company that developed the eye tracker device). The mobile device stand is mounted to the desk (see figure 3).

2. The participant is invited to take a comfortable seat in the location where the study is conducted. The researcher asks to the participants to read the sign the consent form.

    a. If the participant is to complete the study using a computer device, they will sit at the desk where the computer is placed (see figure 2).
    b. If the participant is to complete the study using a phone device, they will sit at the desk facing the mobile device stand, which will be mounted to the desk (see figure 3).

3. The researcher conducts the eye tracker calibration process.

    a. If the participant is to complete the study using a computer device, the researcher will run the calibration software accompanied with the eye tracker. The software will show multiple objects each placed at different locations within the computer screen. The researcher will then ask the participants to look at each object for few seconds. During this stage, the eye tracker will collect the eye gazing information necessary to complete the calibration process.

    b. If the participant is to complete the study using a phone device, the researcher will run another calibration software accompanied with the eye tracker. The mobile device stands comes with a calibration plate printed with numbers at different locations on the calibration plate. The software will inform the researcher to ask the participant to look at each of the numbers printed on the calibration plate for few seconds. During this stage, the eye tracker will collect the eye gazing information necessary to complete the calibration process.

4. After the calibration is completed, the software will indicate the accuracy of the calibration. If the accuracy is low, the process of calibration should be repeated again. The process may be repeated several times until a good calibration accuracy is achieved.

5. The researcher starts the eye tracker data collecting process and starts the study task.

6. The participant then starts the study tasks using the device they are assigned to. The eye tracker will be collecting their eye gazing data while they complete the study tasks.

7. Upon completion of the study, the researcher stops the eye tracker data collecting process.

8. The participant is thanked for his/her participation and receives their remuneration.

At no time during the calibration process nor the during the study will the participants need to physically touch the eye tracker. The eye tracker is screen based, meaning no physical contact is needed for the eye tracker to work. The participants do not need to wear or be attached to any objects for the eye tracker to work, thus minimal to no risks are anticipated.

C. EQUIPMENT

1. Eye tracker

The eye tracking device that will be used is Tobii Pro X3-120. Tobii Pro X3–120 is CE-marked, indicating compliance with the essential health and safety requirements set out in European Directives. The Tobii Pro X3-120 complies with

the Canadian ICES-003 Issue 6:2016 Class B. The location of this device is fixed (CPH 4363). Participants will need to arrive at the study location to complete the study.

An image of the Tobii Pro X3-120 is shown below:

Figure 1: Tobii Pro X3-120. Image Source: www.tobiipro.com

An image of the Tobii Pro X3-120 mounted on the monitor are shown below: This image shows an example of how the eye tracker is used for tasks completed using a computer device.

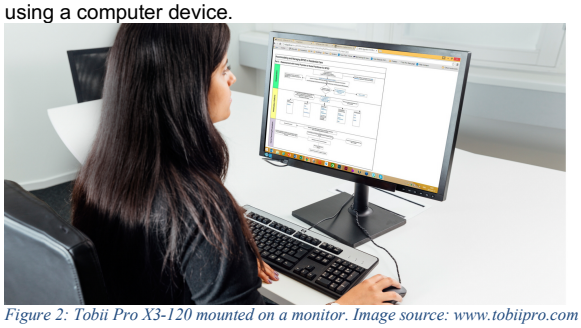


Figure 2: Tobii Pro X3-120 mounted on a monitor. Image source: www.tobiipro.com

2. Eye Tracker Mobile Device Stand

Tobii Pro Mobile Device Stand is designed to be used with Tobii Pro X3-120 Eye Tracker. The propose of this stand is to be able to mount the eye tracker in a stable position, which is necessary to get accurate eye tracking data. The location of this stand is fixed (CPH 4363). Participants will need to arrive at the study location to complete the study.

The Tobii Pro Mobile Device Stand image is shown below.

Figure 3: Tobii Pro Mobile Device Stand mounted to a desk. Image source: www.tobiipro.com

3. Computer device

A Dell computer and monitor will be provided by the researcher to participants who are assigned to use a computer to complete the study tasks. The location of this device is fixed (CPH 4363). Participants will need to arrive at the study location to complete the study.

