The Dark Side of Relevance: The Effect of Non-Relevant Results on Search Behavior

Mustafa Abualsaud
School of Computer Science
University of Waterloo
Waterloo, Canada

Mark D. Smucker
Department of Management Sciences
University of Waterloo
Waterloo, Canada

ABSTRACT

Understanding and modelling user behavior with search results is important to both search engine designers and the design of effectiveness measures. It is well established that users are less likely to view lower ranked search results, and recent research has shown that the type of relevant documents can influence when people stop examining results. However, while existing measures and research consider that relevant documents vary in utility and make use of relevance grades or preference judgments, non-relevant documents are largely all treated the same. In this paper, we show that the nature of non-relevant material affects users' willingness to further explore a ranked list of search results. We first broaden our notion of non-relevant documents and define a spectrum of possible search engine result pages (SERPs). At one end of the spectrum, the search results were filled with off-topic non-relevant documents, and at the other end, the non-relevant documents were all on-topic, but failed to match the required sub-topic of the search task. We conducted a user study where participants used a mobile search interface to find answers to questions, and collected participants' behavior while interacting with different SERPs on our spectrum. Our results show 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. When the SERP is coherent on an egregious topic, 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. Our research implies that to improve information retrieval evaluation, we should be assessing the degree of non-relevance in search results as well as the degree of relevance.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Users and interactive retrieval**; **Relevance assessment**; **Presentation of retrieval results**.

CHIIR '22, March 14–18, 2022, Regensburg, Germany

ACM ISBN 978-1-4503-9186-3/22/03…$15.00
https://doi.org/10.1145/3498366.3505770

KEYWORDS

Non-relevant documents; user study; user behavior

ACM Reference Format:
Mustafa Abualsaud and Mark D. Smucker. 2022. The Dark Side of Relevance: The Effect of Non-Relevant Results on Search Behavior. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22), March 14–18, 2022, Regensburg, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3498366.3505770

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. In the simplest sense, users are more likely to examine further down a ranked list if they cannot find a relevant result, and if they find relevant results, they are less likely to continue down the ranked list. This gross simplification, 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 [1], and Azzopardi et al. [5] have shown that the degree of document relevance influences the extent to which people continue to examine search results or not.

Moffat and Wicaksono [27] 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 paper 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 (e.g., how far in the SERP are users willing to examine?). We first define a spectrum of different possible ways a SERP can be designed and constructed (see Figure 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 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 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 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 1A), search results on multiple egregiously non-relevant topics and one within-topic non-relevant result (Figure 1C), or search results on multiple topics that are within the topic of the search but not relevant to the task (Figure 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.
- While not statistically significant differences, we found across the three non-relevant SERP conditions of our study that users were quicker to abandon the SERP the worse the SERP's quality was. 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 7.1 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 are likely to 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 information retrieval (IR). We discuss these implications further in Section 6.

2 RELATED WORK

2.1 User-model Based Effectiveness Measures

From previous research, it is well known that users are less likely to view lower-ranked search results [7, 12, 16], and this behavior is captured in the widely used normalized discounted cumulative gain (nDCG) effectiveness measure [14]. Since the release of nDCG, many other user-model based effectiveness measures have been proposed [15, 26, 28, 31, 33, 39], and each of these measures aims to better evaluate search quality through improvements to the measure's model of user behavior.

An important aspect of behavior to model is the degree to which users will continue down a ranked list to examine search results. If a user stops examining search results, there is no chance for them to find any relevant documents beyond where they stop. Position-based models, such as nDCG and rank-biased precision (RBP) [29], model user's examination by introducing a discount based on the rank of the item in the ranked-list, regardless of its relevance. Lower-ranking items have a higher discount that reflect lower chances of examination. Cascade-based models, such as expected reciprocal rank (ERR) [6], model the chances of examining a search result based on previously examined results. Additionally, after a user examines a relevant search result, existing cascade models [6, 38] believe that users are more likely to be satisfied and thus more likely to stop their examination, as opposed to when they examine non-relevant results. When a search result has a high level of relevance, it is more likely to satisfy users, and thus follow-up results have a lower chance of being examined.