Figure 4: Dell computer used for completing the study tasks. Image source: www.dell.com

4. Phone device

A Google Pixel 2 phone device will be provided by the researcher to participants who are assigned to use a computer to complete the study tasks.
The location of this device is fixed (CPH 4363). Participants will need to arrive at the study location to complete the study. The phone device does not need a SIM card to work. We will only be using the phone WiFi to access and complete the study.

Figure 5: Google Pixel 2 used for completing the study. Source: www.google.com

D. DESCRIPTION TO STUDY PARTCIPANTS

1. The participants are welcomed in the study and offered a comfortable seat.

2. The participants will be asked to read the consent form and sign it.

3. In the information-consent letter participants will be informed:

    a. An eye tracked will be used to collect eye gazing data during the study.
    b. There is minimal risk for participants to participate in the study.
    c. Questions are welcome at any time before the study starts.
    d. Participants may request to stop the study at any time.

4. Once the participants sign the consent form, the research will ask the participant to prepare for the calibration process.
5. Participants will be instructed on the calibration process procedure and what they are excepted to do.
6. Participants will complete the calibration process by looking at different areas for few seconds while the eye tracker collects the data necessary for calibration.
7. Once the calibration is completed, the participants will complete the study using the device they are assigned to (phone device or computer device).
8. Upon completion of the study tasks, the participant is thanked and will receive their study remuneration.

Photos of how the equipment will be used by a study participants are shown in figure 2 and 3.

E. RISKS

1. PARTICIPANTS

    a. General anxiety.
    b. Tiredness or fatigue.

2. RESEARCHERS

a. There are no known risks to the researchers implementing the procedure as a result of the procedure itself, or the equipment.

F. SAFEGUARDS/SAFETY PROCEDURES

1. PARTICIPANTS

a. To minimize general anxiety, a familiarization period will be used to explain the eye tracker procedure to participants and to inform participants that the device is not harmful.
b. General precaution by the researcher will be applied at all times during the study.
c. Participants are informed in the consent form that they are allowed to withdraw from participating in the study at any time and for any reason they provide.

2. RESEARCHERS

a. The researcher is to have completed:
    i. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2)

G. REFERENCES (if applicable)

- *Tobii Pro X3-120*
  https://www.tobiipro.com/product-listing/tobii-pro-x3-120/
- *Tobii Pro Mobile Device Stand*
  https://www.tobiipro.com/product-listing/mobile-device-stand/
- *Google Pixel 2*
  https://store.google.com/product/pixel_2
- *Dell Computer*
  http://www.dell.com/en-ca/shop/desktops/sc/desktops

Appendix E

Visualizing Searcher Gaze Patterns

E.1 Visualization Code

```
library(ggplot2)
library(reshape2)
#
# load data.
#
dta <- read.csv("transformed_data.csv", header=TRUE)
first_x_seconds = 10 # only show first x seconds.
dta = dta[,0:(first_x_seconds*4 + 2)]
dta <- dta[nrow(dta):1,]
dta$Rank = factor(dta$Rank, levels=rev(sort(dta$Rank)))
#
# create and plot the visualization
#
(p <- ggplot(melt(dta), aes(variable, Rank)) +
  geom_point(aes(colour = value), size=3) +
  theme(axis.text.x = element_text(angle = 90,
  hjust = 1, vjust=0.5, size=7)) +
  theme(panel.background = element_rect(fill = "white",
  colour = "white", size = 0.5, linetype = "solid")) +
  scale_color_gradientn(limits=c(0, 1),breaks=seq(0, 1, .2),
  colours =c("#e31a1c", "#fd8d3c", "#fecc5c", "#ffffff"),
  values = c(1, .60 ,.30 ,0), labels=seq(0, 100, 20)) +
  scale_y_discrete(labels=seq(10,1,-1)) +
  labs(x="Time (seconds)", y="Rank"))
```

Listing 1: Sample R script to create the visualization.

Appendix F

The Effect of Non-Relevant Results on Mobile Search Behavior

#42519 - Web Search on Mobile Devices

Protocol Information

Review Type	Status	Approval Date	Renewal Date
Expedited	**Approved**	**Dec 23, 2020**	**Oct 07, 2021**

Expiration Date	Initial Approval Date	Initial Review Type
Oct 30, 2021	**Oct 29, 2020**	**Expedited**

Feedback

Approval Comment

This amendment has received ethics clearance. If you have any questions please contact Karen Pieters at karen.pieters@uwaterloo.ca.

Protocol Amendment Form

Amendment

Justification

Adding a new source of recruitment through social media (Reddit). The post will be posted in two subreddits (/r/UWaterloo and /r/Waterloo). The reason for this change is to help get more users to join the study.

General Information

Only the Principal Investigator/Faculty Supervisor can submit the application. This acts as a signature indicating approval of the application.

Principal Investigator / Faculty Supervisor

Mark Smucker

Department

Management Sciences

Study title

Web Search on Mobile Devices

General Questionnaire

Indicate the type of application you would like to complete

Standard application *

* The Standard application is for faculty level research and thesis level research.

** The course project application is for (non-thesis) course based research and can be completed by students or the course instructor

Please confirm:

I understand that the type of applications listed above determine the form I am about to complete. If I have chosen the incorrect form I acknowledge that I may need to complete a new application.

People

University of Waterloo research team

Person

Mark Smucker

Waterloo Department

Management Sciences

Email Address

msmucker@uwaterloo.ca

Phone

Researcher Role

Principal Investigator

Permissions

Full Access

Has this person completed the CORE (TCPS2) tutorial?

Yes

Date of completion on TCPS2 certificate (Required)

June 13, 2014

Upload a copy of the TCPS2 certificate (Highly recommended, optional at this time)

As per the Waterloo policy on mandatory research ethics training, if you completed the TCPS2 tutorial more than 5 years ago, you may be asked to update your training within the next 6 months. You will be notified by email if this is the case.

Person

Mustafa Abualsaud

Waterloo Department

Faculty of Mathematics

Email Address

m2abuals@uwaterloo.ca

Phone

Researcher Role

Student investigator

Permissions

Full Access

Has this person completed the CORE (TCPS2) tutorial?

Yes

Date of completion on TCPS2 certificate (Required)

June 16, 2016

Upload a copy of the TCPS2 certificate (Highly recommended, optional at this time)

TCPS2_CORE_CERTIFICATE.PDF

As per the Waterloo policy on mandatory research ethics training, if you completed the TCPS2 tutorial more than 5 years ago, you may be asked to update your training within the next 6 months. You will be notified by email if this is the case.

Do you have any investigators external to the University of Waterloo

No

General details

Is this new study related to any previous application?

Yes

Previous UWaterloo ORE/Kuali #

30836, 21930

What is the estimated start date and end date for the study?

Start Date

November 15, 2020

End Date

August 1, 2021

Does this research require approval from a UWaterloo departmental committee?

Not a department requirement

What is the level of the research to be conducted? Choose one.

PhD dissertation research

Will this study involve Wilfrid Laurier University, Western University, Conestoga College or Local hospitals covered by the Tri-Hospital Research Ethics Board (Cambridge Memorial Hospital, Grand River Hospital and St. Mary's General Hospital)?

No

Has a version of this study been disapproved or rejected by any Research Ethics Board/Committee?

No

Study description

State your research question(s)

State your research question(s)

RQ1: To what extent, if any, does coherence or diversity of documents affect users' examination and therefor their abandonment rate in interactive information retrieval?

Provide a clear, detailed description of the purpose, hypothesis, aim, and objectives of this study

Purpose: To collect information on how people use web search for the purpose of answering fact-based and informational questions. Hypothesis: We hypothesize that search behavior will differ when users are presented with a diverse set of search results as opposed to a coherent set of results. Justification for the Study: While the IR field offers good analysis on how user search for documents, it is lacking information on how much users, given a set of documents, are motivated to find answers, and when do they decide to give up on the list of documents they are provided with and decide to generate a new list. Objectives: This project will collect data allowing us to estimate a conditional probability of user changing search query and searching for new results given the search result lists with varying ranking quality. With this data, we can construct a model of human performance and compare its predictions to actual human performance also measured as part of this project.

Provide background information, a rationale, and justification for conducting this study. Describe why the research is being done and what research has already been done in this area. Be sure to explain why this research is important.

In our previous studies, we have examined relationship between the quality of search results and their effect on search behaviour, and have shown that as search result quality decreases, the rate of users abandoning their search increases. In this study, we hope to refine our definition of quality of search results and to study search behaviour when users are presented with diverse or a coherent set of search results.

In a maximum of 250 words, provide a non-scientific lay language description that summarizes the project outlining the purpose, anticipated benefits, and basic procedures. Write this summary as if it would be read by members of the general public who are not familiar with academic terms or acronyms. Use language suitable for a media release.