Other measures, such as INST [25], model examination in an adaptive manner based on the aggregate volume of relevance viewed. When the user views more relevant documents, they are less likely to examine more results. de Vries et al. [8] proposed to model examination behavior based on users' tolerance to non-relevant results. This is expressed using a parameter representing the maximum time that we expect a user would keep reading non-relevant material. As the user is exposed to more non-relevant results, they are more likely to stop examining further results. More recently, Jiang and Allan [15] proposed that the probability of examining results at different ranks should be adaptable to the overall quality of the SERP. This is motivated by the behavior that users are more likely to stop browsing when the SERP is of low quality. Moffat and Wicaksono [27] note that the approach of of Jiang and Allan [15] requires users to thoroughly understand and recognize the quality

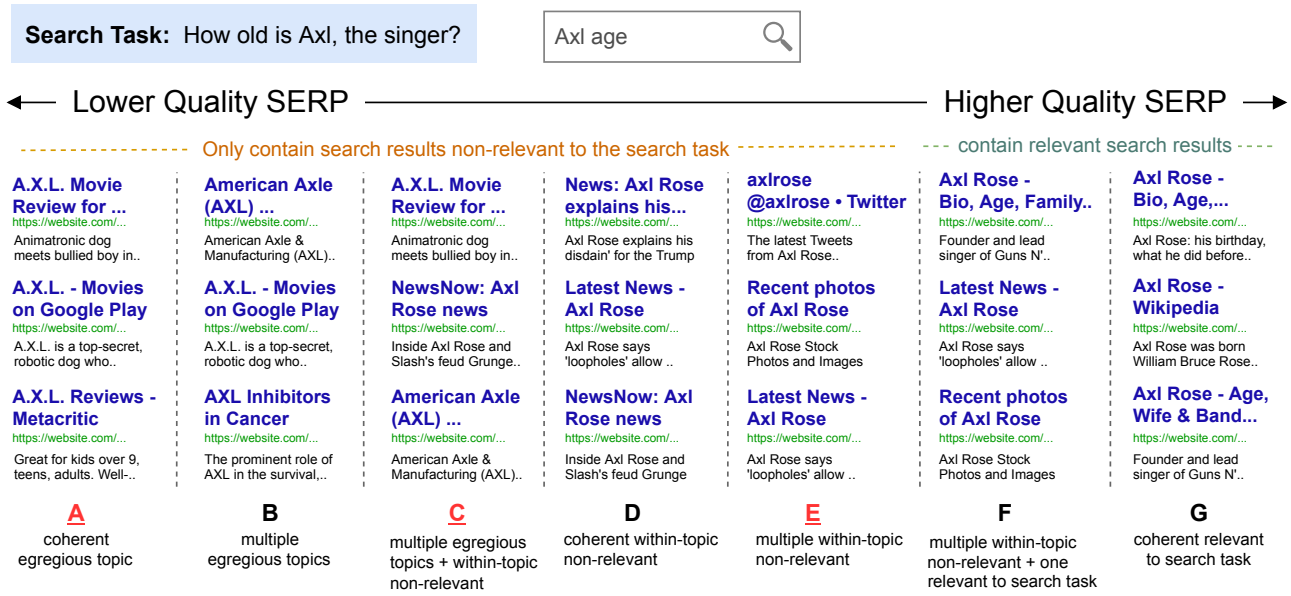

Figure 1: A spectrum of search engine result pages (SERP) representing different overall qualities for the search task "How old is Axl, the singer?". For the SERPs in the left side of the spectrum (A to E), every search result is considered not relevant to the search task. SERPs F and G contain search results relevant to the search task. In this paper, we focus on studying user behavior in the three underlined scenarios: A, C and E. In Section 3, we describe each SERP and explain the quality it represents.

of the SERP before the user is able to examine any of its document. In Moffat and Wicaksono's work, the authors extended INST with a mechanism to model the observation that some users may exit SERPs quickly due to the accumulation of low-quality results. They conclude by saying that further user-based evidence is needed to understand how the nature of non-relevant documents can affect user behavior. Our paper extends Moffat and Wicaksono's work by investigating user behavior when users are presented with different notions of non-relevant results.

2.2 User Examination of Search Results

2.2.1 Studying User Behavior. Several researchers have studied user search interaction [4, 7, 10–13, 17, 18, 20]. One of the earliest and notable research is the work of Granka et al. [12]. They found that users spend most of their attention on search results placed higher in the SERP, and generally examine results from top to bottom. Other researchers have found that users are biased towards clicking top results [17, 21].