When a search result does not satisfy a user’s needs, the user often abandons their query and submits a reformulated query in the hopes of receiving better search results. The action of abandoning search results is termed “query abandonment”. This research project investigates the rate in which people

abandon their queries and how the quality of their queries may affect their decision to abandon search results.

What is the study design?

Randomized controlled user study.

Is this a pilot study?

No

Sample Size

What is the expected sample size? Outline the number of participants anticipated to take part in the study.

24-36

Was a formal sample size calculation completed?

No

Provide a rationale for the number of participants specified

Each user will complete 12 tasks balanced using a 12x12 Graeco-Latin square. Therefore the number of participants needs to be a multiple of 12. We anticipate that 24/36 users would suffice.

Study sites

Where is this study taking place?

University of Waterloo

Are there any permissions required to conduct this study on campus?

No

Funding

Is the study funded/will it be funded?

Yes

Funding
List all funding sources that are new or ongoing

Funding status

Ongoing funding

Funding source is

Tri-agency / Canadian Government sponsor

Canadian Government agency

NSERC - Natural Sciences and Engineering Research Council of Canada

Program name if applicable

Discovery

Work-order or award number, if known

50503-11157

What is the expected period of funding

Funding from

April 1, 2020

Funding to

March 31, 2025

Conflict of interest

Are there any potential, perceived, or actual financial or non-financial conflicts of interest of the research team in undertaking the proposed research?

No

No

Benefits

Are there direct benefits of the proposed research to the study participants?

No

What are the scientific and/or scholarly benefits of the proposed research?

Information retrieval (text search) has become part of daily life for many Canadians, as well as people around the world. This study has the long term potential to allow researchers to better evaluate retrieval systems. With better evaluation tools that allow for faster and more accurate evaluations, the rate at which retrieval systems improve should increase. With better retrieval systems, people are able to find information previously hidden and the better relevant information sorted, the better decisions they are able to make.

Participants

Participant general categories

University of Waterloo undergraduate and/or graduate students
University of Waterloo staff and/or faculty
Adults (age 18-64 years)

Describe the sample in detail and list any specific inclusion/exclusion criteria for the study

1) The participant must be proficient in English. 2) People who have previously participated in our eye-tracking study (ORE#21930) will be excluded from our study. 3) Require no assistance with using a phone device or a computer. 4) Must have a smartphone (Android or iOS) with Zoom installed

If you are excluding people on certain characteristics provide a justification for the exclusion.

The user study tasks require proficiency in English. The new study may include some of the study questions that were asked in our previous user study (ORE#21930). To eliminate any confounding factors due to familiarity, we plan to recruit people that have not previously participated in our study

We plan to recruit people that have not previously participated in our study.
The study will require people to use their own smartphone to complete the study. Participants should be familiar with how to use their own devices.

Will a screening process be used to determine eligibility in the study based on the inclusion and/or exclusion criteria identified above?

Yes

Is a screening questionnaire to be used?

No

How is the screening to be conducted?

Our recruitment email will include eligibility criteria. Participants that email back with interest in participants will be asked to verify that they meet the eligibility criteria.

When is the screening to be conducted in relation to other study procedures (i.e., consent)?

People will be asked to verify that they meet the eligibility criteria before agreeing to participate in the study.

What are individuals told if they do not meet the eligibility criteria?

"Thank you for showing interest in our study. Unfortunately, you do not meet the eligibility criteria set for our study. If you have any further questions, please do not hesitate to email us."

What will be done with the information or data collected if an individual is deemed ineligible to participate in the study?

We plan not to include ineligible individuals in our data collection process.

Recruitment

Identify from where/what sources potential participants will be recruited.

Through email/internet (e.g., social media networks)
Other

Indicate what email listing, internet site or network you intend to recruit from

cs-grad mailing list. Graduate studies and postdoctoral affairs - Call for study

participants website Reddit (/r/UWaterloo and /r/Waterloo)

Provide details on your other recruitment source

NA

What recruitment materials will be used?

Email script
Social media

Describe how social media will be used

A Reddit post will be posted in the following subreddits: UWaterloo, Waterloo. The post information is in the file attached.

Upload your recruitment materials

Upload your recruitment materials

EMAIL-MESSAGE (1) (1) (1) (1).DOCX

Study group

Upload your recruitment materials

REDDIT_POST.DOCX

Study group

Upload your recruitment materials

SCREENCAPTURE-UWATERLOO-CA-GRADUATE-STUDIES-POSTDOCTORAL-AFFAIRS-CURRENT-STUDENTS-CALL-STUDY-PARTICIPANTS-QUESTION-ANSWERING-PERFORMANCE-SEARCH-ENGINES-2020-12-23-12_02_07.PDF

Study group

Will potential participants be recruited through pre-existing relationships with members of the research team (e.g., employees, students, or patients of research team, acquaintances, own children or family members, colleagues, etc.)?

No

Methods and procedures

Which of the following will be conducted for this study?

Surveys/questionnaires
One-on-one interviews
Observation
Other

Describe the other procedure

Interactions with our software will be logged.

How will the survey(s) or questionnaire(s) be administered?

Online or web

Provide the URL of the survey, if available

Attached

How will the one-on-one interviews be conducted?

Online – Video Chat

Will quotations be used in the write-up of the study

Yes

What type of quotations will be used?

Anonymous

What type of observations are planned?

Participant observation (where the researcher engages in, and observes, the action; participant knows they are being observed)

Will the people being observed have an expectation of privacy?

Yes

How will this expectation of privacy be upheld for participants?

Anything unrelated to the study will not be collected or observed.

For each of the procedures indicated above, provide a detailed, sequential description of how they will be used in the study.

Recruiting: We will send an email to the CS-mailing list to ask for participants interested in the study. People who qualify and are interested in participating will be provided with a link that contains a participant id (e.g. participant_1, participant_2). The participant id will be used to fill the consent form, collect questions answers and data. Data collected: We will collect demographic data, and data of participants interacting with the search system (e.g. queries submitted, clicks). The interaction data will help us determine how people interact with search results on mobile devices. We plan to use this data to analyze user search behaviour on mobile devices. At the end of the study, participants will be asked about their experience and if they have any feedback. Quotes from the participants answers may be used in the final analysis of the study. Study Design: The study includes 12 search tasks. The order of the search tasks is randomized to reduce any order bias. Randomization will be done in accordance to a Graeco-Latin square.

Please upload any study materials related to the procedure(s)

Study material

PROCEDUREQMOBILE.DOCX

Study material

INTERVIEW GUIDE.DOCX

Does the study involve the administration or use of an approved drug or natural health product?

No

Will you be collecting any biological specimens?

No

Will you be creating or contributing to a bio-bank, bio-repository, registry, as part of the study?

No

Will you be doing any genetic testing or analysis?

No

Incidental and secondary findings

See Guideline for reporting incidental and secondary findings to study participants

Are any of the methods or procedures used likely (i.e., a real possibility and probability) to reveal an incidental finding (i.e., discoveries made in the course of research but that are outside the scope of the research and/or results that are outside the original purpose for which a test or procedure was conducted)?

No

Are any of the methods or procedures used likely to reveal a secondary finding (i.e., findings that are not the primary target of the test or procedure; rather, it is an additional result that is actively sought)?

No

Equipment use

Will there be any equipment used as part of this study?

No

Deception

Does the study involve deception or partial disclosure?

No

Risks and safeguards

Considering each method or procedure to be used in this study, indicate if participants might experience any of the following risks or harms

No known or anticipated risks

Outline the criteria for stopping the study early due to safety concerns/other issues.

none

Privacy

Will demographic and/or background information be asked of participants?

Yes

What demographic/background information will be collected?

Age
Gender
Education

Will demographic/background information be collected separately from names and other identifying information?

Yes

Participant identification

Participants will be identified with a unique number, e.g. P01, P02, etc.

If applicable how will the key/list that links participants' codes with their actual name and/or consent forms be stored and protected?

The mapping from a participant's name to the ID will be maintained for the length of the study in case the participant forgets the ID. This mapping will be kept in a secure computer system during the study and will be destroyed at

the completion of the study. After the study concludes, there will be no way to identify a participant to the data. All data collected will be retained for 7 years and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to “participant 1” or some other similar anonymous name. Participants' names will never appear in any publication that results from this study

Are there any limitations to the promise of confidentiality?

No

Will any study data be leaving the University of Waterloo, the province, or country (e.g., member of research team is located in another institution, province, or country, etc.)?