User type can also play a role in how people examine results. Using eye tracking, previous researchers [2, 4, 10, 18] have found that users fall into multiple categories. Aula et al. [4] classified users as "economic" or "exhaustive" depending on what strategy they follow while examining search results. Economic users follow an economic strategy, e.g., in more than half of tasks, they would scan at most the first three results before making a decision. Exhaustive users would examine more search results than economic users and sometimes scroll down the page to view more results. Like Aula et al. [4], Dumais et al. [10] found that users have different examination strategies and that there is a significant difference between how economic and exhaustive users spend their time examining search results.

2.2.2 Query abandonment. One related aspect that can affect examination is the behavior of abandoning search results, commonly known as query abandonment [36, Chapter 3.2.2.1]. Query abandonment is when a user decides not to click on any search results and either reformulates their query or quits the search process. There are many reasons why people might abandon their queries [9, 34]. The most common reason is due to dissatisfaction with the search results [9]. Abandoning the page while unsatisfied is referred to as "bad abandonment". The other type of abandonment, referred to as "good abandonment", occurs when a direct answer on the SERP satisfies the user's information need [19].

Many researchers have investigated query abandonment as part of experiments seeking to understand how users search information in the web [1, 7, 30, 37, 40]. A particularly related work is that by Wu et al. [37]. In their work, the authors conducted a user study where participants are tasked to find answers to open-ended questions. When people submit their queries to the search engine, they were shown manipulated search results that contain different amounts of relevant results ranked according to different patterns. They found that the number of relevant documents in the SERP affects the rate of query abandonment. Zhang et al. [40] conducted a user study to investigate the rate at which people abandon their search results on different SERPs. In their experiment, users were shown carefully manipulated SERPs that contain either no relevant documents or a single relevant document placed at ranks 1 to 10. The authors found that as the top-most relevant document is placed lower on the search page, the probability of query abandonment increases. Zhang

et al. [40] hypothesized that user type (whether an economic or an exhaustive user) may influence a user's examination but lacked eye-tracking data to verify this behavior. In a follow-up work [1], we used an eye tracker to investigate how far in the SERP people examine before abandoning their results and what factors influence their decision to continue or stop their examination. We confirmed that examination of search results is influenced by user type, and also showed that regardless of where the relevant document is placed in the SERP, the type of query submitted affects examination. If a user enters an ambiguous query, they are likely to examine fewer results.

In another related work, Maxwell et al. [23] conducted multiple simulation experiments to determine how far down the rank list people examine before stopping their search. The authors created different stopping rules (e.g., a threshold on the total number of results the user has examined so far) as part of their simulated user model. They found that stopping rules representing disgust or frustration (e.g., the number of continuous non-relevant documents seen) provide the closest approximation to actual user behavior. However, the non-relevant documents in Maxwell et al. [23] simulations, as well those in [1, 30, 37, 40], were all regarded the same. In our work, we focus on how the nature of those non-relevant documents can affect user interaction.

2.2.3 Search result representation. Search result diversification is an important area in IR. Given an ambiguous query, it is a common strategy for search engines to diversify search results to include different possible topics. For example the query "jaguar" could have multiple meanings, such as the car manufacturer or the animal. While some of the topics might be considered off-topic by the user, the benefit of diversification is that it increases the chances of retrieving relevant material. Some research has looked into how diversification can affect search behavior [3, 24]. Maxwell et al. [24] conducted a user study to investigate how diversification can affect the number of queries and document clicks issued by users. In their experiment, participants used two different search engines, one with and without diversification, to complete ad-hoc and aspectual search tasks. They found that participants issued more queries and clicked on fewer documents per query when using a diversified search engine. Arguello and Capra [3] 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, Arguello and Capra [3] focused on the coherence between two search components: images and web results. They found that when web results are diversified, image results in the SERP can have significant effect on user interaction with the web results.

3 DIFFERENT SEARCH RESULTS SCENARIOS

The utility of a search engine result page (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 1, we show possible scenarios a search engine might return for the search task "Axl singer 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 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 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 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.