Yes

Will any identifiable participant information be leaving the University of Waterloo, the province, or country (e.g., member of research team is located in another institution, province, or country, etc.)?

No identifiable information being collected

Where will the study data be sent? Why is it necessary for it to leave the University of Waterloo?

We may share the data collected from participants with other researchers.

Explain what data will be leaving the University of Waterloo, who will receive it, why they need access, and what safeguards will be used to protect the identity of participants and the privacy of their data.

No identifying information will be shared. Only queries submitted by participants and their user behaviour data may be shared with other researchers. We believe that sharing the data can advance the research in this area.

Describe the measures in place to ensure secure transfer of study data outside of the University.

If the data is going to be shared, it will be compressed and available to download online. We do not believe that we need to ensure a secure transfer of the study data as it will not include any information that could be used to identify or harm the participants.

Has a research data agreement or data transfer agreement been created?

No

No

Will any collected data or information be entered into a database for future use?

No

Are there other members of the research team who are not named on this application (e.g., co-op students, research assistants, or other temporary personnel) who may carry out specific tasks involved in your study?

No

Will individual participant identities be confidential in the publication or release of the study findings?

Yes

Data storage

What type(s) of data will be collected for this study?

Electronic files

For each type of information collected, identify where the data will be stored

Data will be stored in the investigator personal computer while a participant is conducting the study. The data will then be remotely transferred to a UW computer stored in a safe location in the university.

For each type of data collected, identify the minimum retention period

7 years.

Data Management

Are there plans to link the data collected with other data sets, databases, or registries?

No

The Tri-Agency Open Access Policy on Publications and some journals are requesting that

The TT Agency Open Access Policy on Publications and some journals are requesting that research data be provided to an open access repository to promote the availability of findings, to enhance transparency and share with the widest possible audience.

Do researchers plan to make the anonymized data-set available in an online repository?

Unsure at this time

Do you have a data management plan?

No

Data management planning is necessary at all stages of the research project lifecycle, from design and inception to completion. Data management plans are key elements of the data management planning process. They describe how data is collected, formatted, preserved and shared, as well as how existing datasets will be used and what new data will be created. They also assist researchers in determining the costs, benefits and challenges of managing data.

Consent and Withdrawal

What member(s) of the research team will be responsible for obtaining informed consent?

Mustafa Abualsaud

Is there a relationship between the potential participant(s) and the person obtaining consent?

No

How will consent be obtained

Online consent (e.g., click one of two radio buttons)

Upload Information and Consent Materials

Upload Information and Consent Materials

CONSENTMOBILE (1) (1) (1) (1).DOCX

Study group

Do you anticipate that you will need to make special accommodations for your participant group?

No

No

Do you anticipate needing to put in place any special procedures when obtaining informed consent?

No

Will consent need to be re-documented throughout the life of this study?

No

Describe how participants will be informed of their right to withdraw from the study.

In the consent form shown at the beginning of the study, participants will be told that they can "You can terminate the experiment at any time. Withdrawing from the study will not result in any negative consequences for you."

Outline what will be done with the participant's data if they withdraw from the study.

It will be discarded.

Will any individuals taking part in this study be unable to provide their own informed consent?

No

Remuneration

Will there be remuneration provided to show appreciation for a participant's time, effort, skills, etc. to take part in the study?

Yes

Type of remuneration

Other

Explain the other remuneration

Cash (i.e., $15 CAD) which is being provided via e-transfer.

If a participant withdraws from the study will remuneration be pro-rated?

Yes

Explain your plans for pro-rating the remuneration

The study consists of 12 tasks. Pro-rating will depend on how many tasks were completed (1.25$ for each task)

Will participants incur any expenses by participating in the study?

No

Feedback and Appreciation

How will you show appreciation to participants for taking part in the study?

After participants complete the study, they will be shown the letter of appreciation.

When will feedback/appreciation be provided to participants (e.g., immediately after the session, at the end of a survey, mail results at time X.)?

At the end of the study.

Upload Feedback/Appreciation materials

Upload Feedback/Appreciation materials

FEEDBACK (3) (1).DOCX

Study group

How can participants learn about the study results/obtain a summary of the findings if interested?

Participants can email the investigator for more information on the findings.

Other Details

Provide any other information relevant to this study you wish to explain to the Research Ethics Committee reviewers or to the staff in the Office of Research Ethics

Committee reviewers or to the staff in the Office of Research Ethics.

Na

Other Attachments

Upload any additional study documents

Attachments

Attestation

As the Principal Investigator/Faculty Supervisor/Local Investigator, I attest to the following:

- I will ensure all co-investigators, collaborators, and student investigators listed on this application have reviewed the application contents and will conduct the study according to the application/protocol.
- I am aware that any changes made to the research must be reviewed and provided clearance before the changes are implemented. Change requests (i.e., an amendment) are to be submitted through the system. I am also aware ethics clearance for this study is valid for only 12 months unless I renew the study prior to the ethics clearance expiry date. If an annual renewal report is NOT submitted through the system prior to the expiry date, the study will be suspended, all work on the study must stop, and Research Finance will be notified which will result in a hold being put on the funds associated with this study.
- I agree to comply with the Tri-Council Policy Statement (TCPS2) for conducting research with human participants and with University of Waterloo policies and guidelines when conducting this study (e.g., statement on human participant research, IST policies, etc.).
- I confirm I have read the University of Waterloo Research Integrity guidelines and I agree to comply with the policies and guidelines of my profession or discipline regarding the ethical conduct of research involving humans.

By submitting this application I agree to the above attestations and will ensure the research is conducted accordingly

Only the Principal Investigator/Faculty Supervisor can submit the application. This acts as a signature indicating approval of the application.

This is the end of the application form. Click submit in the right menu if you are ready to send it to the Research Ethics Office.

Title of Project: Question Answering Performance of Search Engines

Principal Investigator (Supervisor)

Mark D. Smucker (msmucker@uwaterloo.ca) 519-888-4567 x38620

Student Investigators

Mustafa Abualsaud (Computer Science department), m2abuals@uwaterloo.ca

Summary of the Project:

This is a research study on question answering using web search. In this study, we will ask participants to use our specifically designed web-search engine for answering various fact-based and informational questions.

For the study, you will be given a set of simple factoid and informational questions (e.g. what is the height of Mount Everest in miles?). Your task is to search for answers to the given questions as you would normally do using a web search engine. For this study, you will be using our search engine interface to search and find answers. Participants will be using their own mobile device to complete the study. Participants will be allowed to enter their own queries and look for web documents that they feel will help them find answers. We will measure participants' behaviour (clicks and queries submitted) and performance. With this data, we plan to build better models of human performance.

You will need a smartphone and personal laptop/desktop with internet access and a browser to participate and complete the study. Participation require downloading Zoom (an secured video conferencing application) on your phone. Your interactions with the browser on the phone screen for the study session will be video recorded. Only your interaction with the webpage associated with the user study will be screen recorded. You will not be required to turn on your camera. However, you are required to share your phone screen while you are using our software. The study will be conducted fully online using Zoom. Unique meeting links will be generated and provided before the study. Personal identifying information will kept completely confidential, and all your interactions with the browser and your feedback will be kept completely confidential.

Study Eligibility:

In order to participate in the study, you must:

1. Be proficient in English
2. Require no assistance with using a phone device or personal computer.
3. Have a smartphone (Android or iOS) with Zoom installed.
4. Have access to Zoom in your personal computer.
5. Have not participated in a similar information retrieval user study.

Procedure:

Your participation in this study is voluntary. Participation involves using a search engine to answer questions. One example question is that what is the height of the CN tower?

You will be asked to complete several brief questionnaires and to search for and save answers towards given search questions for 12 topics using a search engine. The questionnaires that you will be asked to complete consist of a demographic questionnaire and a questionnaire concerning the search topic before each search topic task and a questionnaire about the task after each search topic task. Questionnaires will be collected via our own web application.

The study will take approximately 1 hour.

We will record both your answers and your interaction with the phone used to complete the study. We may also make note of and record anything we observe, including what you say, while you are participating in the study.

You may decline to answer any question that you prefer not to answer.

You may stop participating in the study at any point and withdraw your consent without penalty.

At the end of the study, you will be asked few questions regarding your experience with the user study and any feedback you might have.

Expectations for your Participation:

Please focus on your own work and continue to work at your own pace. Please work on a given task form start to finish. If you need to take a break, please do so between tasks. Once you have answered a question, do not attempt to go back and change your answer. All answers are final.

This scientific research study requires your full attention. If you are unable to give this research your full attention, please excuse yourself from the study. In particular:

Your phone must be in do not disturb mode.