While the concept of relevance used above and in Figure 1 is shown as one-dimensional, the nature of relevance can be multi-dimensional (e.g., involving multiple aspects such as recency, novelty, etc.) and complex [22, 32]. In this work, we focused on non-relevance and adopted a notion of non-relevance that reflects some

aspects of topicality and usefulness. Egregiously non-relevant results can be considered neither topically relevant nor useful to the task the user is trying to accomplish, whereas within-topic non-relevant results can provide the user with some topical information but are not useful when it comes to helping them with their search tasks. In our design of Figure 1 and construction of search results, we selected documents we think will be judged by a user to be of these various grades.

4 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 "More results" button (i.e., button to show extra results in the SERP) 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.

5 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 5.3 discusses the user interface in more detail.

5.1 Experimental Protocol

Figure 2 shows an overview of the experiment protocol. 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

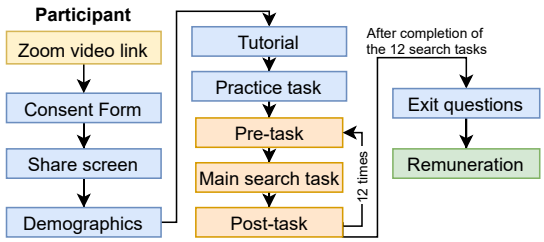

Figure 2: Overview of the experimental protocol.

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 in a semi-structured interview from the researcher regarding their search behavior.

5.2 Search Tasks

We asked participants to complete 12 search task questions. The questions required no prior knowledge and were likely to be familiar to participants. The questions are shown in Table 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 [1]. 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 earlier work [1] and the TREC Web 2012 and 2014 Tracks. The search tasks were designed to be simple, so as not to confuse people and to reduce confounding variables that may arise.

5.3 Search Interface

Figure 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 our initial inspections show that most phones can fit three search results within the page fold.

[1] azure.microsoft.com/services/cognitive-services/bing-web-search-api/

Table 1: A list of the search tasks used in our user study. The related subtopics and egregious topics are used to construct the search engine result page for our tasks scenarios. The cells shaded in blue in the egregious 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). The related subtopics column is what we consider within-topic non-relevant.

#	Type	(Topic) / Search Task	Related subtopics (not relevant to search task)	Egregious topics
1	Factoid	(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	Factoid	(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	Factoid	(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	Factoid	(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	Factoid	(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	Factoid	(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	Info.	(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	Info.	(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	Info.	(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	Info.	(Fidel Castro) **Find some quotes from Fidel Castro, the 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	Info.	(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	Info.	(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

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.

The web application in Figure 3 was built using the Django and JavaScript (JS). JS was used for client-time tracking of user behavior, such as clicks, keystrokes, and dwell time.

5.4 Constructing Search Results Pages

For each topic in Table 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.

5.4.1 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 1 for that task (shaded in blue). For scenario **C**, we selected the two egregious topics from Table 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 1 for that task.

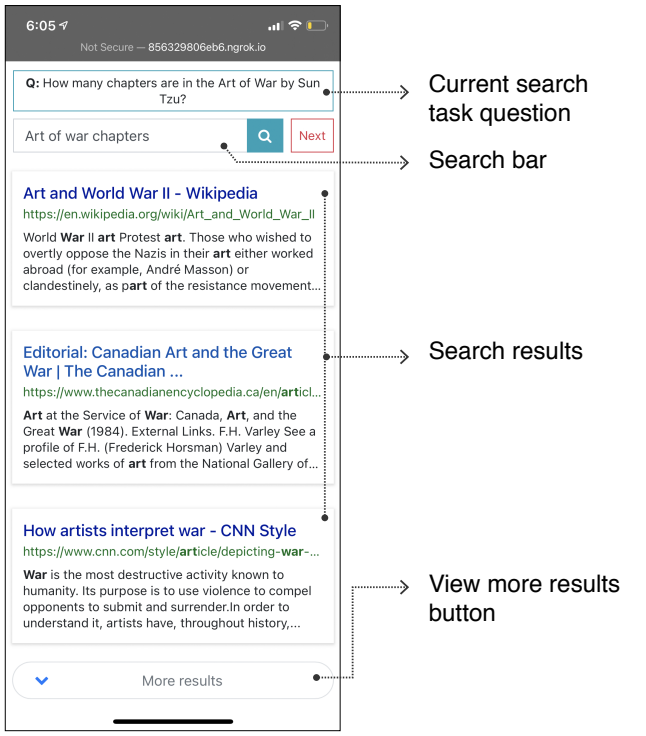

Figure 3: Screenshot of the search interface. The interface shows 3 search results by default. The more results button shows an extra three results, up to 15 results for each query.

5.4.2 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 1 for task #6 and in Figure 3 for task #3 under the single egregious topic (**A**) scenario.

5.4.3 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. For example, for the task on Axl Rose, any query that contains the term "Axl" (case-insensitive) 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.

5.5 Study Design

The study followed a within-subject 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.

5.6 Participants

After receiving ethics approval from our university's Office of Research Ethics, we advertised the 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.

5.7 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 some of the search snippets. Bad abandonment is when a user leaves the search results because they are unsatisfied, without clicking at 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 RESULTS

In our study, participants used our search engine to find answers to 12 search tasks shown in Table 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 3. 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 predetermined relevant term to the task. All subsequent queries results were fetched from the Bing API. There were one task in each condition where a participant was shown Bing results instead (i.e., user entered other terms), and these were excluded from the analysis.

Figures 4, 6, 5, and 7 show our main results. In Figure 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 of requesting more results is the lowest (0.28) compared to **C** (0.41),

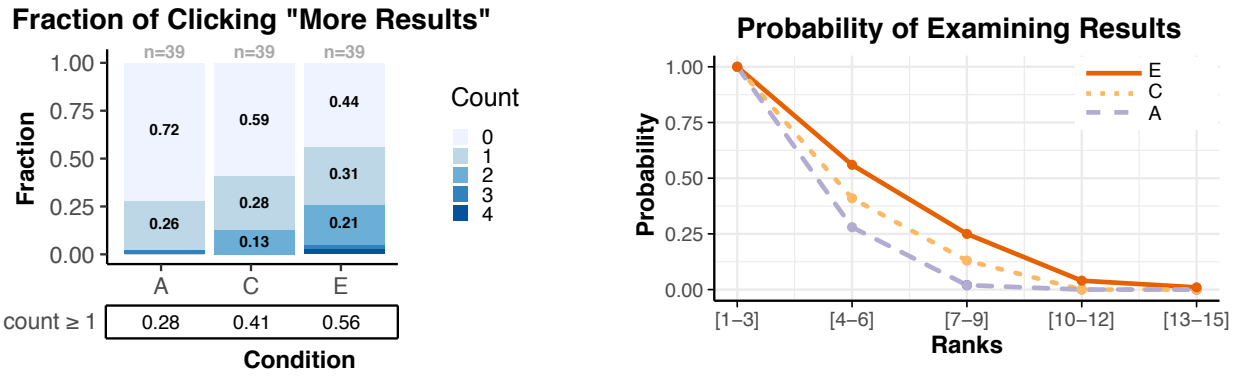

Figure 4: Left: The fraction of clicking at "More results" button under each condition. The count indicates the number of times a user requested to view more results. Among our three conditions, users' lowest fraction of viewing more results is when they are presented with results coherent on a single 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. Right: 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.

and **E** (0.56). Using a chi-square test (with Yates correction), Table 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 (χ^2 = 5.25, p = .02). 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 and appears to be similar across the conditions, with condition **E** slightly lower than the other conditions. We did not find any statistically significant difference in the time to first "more result" click.

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 more, the ranks are grouped into sets of three. Figure 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 5 shows the probability of the first action in each condition. The probability of abandonment 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**). Using a z-score test, the difference between **A** and **E** is statistically significant (z = 2.0321, p = .042). 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 to the search task. Users who clicked on a wrong document returned back to the SERP and reformulated their query.

Figure 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 Kruskal-Wallis test. The results of the test showed that the effect of the type of condition on the time to query abandonment was not significant ($\chi^2(2)$ = 3.8492, p = 0.0.14).

Figure 4 and Table 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 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 not find any statistically significant difference between the conditions. While the results in Figure 7 are not statistically significant, we see a trend that reflects what we hypothesized.

Table 2: Result of chi-square test of independence (with Yates correction) between experimental conditions and requests for more search results. Star symbol indicates statistically significance (p < 0.05).

	A	C
C	χ^2 = 0.91, p = .34	
E	χ^2 = 5.25, p = .02*	χ^2 = 1.28, p = .26

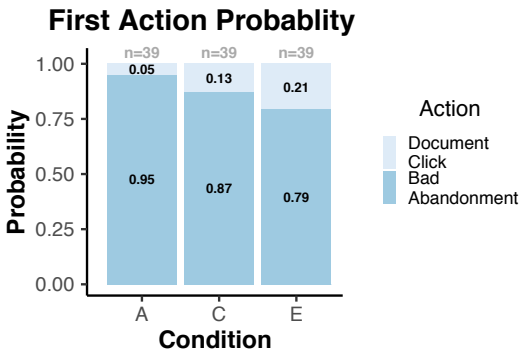

Figure 5: The probability of the first action users would make after being presented with the search results.

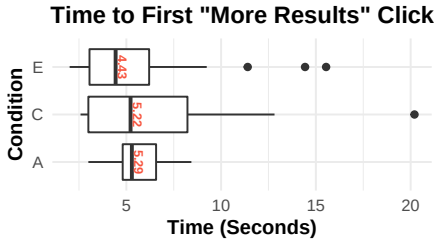

Figure 6: Time to first click on "More results" button.

7 DISCUSSION

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 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 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. [35]). 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).

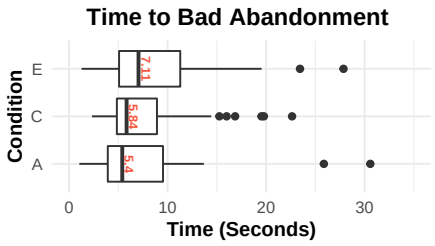

Figure 7: Time to bad abandonment under each condition.

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 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 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*".

8 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 confounding variables. More complex questions, such as those that require more understanding or memory recall are shown to require different search behavior [25, 26].

9 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 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.

ACKNOWLEDGMENTS

We thank Charles Clarke for his feedback. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (RGPIN-04665-2020).

REFERENCES

[1] Mustafa Abualsaud and Mark 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* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1833–1842. https://doi.org/10.1145/3357384.3358041

[2] Mustafa Abualsaud, Mark D. Smucker, and Charles L. A. Clarke. 2021. *Visualizing Searcher Gaze Patterns.* Association for Computing Machinery, New York, NY, USA, 295–299. https://doi.org/10.1145/3406522.3446041

[3] Jaime Arguello and Robert Capra. 2012. The Effect of Aggregated Search Coherence on Search Behavior. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, Hawaii, USA) *(CIKM '12)*. Association for Computing Machinery, New York, NY, USA, 1293–1302. https://doi.org/10.1145/2396761.2398432

[4] Anne Aula, Päivi Majaranta, and Kari-Jouko Räihä. 2005. Eye-Tracking Reveals the Personal Styles for Search Result Evaluation. In *Human-Computer Interaction - INTERACT 2005*, Maria Francesca Costabile and Fabio Paternò (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1058–1061.

[5] Leif Azzopardi, Ryen W White, Paul Thomas, and Nick Craswell. 2020. Data-driven evaluation metrics for heterogeneous search engine result pages. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 213–222.

[6] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[7] Edward Cutrell and Zhiwei Guan. 2007. What Are You Looking for? An Eye-Tracking Study of Information Usage in Web Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 407–416. https://doi.org/10.1145/1240624.1240690

[8] Arjen P. de Vries, Gabriella Kazai, and Mounia Lalmas. 2004. Tolerance to Irrelevance: A User-Effort Oriented Evaluation of Retrieval Systems without Predefined Retrieval Unit. In *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval* (Vaucluse, France) *(RIAO '04)*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, FRA, 463–473.

[9] Abdigani Diriye, Ryen White, Georg Buscher, and Susan 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* (Maui, Hawaii, USA) *(CIKM '12)*. Association for Computing Machinery, New York, NY, USA, 1025–1034. https://doi.org/10.1145/2396761.2398399

[10] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *Proceedings of the Third Symposium on Information Interaction in Context* (New Brunswick, New Jersey, USA) *(IIiX '10)*. ACM, New York, NY, USA, 185–194. https://doi.org/10.1145/1840784.1840812

[11] Carsten Eickhoff, Sebastian Dungs, and Vu 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* (Santiago, Chile) *(SIGIR '15)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/2766462.2767703

[12] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking Analysis of User Behavior in WWW Search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, United Kingdom) *(SIGIR '04)*. ACM, New York, NY, USA, 478–479. https://doi.org/10.1145/1008992.1009079

[13] Zhiwei Guan and Edward 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* (San Jose, California, USA) *(CHI '07)*. ACM, New York, NY, USA, 417–420. https://doi.org/10.1145/1240624.1240691

[14] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[15] Jiepu Jiang and James Allan. 2017. Adaptive Persistence for Search Effectiveness Measures. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 747–756. https://doi.org/10.1145/3132847.3133033

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri 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* (Salvador, Brazil) *(SIGIR '05)*. Association for Computing Machinery, New York, NY, USA, 154–161. https://doi.org/10.1145/1076034.1076063

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri 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* (Salvador, Brazil) *(SIGIR '05)*. ACM, New York, NY, USA, 154–161. https://doi.org/10.1145/1076034.1076063

[18] Kerstin Klöckner, Nadine Wirschum, and Anthony Jameson. 2004. Depth- and Breadth-first Processing of Search Result Lists. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. ACM, New York, NY, USA, 1539–1539. https://doi.org/10.1145/985921.986115

[19] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good Abandonment in Mobile and PC Internet Search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 43–50. https://doi.org/10.1145/1571941.1571951

[20] Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From Skimming to Reading: A Two-stage Examination Model for Web Search. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (Shanghai, China) *(CIKM '14)*. ACM, New York, NY, USA, 849–858. https://doi.org/10.1145/2661829.2661907

[21] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology* 59, 7 (2008), 1041–1052. https://doi.org/10.1002/asi.20794 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20794

[22] Kelly L. Maglaughlin and Diane H. Sonnenwald. 2002. User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology* 53, 5 (2002), 327–342. https://doi.org/10.1002/asi.10049 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.10049

[23] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 313–322. https://doi.org/10.1145/2806416.2806476

[24] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2019. The impact of result diversification on search behaviour and performance. *Inf. Retr. J.* 22, 5 (2019), 422–446. https://doi.org/10.1007/s10791-019-09353-0

[25] Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. 2017. Incorporating User Expectations and Behavior into the Measurement of Search Effectiveness. *ACM Trans. Inf. Syst.* 35, 3, Article 24 (June 2017), 38 pages. https://doi.org/10.1145/3052768

[26] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus Models: What Observation Tells Us about Effectiveness Metrics. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 659–668. https://doi.org/10.1145/2505515.2507665

[27] Alistair Moffat and Alfan Farizki Wicaksono. 2018. Users, Adaptivity, and Bad Abandonment. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 897–900. https://doi.org/10.1145/3209978.3210075

[28] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (2008), 27 pages. http://doi.acm.org/10.1145/1416950.1416952

[29] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (Dec. 2008), 27 pages. https://doi.org/10.1145/1416950.1416952

[30] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk 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* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. ACM, New York, NY, USA, 295–304. https://doi.org/10.1145/3077136.3080817

[31] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. 473–482. http://doi.acm.org/10.1145/2484028.2484031

[32] Tefko Saracevic. 2016. The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really? *Synthesis Lectures on Information Concepts, Retrieval, and Services* 8, 3 (2016), i–109. https://doi.org/10.2200/S00723ED1V01Y201607ICR050

[33] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-based Calibration of Effectiveness Measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. 95–104. http://doi.acm.org/10.1145/2348283.2348300

[34] Sofia Stamou and Efthimis N. Efthimiadis. 2009. Queries without clicks: Successful or failed searches. In *In Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 13–14.

[35] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy 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* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 565–574. https://doi.org/10.1145/2766462.2767760

[36] Ryen W White. 2016. *Interactions with search systems*. Cambridge University Press, New York.

[37] Wan-Ching Wu, Diane Kelly, and Avneesh 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* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/2600428.2609626

[38] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. 2010. Expected Browsing Utility for Web Search Evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1561–1564. https://doi.org/10.1145/1871437.1871672

[39] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating Web Search with a Bejeweled Player Model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. 425–434. http://doi.acm.org/10.1145/3077136.3080841

[40] Haotian Zhang, Mustafa Abualsaud, and Mark D. Smucker. 2018. A Study of Immediate Requery Behavior in Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/3176349.3176400