You may not listen to music during the study.

You may not use your phone device for activities other than that related to the study.

Confidentiality and Data Security:

You will be issued an anonymous identifier (ID) as a participant in this study. The mapping from your name to the ID will be maintained for the length of the study in case you forget the ID. This mapping will be stored digitally in a secure location during the study and will be destroyed at the completion of the study. After the study concludes, there will be no way to identify you to the data. All data collected will be retained for a minimum of 7 years and will be used for research purposes. We may refer to individual participants when describing the results or the study, and in these cases, we will always refer to “participant 1” or some other similar anonymous name. Your name will never appear in any publication that results from this study.

We may choose to distribute the data collected to other researchers. All data will be anonymized at the conclusion of the study and prior to any distribution, but each participant’s data will remain identifiable as coming from an individual, i.e. “participant 1”, “participant 2”, etc. We will not publicly share this data, i.e. the data would only be made available to other researchers for research purposes.

You will be completing the study by an online survey operated by Zoom. When information is transmitted or stored on the internet privacy cannot be guaranteed. There is always a risk your responses may be intercepted by a third party (e.g., government agencies, hackers). Zoom temporarily collects your Zoom ID and computer IP address to avoid duplicate responses in the dataset but will not collect information that could identify you personally.

Remuneration for Your Participation:

You will be provided with an $15 CAD. Should you stop before completing the study, you will be paid on a pro-rated basis based on the number of tasks completed (1.25$ for each task). The amount received is taxable. It is your responsibility to report this amount for income tax purposes. If your participation is ended early due to your actions (e.g. using the device for non-study related activities), you will be paid on the pro-rated basis detailed above.

Risks and Benefits:

There is minimal risk to you from participation in this study. Phone use and searching for relevant documents are common everyday activities and pose no anticipated risk greater than that encountered in everyday activities. The search topics that will be utilized are those that might be posed by a lay person in regular everyday use of a search engine. All documents come from web sites.

There are no direct benefits to you from participation. However, we hope the study will provide results that can lead to advances in the evaluation and development of advanced text retrieval systems that will benefit society at large.

Research Ethics Clearance:

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #42519). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

For all other questions please contact the researchers at the email/phone number provided on the first page of this letter. Thank you for your assistance in this project.

ONLINE CONSENT FORM

I agree to participate in a study being conducted by Mustafa Abualsaud, a PhD student in the University of Waterloo's Department of Computer Science. I have made this decision based on the information I have received in the information letter. I have had the opportunity to ask questions and request any additional details I wanted about this study.

If I participate in this study, I will be asked to complete several brief questionnaires and to search for and save answers towards given questions with a text retrieval system.

As a participant in this study, I am aware that I may decline to answer any question that I prefer not to answer and that I may stop participating in the study at any point and withdraw my consent. Should I stop before completing the study, I will be paid on a pro-rated basis for the number of questions answered (1.25$ for each task).

I am aware that any identifying information I provide will be kept confidential, and that any data presented, published or shared will be anonymized.

I agree to participate in this study [Question Answering Performance of Search Engines (approximately 60 minutes)]

[ ] YES [ ] NO **(Please check your choice)**

Title:

Call for Participation in a Web Search User Study ($15 for 1 hour).

Post:

This post is sent on behalf of the researchers.

Call for Participation in an Web Search User Study

This is a research study on question answering using web search. In this study, we will ask participants to use our specifically designed web-search engine for answering various fact-based and informational questions.

What do I need to do?

You will be asked to complete several brief questionnaires and to search for and save answers towards given search questions for 12 topics using a search engine. The questionnaires that you will be asked to complete consist of a demographic questionnaire and a questionnaire concerning the search topic before each search topic task and a questionnaire about the task after each search topic task. Questionnaires will be collected via our own web application.

What I will get?

You will be provided with an $15 CAD. You will need a bank account that accepts e-interac transfer.

Who can join?

In order to participate in the study, you must:

1. Be proficient in English.
2. Require no assistance with using a phone device or personal computer.
3. Have a smartphone (Android or iOS) with Zoom installed.
4. Have not participated in a similar information retrieval user study.

How to join

For more information about the study and how to sign up, please see:
https://uwaterloo.ca/graduate-studies-postdoctoral-affairs/current-students/call-study-participants/question-answering-performance-search-engines

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee.