Effects of Time Constraints and Search Results Presentation on Web Search

M.F. Beijen

MSc Thesis

TU Delft

Effects of Time Constraints and Search Results Presentation on Web Search

by

M.F. Beijen

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday July 16, 2021 at 13:30.

Student number:	4542339	
Project duration:	November 10, 2020 – July 16, 2021	
Thesis committee:	Prof. dr. ir. G. J. P. M. Houben,	TU Delft, chair
	Dr. P. Pawełczak,	TU Delft
	Dr. U. K. Gadiraju	TU Delft, supervisor

Preface

Before you lies my thesis looking into the effects of time constraints and search results presentation on web search. This thesis concludes my MSc Computer Science at the faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) at the Delft University of Technology.

Without a doubt, I would like to express my gratitude to Ujwal Gadiraju for introducing me to the topic of the thesis and for his excellent guidance and advice throughout my entire thesis. Also, I would like to thank David Maxwell for the critical and instructive feedback provided during his involvement. Unmistakably, the critical yet constructive discussions we have had together were beneficial for this work.

Working on this thesis during the Covid-19 pandemic has proven to be challenging. Needless to say, I want to thank my friends and family for their moral support during this unusual period of time.

Mike Beijen
Delft, July 2021

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Abstract

Everybody experiences time constraints in their day-to-day lives. These time constraints may induce stress, possibly influencing our judgment abilities and behavior. In contemporary daily life, search engines are regularly consulted to look for information online with searchers commonly experiencing time constraints in the web search process. With an increasing percentage of the global population having access to the internet and thus a search engine, more people will experience being time-constrained during the search process.

The generally unfavorable consequences of these time constraints have been examined in various stages of the search process. Despite the importance of this issue, how the design of the web page showing the search results, or the Search Engine Results Page (SERP), may benefit in time-constrained web searches constitutes a substantial and important knowledge gap that this work aims to address. Concretely, we investigate how different time constraints influence task performance and search behavior. Consequently, we aim to determine to what extent various SERP interfaces are susceptible to the effects of time constraints. We also want to know how different SERP interfaces impact user experience and we examine to what extent affinity for technology interaction moderates the relationship between time constraint and task performance.

We aim to address these questions through a crowdsourced 4 (SERP interfaces) × 4 (time constraints) between-subjects factorial design user study in which participants are tasked with searching the web to find a list of arguments supporting or opposing a controversial topic. To examine to what extent user interfaces are susceptible to the effects of time constraints, participants make use of a mock search system displaying one of four SERP interfaces depending on the experimental condition: a list interface (baseline, traditional interface), a grid interface, a list-like interface without the snippet, and an interface with the snippet placed to the right of other data. The used time constraints are 2, 5, and 8 minutes in addition to a condition without a time constraint. The effects of SERP interfaces and time constraints are evaluated in terms of task performance metrics, search behavior, and user experience.

Results have shown that task performance is considerably decreased by stricter time constraints. Also, as time constraints tightened, the rate at which participants issued queries increased. Exploratory results suggest this comes as the cost of reduced depth to which individuals click on results in the ranked list. As for the interaction between SERP interfaces and time constraints, SERP interfaces have not been found to be susceptible to the effects of time constraints. Interestingly, user experience was neither worsened nor improved because of the experimental SERP interfaces and affinity for technology interaction was not found to play a moderating role in the relationship between time constraints and task performance.

List of Figures

List of Tables

Contents

1 Introduction

In this chapter, the motivation and the research questions forming the foundation of this thesis are laid out. Next, the contributions and outline of this work are presented.

1.1. Motivation

We live in a world in which time is becoming an ever more important aspect in the daily lives of many people. Being under time pressure can have many causes such as deadlines, public transport disruptions, or family obligations. Time constraints may induce stress, possibly influencing cognitive processes and mental well-being. In contemporary daily life, we regularly consult search engines to look for information online, thus we regularly experience time constraints in the web search process.

An increasing share of the global population has access to the internet and hence a search engine. The possibilities of search engines have grown to such an extent that they are used for much more than just search, further stimulating search engine usage. An unusual feature of search engines is that they have a wide target audience containing all layers of the population: people of varying ages, ethnicity, religion, income, education, or employment status all use search engines; hence, the use of search engines is becoming practically ubiquitous. Therefore, more people will experience being time-constrained during the web search process.

The effects of these time constraints are reflected in web searches. In terms of the search process, experiencing time pressure has, among other effects, resulted in changed search strategies [47], faster-made decisions [15], worsened user experience [14], and differences in search behavior such as query rates, dwell times, and time spent examining documents [13, 15]. The search outcome, decisions made, knowledge gained, news articles studied, recommendations made, etc. could suffer from time constraints as a potential source of influence. As an illustration of the possible consequences of time constraints in search: the use of a search system in a clinical decision-making study decreased gained accuracy from 32% to only 6% as time pressure increased [79]. Moreover, time pressure has been shown to shape the length and specificity recommendations made in a series of decision-making tasks [15]. Hence, time pressure in information retrieval is a facet that is of significant importance. The common denominator between most research regarding time pressure is the fact that they vary only the presence of a time constraint. How the *duration* of time constraints affect the search process and search outcome constitutes a promising area of inquiry.

As [14] suggests, considering the consequences of time-pressured search on the search process, search outcome, and human well-being, directing research efforts into how the user interface of a search engine may assist those under time pressure is justified. While search engines cannot change the time pressure their users are experiencing, they can change the web page on which search results are presented - the *Search Engine Results Page* (SERP). The current body of literature provides only a limited insight into the relationship between user interfaces and search behavior. For example, usage of a grid interface has been shown to reduce the impact of the position of the search result [40] and more equally divide the attention of the user over the search results compared to using a list interface [38]. Regarding search result snippets in the user interface: while longer snippets gave searchers the feeling they performed better, it did not result in a significant performance improvement [56]. Yet, to what extent elements of the SERP interface may cater people under time constraints and perhaps at what costs constitutes a knowledge gap.

The relevance of an investigation into time constraints and the extent to which those effects are susceptible to different SERP interfaces is now established. The next section will outline the utterances in this section translated into research questions and present the hypotheses.

1.2. Research Questions and Hypotheses

Using a crowd-sourced user study, this thesis aims to explore how user interfaces and different levels of time constraints influence task performance, user behavior, and user experience. A conceptual model visualizing the relationships between the hypotheses can be found in Figure 1.1. Based on the motivation presented in the previous section, five research questions (**RQ**) and corresponding hypotheses (**H**) were defined below.

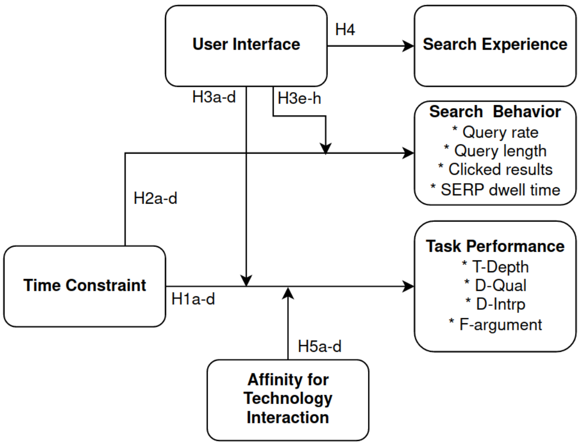


Figure 1.1: Conceptual model of the variables and hypothesis to be tested.

- **RQ1: How do different time constraints influence task performance?**
  Task performance will be determined by evaluating the outcome of a time-constrained search. Existing literature tends to only vary the presence of a time constraint (i.e., a time constraint of *n* minutes versus no time constraint), not its duration. These hypotheses are based on the interpolation of results of works like [13, 79] reporting a decrease in task performance with the introduction of time constraints and the approximately linear decrease in task performance as time constraints tighten in [79]. Due to the meager amount of literature available on this topic, the effects of variations in time constraint duration are promising to investigate. We therefore propose the following four hypotheses.
  - *H1a:* Stricter time constraints reduce the level of topic focus (T-Depth).
  - *H1b:* Stricter time constraints reduce the quality of arguments (D-Qual).
  - *H1c:* Stricter time constraints reduce interpretation of data into arguments (D-Intrp).
  - *H1d:* Stricter time constraints reduce the number of arguments extracted (F-Argument).
- **RQ2: How do different time constraints affect search behavior?**
  Crescenzi et al. found that, when searchers were placed under time constraints, query rates increased and fewer search results per query were examined [13] and decisions in a decision-making task were made quicker [15]. Similar findings have been noted by Liu and Wei [47] who showed that searchers' strategies moved from "economic" to "cautious" as users were presented with time constraints in the sense that more time was spent on SERPs trying to find relevant search results and the number of search results viewed per query decreased. Note that these works only varied the presence of a time constraint, not the actual duration of time constraints. We therefore propose the following four hypotheses.
  - *H2a:* Stricter time constraints increase the query rate (number of queries issued per minute).

  - *H2b:* Stricter time constraints decrease the average length of queries.
  - *H2c:* Stricter time constraints increase the depth to which individuals will click on results in the ranked list.
  - *H2d:* Stricter time constraints increase the time spent on the SERP on average, at the cost of time spent reading web pages.

- **RQ3: In what way are different SERP interfaces susceptible to the effect of time constraints?**
  Various experimental UIs are shown to have an effect when used not under time constraints; Kammerer and Gerjets [38, 40] have shown the advantages of a grid interface in terms of source trustworthiness and search result position and Cutrell and Guan [16] and Clarke et al. [11] have shown snippet features such as query terms presence and readability may influence web search behavior. In terms of search behavior, Kammerer and Gerjets [39] showed that the SERP interface influences dwell time and clicked search results depending on search result trustworthiness. Also, Joho and Jose [37] suggest the experimental SERP interfaces used in their study resulted in increased engagement in query (re)formulation. Hence, the SERP interface is expected to moderate task performance and search behavior. In exploring this moderation, the proximity to the results observed in a control condition that emulates the typical search experience is a measure of the susceptibility to the effects of time constraints. We propose the following hypotheses.

  - *H3a:* SERP interface moderates level of topic focus (T-Depth).
  - *H3b:* SERP interface moderates quality of arguments (D-Qual).
  - *H3c:* SERP interface moderates interpretation of data into arguments (D-Intrp).
  - *H3d:* SERP interface moderates number of arguments extracted (F-Argument).

  Because of the tentative nature of works examining the effect of SERP interfaces on web search behavior, we propose the following exploratory hypotheses.

  - *H3e:* SERP interface moderates query rate.
  - *H3f:* SERP interface moderates average length of queries.
  - *H3g:* SERP interface moderates the depth to which individuals will click on results in the ranked list.
  - *H3h:* SERP interface moderates the time spent on the SERP on average, at the cost of time spent reading web pages.

  RQ4: How do elements of SERP interfaces impact user experience?
  Various design features of the SERP have been the subject of user studies. Kelly and Azzopardi [42] evaluated the number of search results on the SERP and reported non-significant differences in terms of difficulty and workload. The addition of a knowledge module (area on the SERP containing facts about a named entity) when applicable, as investigated by Arapakis et al. [4], resulted in more satisfaction with the retrieved search results, and was found to be more helpful to users. Regarding user experience, Marcos et al. [51] found that rich snippets experienced increased noticeability in lower-ranked results as opposed to their respective plain snippets which did not matter in higher-ranked positions. Moreover, snippet length has been shown to influence user experience by Maxwell et al. [56]: users preferred longer snippets and felt they were more informative (albeit longer snippets did not result in better performance). It is important to uncover how the SERP interface influences the search experience since search engines causing dissatisfaction among its users probably cannot count on many searchers - regardless of how well it may benefit time-pressured searches. Based on the aforementioned works we conjecture different elements of the SERP impact user experience and propose the following hypothesis.

  - *H4:* Different elements and their presentation on the SERP interface affects user experience.

- **RQ5: To what extent does affinity for technology interaction (ATI) serve as a moderating variable for task performance?**
  Users' search experience is shown to relate to a more critical attitude towards verification strategies and increased probability of clicking lower-ranked results [87], and thus they are expected to be more familiar and efficient in extracting the right and necessary information and gaining knowledge from it. We therefore propose the following four hypotheses.

  – *H5a:* Affinity for Technology Interaction moderates the relationship between time constraints and level of topic focus (T-Depth).

  – *H5b:* Affinity for Technology Interaction moderates the relationship between time constraints and quality of arguments (D-Qual).

  – *H5c:* Affinity for Technology Interaction moderates the relationship between time constraints and interpretation of data into arguments (D-Intrp).

  – *H5d:* Affinity for Technology Interaction moderates the relationship between time constraints and the number of arguments extracted (F-Argument).

1.3. Contributions

With this thesis, we aim to address a knowledge gap in the field of information retrieval. More precisely, the following contributions are made:

1. A comprehensive literature review into the state of the art in the area of web search in information retrieval.
2. A pre-registered 4 × 4 factorial design user study investigating the influence of time constraints and user interfaces on task performance, search results, and user experience[1].
3. Implementation of a search system called *BBTFind* used in this work.
4. Publication of all gathered anonymized data.

1.4. Outline

The remaining part of this thesis is structured as follows. In Chapter 2, related literature from the field of information retrieval is studied to form a solid background and position this work in its context. Next, Chapter 3 introduces the methodology used to conduct the experiments. Chapter 4 outlines the steps that are taken after the experiment was conducted and before the results are presented in Chapter 5. Then, Chapter 6 interprets the results and presents their implications together with limitations encountered during this study. Lastly, Chapter 7 concludes this work and proposes directions for future research.

[1] The time-stamped pre-registration can be found at https://osf.io/25ksb

2 Related Work

This chapter reviews the related literature in the field of information retrieval. Firstly, the web search process and its effects on users are discussed. Then, the means and measures to quantify effects and the topics of search tasks in web search behavior literature are examined. Subsequently, the influence of user interfaces is investigated. The chapter is concluded with a discussion on time in web search.

2.1. Searching the WWW

The web search process has been subject to much research over the last decades. The smallest change in the nuts and bolts of a search engine may significantly influence the resulting beliefs and judgments of its users. Conversely, unconscious underlying assumptions of search engine users may affect interaction with the SERP. This section aims to identify research on the effects of web search and its impact on users in practice.

Perhaps the most well-known inclination in web search is that of users being more likely to pay increased attention to higher-ranked results: position bias or trust bias [40, 51, 65–67]. This is exemplified in early work by O'Brien and Keane [65] who presented participants of a user study with varying interfaces and positions of the most relevant results. They revealed that the most relevant link was clicked 83% of the time if it was the top link in the search entry result page as opposed to 43% when it was the last link, independent of the interface used. While varying source trustworthiness, Kammerer and Gertjets [40] found that in a SERP whose trustworthiness order was reversed on average 1.53 of the most trustworthy search results were selected in contrast to 2.53 results in a SERP with traditional trustworthiness order. A similar result was found for time spent on the most trustworthy web pages in a normal and reversed order, with 129.89 s and 88.52 s being spent on the web pages for the respective orders this lead to users identifying fewer arguments from trustworthy sources in the context of a controversial medical issue. Pan et al. [66] investigated determinants for whether a search result would be viewed or clicked and found that significance, position, and relevance are the most important factors. Albeit participants in the experiment who were presented with a SERP from Google with reversed results were more critical, this was not reflected in the success rate of the tasks the participants were presented with: in the reversed order the success rate was 63% as opposed to 85% in a normal SERP. In a like manner, Pogacar et al. [67] found that the ranking of correct and incorrect information can significantly negatively influence the accuracy of decisions made on the efficacy of medical treatments, exposing a potential source of harm. A follow-up think-aloud study by Ghenai et al. [30] found that the majority view in the SERP, authoritativeness, and quality of sources were the main factors in deciding on the efficacy of health treatments. Efforts to evaluate the effect of snippets on web search behavior by Marcos et al. [51] revealed that the rank of a result on the SERP is more influential than rich snippets such as multimedia snippets, geolocation snippets, or recommendation snippets. Other research such as [16, 36, 82] reports similar findings that result in higher-ranked positions are viewed and clicked more often than lower-ranked results. Models to estimate such position bias can be created, such as the AllPairs model by Agarwal et al. [1] that can be used to control for relevance.

An important aspect of position bias as identified by Yamamoto et al. [87] is attitude towards verification strategies; searchers with a strong attitude towards verification strategies were found to be more likely to click lower-ranked results. As a measure to improve this attitude, Yamamoto and Yamamoto [88] suggest using query priming (use of keywords stimulating critical thinking) in query autocompletion and query suggestion as this resulted in more careful information seeking behavior and increased query rate. The latter effect was also noticed by Medlar et al. [58] using query suggestion (without query priming) in a scientific literature search task, suggesting that, including other search behavior metrics, “query suggestion dramatically impacts user behavior” in their study. An interesting finding regarding search behavior was made by Maxwell et al. [56]: searchers were more likely to click on search results with longer snippets as they felt those to be more informative, however coming at the cost of an increased likelihood of clicking irrelevant items. Another aspect influencing search behavior is stopping. Maxwell and Azzopardi [55] investigated stopping strategies and found that performances of these strategies varied greatly, with simulated users stopping at fixed depth or until a certain amount of non-relevant snippets in a row had been observed performed best overall while the strategy that turned out to be the closest to actual search behavior was a strategy in which users stops after encountering a certain amount of non-relevant snippets in total. These three strategies were later used again by Maxwell and Azzopardi [53] in examining the influence of information scent, the initial impression of a SERP. A SERP level decision point was proposed as an extension to their considered search model that decides whether or not a user abandons the SERP immediately based on the information scent. It was found that including said decision point in their model leads to more realistic modeling of stopping behavior and more effective search given that the user can distinguish between high and low information scent properly. To quantify the effect of, amongst other aspects, query length and relevance feedback, Azzopardi and Zuccon [5] developed numerous user models that take into account the costs and benefits of user interactions.

Another well-investigated bias is confirmation bias: in the process of obtaining information regarding an issue, information representing an opposing opinion is less or not being taken into account [43, 77, 82]. In two experiments in which Schwind et al. [77] presented participants with a list of eight arguments, predominantly preference-consistent arguments were chosen replicating a natural confirmation bias; yet, presenting the participant with preference-inconsistent recommendations weakened the confirmation bias. On the contrary, when Potthirat et al. [69] presented searchers with preference-inconsistent questions in the “People also ask” section of a SERP, a series of questions and their answers related to the query, no significant effect suggesting mitigation of confirmation bias was established. Also, White [82] finds that participants in their study experience confirmation bias as they provide evidence that people mainly look for confirmatory information and barely adapt their view. In the context of media messages, Knobloch-Westerwick and Meng [43] found that participants clicked on average on 1.9 articles that are of preference-consistent nature, whereas on only 1.4 articles were clicked that are preference-inconsistent nature. Adding to that, Pothirattanachaikul et al. [68] found that documents’ opinion and credibility influence search behavior, but this confirmation bias can be reduced by presenting opinion-inconsistent beliefs at higher-ranked positions.

The use of search engines may not only reinforce current opinions or attitudes but may also change them. When a user changes her opinion or attitude due to a set of search results this may be caused by the novel search engine manipulation effect (SEME), a term first coined by Epstein and Robertson [22]. The effect was first investigated in the context of elections in the United States and India and the results emphasized the significance of this bias: biased SERPs may cause voting shifts of undecided voters up to 20% or more and it can be concealed to prevent people from being aware of it [22]. Hence, Epstein and Robertson conclude with the suggestion that SEME can significantly influence elections and has the greatest effect in countries with one predominant search engine. Yet, there are methods to reduce SEME as found by Epstein et al. [23] in follow-up work discussed later. Regarding bias in web search results, Gezici et al. [29] found that the stance of results regarding a controversial topic returned by two major search engines is not favored towards a specific side; rather, the presence of an ideological bias was shown indicating that search engines favor one stance for a certain topic whereas they may favor another stance for another topic. To find out the cause of SEME, Draws et al. [18] investigated the effect of biased search result rankings on debated topics. The evaluation revealed that they were able to replicate SEME with 70% of participants changing attitudes after viewing a biased SERP for which exposure effects are suggested as an aspect influencing this attitude change. That SEME may be an influential aspect is confirmed by McKay et al. [57] who find that active search is an information interaction behavior that was found in almost all interviewed participants who encountered a change of viewpoint. An element to consider in this discussion that likely negatively influences SEME outside of strictly controlled environment studies is the existence of coverage bias as shown by Vaughan and Thelwall [80] who find that search engines with a large market share have the (unintentional) tendency to cover sites from the US better.

Another interesting line of research considers visual and contextual information biases. Novin and Meyers [63] identified four cognitive biases in the participants of their study: (i) priming effects causing a user's attention to visual features upon opening the SERP, (ii) anchoring effects reflecting bias towards the first result, (iii) framing effects arising from users being influenced by how search results are displayed and (iv) availability heuristic favoring most accessible information. Comparable work was done by Liu et al. [50] who looked at the influence of vertical results (snippets with images, videos, news, etc.) in web search, analyzing how users' examination of snippets right above and below a vertical changes due to the presence of a vertical result they found two vertical biases: a cut-off effect meaning that users pay less attention the result below a vertical if it is relevant and a spill-over effect meaning that users pay less attention to the results above but more attention to the results below the vertical if it is irrelevant. Yue et al. [90] take a more general view and looked at presentation bias, concluding that search results are more likely to be clicked on if they had more bolded query terms in the title, even when controlling for position and relevance.

Lastly, some biases in which the human plays a bigger role are discussed. Ieong et al. [35] recognize that click logs examination is a commonly used approach to obtaining human judgments; however, they suffer from domain bias: increased likelihood of users clicking on results from a familiar domain and decreased likelihood of clicking on results from an unknown or untrustworthy domain. This effect was prevalent even when controlling for relevance. Ieong et al. [35] argue that due to domain bias getting stronger, users are visiting fewer domains and, as a consequence, future research will have to control for domain bias. Most of the views presented until now in this section are drawn from studies where the authors looked at only one search query. By taking a different approach and looking at search behavior in the context of a search session (i.e. multiple queries), Zhang et al. [92] were able to unveil two new biases: query bias and duplicate bias. Query bias occurs when the results of a query strike with the user's intention, with query reformulation and no clicked results as a consequence. Duplicate bias implies that a document that was viewed before in an earlier search query is less likely to be viewed again. Regarding query formulation, Roy et al. [73] find that number of queries issued and average document dwell time best predict a user's learning gain. Also, they find that users with some prior knowledge have the largest knowledge gains as opposed to sublinear learning gains for users with little or no prior knowledge.

In summary, we have seen that the effects of position bias and confirmation bias (including related biases) have been investigated and replicated extensively. Research suggests that confirmation bias can be reduced when presenting users with preference-inconsistent views. A relatively limited amount of literature is available on presentation(-related) biases, albeit the existing literature reporting various noteworthy influences. Furthermore, a novel form of bias, SEME, was discovered with the potential of being of vast societal influence.

2.2. Measuring and Quantifying Effects

Being aware of the possible influences of web searches it is important to know how they are investigated and assessed. In this section, the means and measures used in research revolving around web search as well as its target audience are discussed.

There is copious research looking into the web search process. From the effects studied in the previous section, it appears there are roughly two ways in which the experimental procedures to identify these effects can be distinguished:

- *User studies.* Participants are presented with a search task in a user study that can either take place in a lab setting or online in which they complete one or more search tasks subject to an experimental condition. In these experimental conditions, frequent use is made of mock SERPs that are somehow either controlled in terms of order, search entries, or user interface to facilitate the research going on [39, 40, 65, 66, 66, 76]. Typically, the experimental setups in these works have mock SERPs with ads removed, cache web pages to ensure consistency, have participants perform a practice task first, and assign participants to an experimental conditions at random to avoid learning effects [16, 50, 76]. During the experiment, the approach taken to measure the effects of web search can be further subdivided:

  - Effects measured *at a single point in time*. In this approach, participants of user studies are asked to perform the search task after which some measures are taken or evaluated to identify whether a bias effect was present. This is a procedure frequently taken in identifying position bias: Pan et al. [66] log, amongst others, clicks on search entries on the SERP which are analyzed after the experiment to determine the effect of position in the SERP while controlling for relevance. Varying the trustworthiness of sources, Kammerer and Gerjets [40] use a similar procedure in which the search outcome is investigated after the experiment. This approach is not only taken in position bias research [16, 65], but also with vertical bias [50], source bias [63], and social annotations [62].
  - Effects measured *over a period of time*. This approach is often taken to be able to determine whether exposure to an experimental condition resulted in a change in attitude or preference towards a certain subject and therefore requires at least two data points about participants' preferences. This becomes especially clear in works by Epstein and Robertson [22] and Epstein et al. [23] who managed to shift the voting preferences of undecided voters by exploiting SEME. Similar to that, attitude change in debated topics is achieved by Draws et al. [18] in users with mild pre-existing opinions. Yet, SEME is not the only bias applicable to this approach. In the context of confirmation bias, preference of participants with little prior knowledge about the topic of neuro-enhancement could be shifted due to presenting them with preference-inconsistent recommendations [77]. Pre-task and post-task assessments were also used in determining knowledge change by Bhattacharya and Gwizdka [7] and by Gadiraju et al. [28] in the form of knowledge tests. But that is not the only way to measure knowledge change, Zhang and Liu [91] use vocabulary change in pre-task and post-task made mind maps as a measure for knowledge change and knowledge use. Note that a common denominator between these studies is the involvement of prior knowledge or prior opinion on the subject of the task as a restriction or covariate.

- *Click log analyses.* Analyzing click logs from a commercial search engine provides impressions on a larger scale. However, these click log analyses come at the cost of being able to acquire fewer data about the user issuing a specific query. Yet, click log analyses uncover biases not uncovered in user studies. For example, to prove the presence of presentation bias Yue et al. [90] pre-processed search traffic to create a dataset of Fair Pairs [71] and collect human judgments on them to control for position bias. To establish differences in domain expertise and its influences on domain bias, White et al. [83] first had to come up with a method to identify users with a specific topical interest in a subject and separate expert users from non-expert users before a click log analysis could be performed. In [11], clickthrough patterns were examined, but only the first click after the query was issued is considered to capture the users' relative relevance sufficiently well. An advantage of click logs is that it is easier to incorporate complete search sessions instead of only search queries as is done in some research, as exemplified by Zhang et al. [92] revealing query and duplicate bias through developing a task-centric click model. While providing valuable insights, click log analysis requires enormous amounts of data and is therefore reserved for a limited amount of organizations having access to such quantities of data.

Being aware of the general approaches taken to investigate bias in web search behavior, a look at the measures that these biases are quantified in is taken. The most viable technique recognized is web search behavior logging that records interaction with a SERP or web page. For instance, the fraction of total and unique clicks per search result are used in [67] to examine the effect of rank in different experimental conditions. Similarly, [65] looked at the average location of the first click to investigate position bias. Yet, generally, a mixture of search behavior measures is used: [66] logs query reformulations, the number of results clicked, and time spent on search result pages whereas [39] tracks the number and frequency of selected results. Search behavior can even serve as a predictor for user engagement as Zhuang et al. [93] found. Similarly, Arapakis and Leiva [2] used various neural networks to predict user attention to ads from mouse movements. Next to search behavior logging, eye-tracking techniques are also a prevalent method that can provide useful understandings. Liu et al. [50] use metrics such as eye fixation distribution, mean time of arrival at search results, and mean percentage of fixation duration to study the influence of verticals; Marcos et al. [51] study the effect of rich snippets using metrics as fixation duration, fixation count, and visit duration. Albeit being useful, eye-tracking information can only be used in user studies; this information is not available in click log analyses of commercial search engines. Unmistakably, the outcome of the search tasks as is also of immanent importance. Regarding decisions on the efficacy of medical treatments, conclusions on the influence of biased search results can be drawn only after evaluating whether the decisions made by participants of a user study by Pogacar et al. [67] were correct or harmful. When it comes down to the topic of biofuels, Novin and Meyers [63] analyzed how the number and type of arguments as presented in the summaries written during the experiment reflected the results presented in a SERP before drawing conclusions regarding bias effects. Note that search outcome or success cannot only be determined in user studies but also in click log analyses using heuristics; White et al. [83] determine search success based on whether the last event in a session was a click on a URL or a query being issued. Participants of user studies are often presented with a pre-task or post-task questionnaire to study the collected values as mediating or moderating variables or as control and descriptive variables. Standard questions in such questionnaires are related to demographics and web search experience, but they may also include study-specific questions about epistemic beliefs [39], actively open-minded thinking or mood [18], and political ideology [22]. An important realization is that many aspects in information retrieval are interconnected: As Liu et al. [49] have shown, task design influences information-seeking intentions which may, in turn, be reflected in behavioral metrics.

Lastly, the target audience of the experiments in the investigated literature is focused on. As for user studies, two main groups of people participate. Firstly, many works use university students and staff as their target audience [39, 40, 51, 63, 66, 67]: they are predominantly young and have the same level of education. Hence, this is not a representative sample of a population other than an academic one. A noticeable exception to this is [76] whose participants were recruited from a metropolitan area with varying ages and education. Also, generally less than 60 participants are used in the reviewed user studies. A target group that suffers less from these drawbacks but is employed to a much smaller degree is crowdworkers. Research employing crowdworkers tends to have a higher mean age, a broader standard deviation of mean age, and a greater number of participants [18, 22, 23, 35]. Yet, crowdworkers are no one-size-fits-all solution for representative cross-section as most crowdworkers are from the US or India [17]. To overcome this issue, Barbosa and Chen [6] created a framework for a crowdsourcing platform that reduces demographic biases in terms of country of origin, gender, and age at the cost of a 3-5% increase in incorrect responses given to ground-truth questions. As for click log studies, certain studies restrict themselves to queries issued in the English-speaking United States locale [35, 82, 83], but this is not the case for all click log studies.

To sum up, the two most identified means to investigate bias are click log analyses and user studies, where it must be noted the click log analysis is reserved for authors working for companies that have such an abundant amount of data available. Returning measures used to determine the effects of experimental conditions are search behavior analysis and eye-tracking behavior. As for user studies, recurring observations that were made include the use of mock SERPs, target groups consisting mainly of university students, random assignment to an experimental condition, the presence of a training task, and web page caching to ensure consistency.

2.3. Topics

As established in the previous section, research identifying biases often presents users with search tasks to complete as a means to arrive at some conclusions. In this section, a closer look is taken at the characteristics and nature of these search tasks. Also, we take a look at how search results or search outcomes are judged regarding relevance, bias, credibility, or any other aspect that is looked at from an experimental point of view.

The nature of the topics used in search tasks can be roughly distinguished into two categories: controversial and factual, with controversial topics being discussed first. From the literature investigating SEME, it becomes clear immediately that the topics used are of controversial nature. From an empirical perspective, this is intuitive as preferences towards a fact cannot be influenced. Epstein et al. in [22] and [23] try to manipulate the preferences of undecided voters in elections that were expected to be a close call. For both works, independent raters judged whether search results presented on an experimental SERP were biased towards either one of the expected close-call candidates in the election or none. In [18], the authors decided to use the topics of social networks, zoos, cellphone radiation, bottled water, and obesity after crowdworkers were employed to give their opinion on a variety of controversial topics with the aforementioned topics turning out to be the most controversial. Additionally, crowdworkers were used to judge the relevance and viewpoints present in the search results. The non-factual topics of health, politics, finance, environmental sciences, and celebrity news are investigated by Schwarz and Morris [76] asking, for example, questions regarding a diet's effectiveness or mutual funds to invest in. The web pages used in this work were rated by the authors in terms of credibility. A set of more controversial topics was used in [43] who looked at gun ownership, abortion, healthcare regulation, and minimum wage in the context of confirmation bias. As the necessary controversy and political debates exist around these topics, the biases that the nature of these topics induced were judged by participants from an undergraduate university class. The topic used in [63], biofuels, is in line with topics used in the aforementioned works in terms of controversy. Besides these varying topics, there is also a recurring topic: health. Schwind et al. [77] render the topic of neuro-enhancement useful to investigate confirmation bias. They took an uncommon approach: while the topic is complex with participants having little prior knowledge about the topic, the arguments presented in the task were bogus arguments either in favor or against neuro-enhancement, thus not requiring experts to assess the arguments' viewpoints. In [39] and [40], Bechterov's disease and two controversial competing therapies, radon therapy and infliximab theory. are used. In [40], with position bias being investigated, the trustworthiness of the search results was determined in a pilot study with university students as participants. To identify whether a search result was subjective or objective, [39] used a classifier from another work, an approach not identified in any others works.

Next, topics of factual nature are considered. For position bias, generally, an approach with factual questions is used. For example, in a user study by O'Brien and Keane [65] students were presented with computer science related search tasks whose relevance judgments were made by applying rule-based criteria. A similar method is used by Pan et al. [66], who investigate position bias using accessible informational and navigational tasks of everyday topics such as travel and movies; the relevance assessments were performed by independent raters. In the work by Pogacar et al. [67] we see the topic of health returning. It considers the efficacy of medical treatments as the topic and uses efficacy judgments made by authors from another work who performed a systematic review (Cochrane review). Albeit in another context, that of beliefs in search behavior, [82] uses factual medical yes-no questions as their search tasks. They used an unusual way to gather judgments: initially, crowdworkers were tasked with assessing several medical-related micro-tasks; then, to verify their judgments, physicians were asked to perform the same task and found that consensus between the crowdworkers and physicians was good with Fleiss' multi-rater $\kappa \geq 0.853$. During the investigating of Cutrell and Guan [16] on the effect of snippet length, tasks of navigational and informational nature and varying complexity are used that concern everyday topics such as education or movies. Yue et al. [90] investigating presentation bias uses crowdworkers to assess the relevance of URLs to a query; this task was considered difficult by the authors, not because of the topic itself, but because the two URLs presented to choose the most relevant one from would have a similar quality as they would appear adjacent to each other on a SERP. Another seldomly used method for gathering judgments is that as presented by Ieong et al. [35] using heuristics to gather human judgments from the click logs of a search engine as they form a cheap alternative of user feedback.

Altogether, mainly controversial and factual topics are used depending on the search task. As for SEME and confirmation bias (related) literature, topics are mostly of controversial nature to facilitate and measure attitude shifts. As for position bias, mainly primarily topics of factual nature are used. Controversial topics like social networks, environmental science, and politics tend to be of a more complex nature, whereas the factual topics often vary more in complexity. A wide variety of techniques are used to assess search results including independent raters, experts, authors, crowd workers, university students, click logs, and listed criteria; no patterns were identified except the relevance of tasks of medical nature mainly involve medical experts.

2.4. User Interfaces

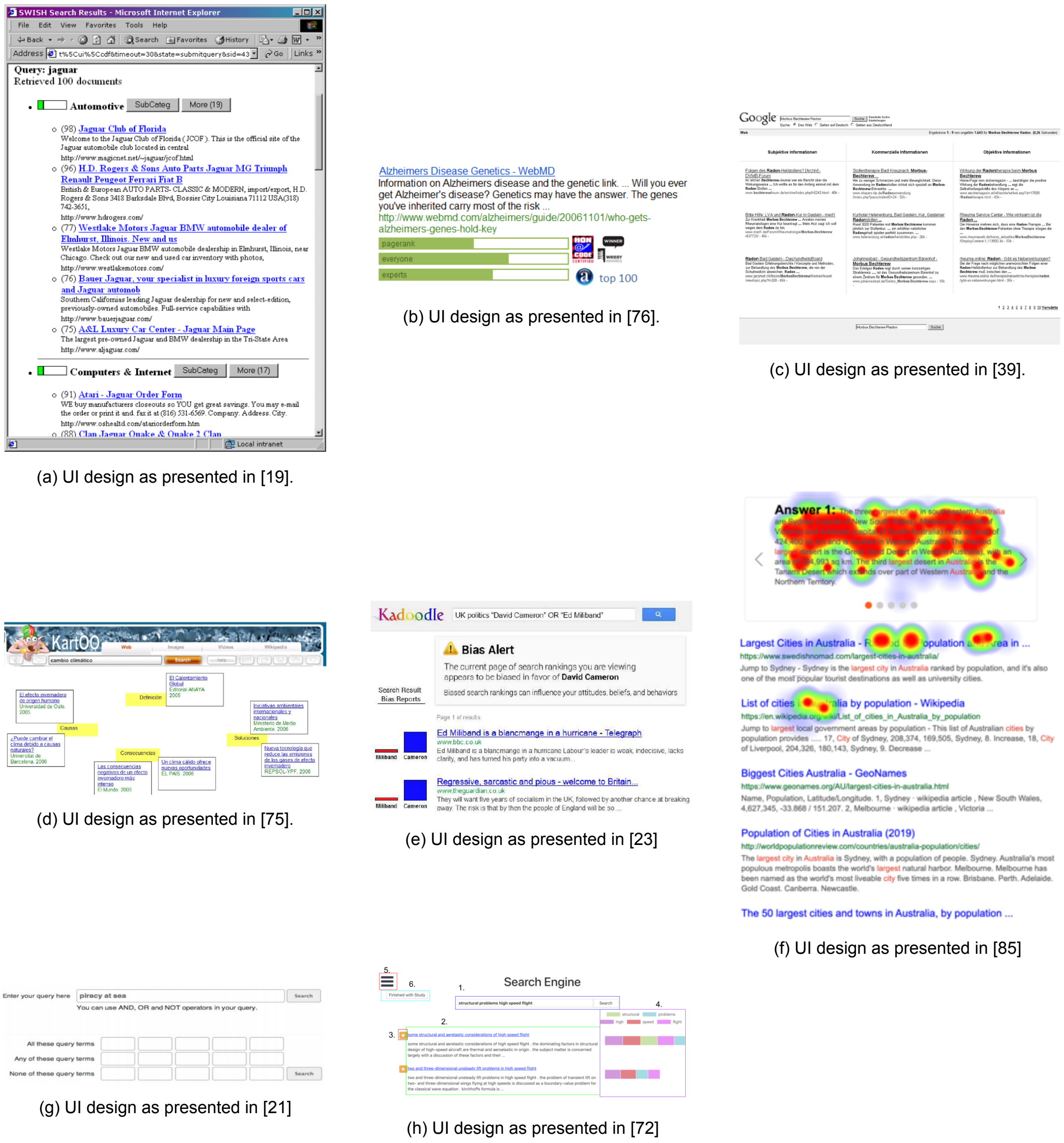
Whereas previous sections have explored literature demonstrating the existence of biases and techniques to assess them, the medium through which the results are presented, the user interface, may also play a role. In this section, existing work regarding user interfaces and web search behavior are reviewed and suggestions from other literature to mitigate bias using the user interface are presented.

To explore the effect of the user interface on search time, Dumais et al. [19] performed a user study in which participants had to complete factual search tasks in different experimental SERP interfaces while search time was tracked. They found that user interfaces with search results grouped by category yielded faster search times than those with list interfaces and that inline summaries are more effective than summaries appearing when hovering over the URL. The interface with the best performance is shown in Figure 2.1a. Schwarz and Morris [76] looked at the addition of contextual information and did not only add contextual information to the SERP, but also to the web page that appears upon clicking on the search result. Information added to the web page includes pagerank, overall and expert popularity, awards won, domain type focus (.gov, .edu, etc.), and temporal and geospatial trends; due to the limited space available on the SERP, search results on the SERP were augmented with only the first three pieces of aforementioned information (Figure 2.1b). The effects are clear: the augmented search results caused a significant improvement in credibility assessments of the search result and increased accuracy to the level obtained when users are viewing an entire page. Kammerer and Gerjets [39] investigated how the user interface may facilitate source evaluations using two interfaces: a list interface and a tabular interface with results sorted into three categories: subjective, objective, and commercial (Figure 2.1c). Behavior in the user interfaces was different: in the tabular interface, less attention was paid to commercial search results and objective results were selected more often. However, for a positive outcome on the search task, the participants also required high epistemic beliefs. In a similar interface, but without the categories, that Kammerer and Gerjets [40] call a grid interface, more results were inspected before the first result was clicked, thereby decreasing the role of search result position in an effort to combat position bias.

In trying to reduce the effects of SEME, Epstein et al. [23] experimented with a low and high bias alert (Figure 2.1e) and found that these bias alerts managed to reduce SEME in vote manipulation. A different design intervention is tested by Salmeron et al. [75] who are interested in identifying how the user interface affects learning. Students were presented with two types of SERPs: a conventional list SERP and a graphical overview SERP providing the relationship between web pages (Figure 2.1d) which had to be studied as if the student was preparing for a test. Afterward, students performed tasks to measure inter-text comprehension which suggested signaling the relationships between web documents resulted in increased inter-text comprehension. Wu et al. [85] investigated how the presence of an answer module on a SERP influences user behavior and found that it reduces search task completion time as well as user effort and increases user engagement. The increase in user engagement grew even further as users were presented with an answer module containing multiple answers (Figure 2.1f). The effects of ads on a SERP was investigated in a user study by Foulds et al. [26], finding that the presence of ads leads to lower recall of relevant concepts and more negative user experience; interestingly, participants in both the experimental condition and the control condition retrieved a similar number of relevant documents, however, taking a significantly longer time with ads present. Next to the way search results are presented on the SERP, the query interface can be altered to. A study by Edwards et al. [21] focused on the query interface instead of the SERP interface and used a conventional query interface and structured interface (Figure 2.1g), finding that participants of the laboratory experiment using the structured interface self-rated their success higher and reported less workload. Surprisingly, no statistically significant differences between stress levels and search behaviors were found. Regarding search behavior, Roy et al. [74] found that the inclusion of active learning tools in the user interface changed several search behavior metrics, but most notably that note-taking increases the number of facts covered in post-task written essays by 34% and highlighting resulted in 34% more subtopics covered.

An approach not yet discussed is taken by Wang et al. [81] recognizing how the use of various types of verticals has changed web search: to optimize whole-page presentation in user satisfaction, a framework that learns the optimal presentation is presented. A scoring function unique to each user calculates an optimal presentation with the results supplied by the search engine, with the authors claiming to outperform contemporary search engines [81].

Next to the aforementioned literature proving through experimentation the effect of a UI on some bias, some unproven yet potentially useful suggestions and their possible implications from research establishing these biases are discussed now. Two similar suggestions for more transparency in SERPs arose. Pan et al. [66] suggest informing users about how search engines crawl results and rank them is coined claiming possible benefits for a limited amount of users. A related suggestion by Novin and Meyers [63] calls for greater transparency in SERPs to improve users' understanding of "the relation-



SWISH Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address t%5Cui%5Ccdf&timeout=30&state=submitquery&sid=43
Go Links
Local intranet

Query: jaguar
Retrieved 100 documents

- **Automotive** SubCateg More (19)
  - (98) Jaguar Club of Florida
    Welcome to the Jaguar Club of Florida (JCOF). This is the official site of the Jaguar automobile club located in central
    http://www.magicnet.net/~jaguar/jcof.html
  - (96) H.D. Rogers & Sons Auto Parts Jaguar MG Triumph Renault Peugeot Ferrari Fiat B
    British & European AUTO PARTS- CLASSIC & MODERN, import/export, H.D. Rogers & Sons 3418 Barksdale Blvd, Bossier City Louisiana 71112 USA(318) 742-3651,
    http://www.hdrogers.com/
  - (77) Westlake Motors Jaguar BMW automobile dealer of Elmhurst, Illinois. New and us
    Westlake Motors Jaguar BMW automobile dealership in Elmhurst, Illinois, near Chicago. Check out our new and used car inventory with photos,
    http://www.westlakemotors.com/
  - (76) Bauer Jaguar, your specialist in luxury foreign sports cars and Jaguar automob
    Southern Californias leading Jaguar dealership for new and select-edition, previously-owned automobiles. Full-service capabilities with
    http://www.bauerjaguar.com/
  - (75) A&L Luxury Car Center - Jaguar Main Page
    The largest pre-owned Jaguar and BMW dealership in the Tri-State Area
    http://www.aljaguar.com/
- **Computers & Internet** SubCateg More (17)
  - (91) Atari - Jaguar Order Form
    WE buy manufacturers closeouts so YOU get great savings. You may e-mail the order or print it and. fax it at (816) 531-6569. Company. Address. City.
    http://www.oshealtd.com/atariorderform.htm
  - (88) Clan Jaguar Quake & Quake 2 Clan

(a) UI design as presented in [19].

Alzheimers Disease Genetics - WebMD
Information on Alzheimers disease and the genetic link. ... Will you ever get Alzheimer's disease? Genetics may have the answer. The genes you've inherited carry most of the risk ...
http://www.webmd.com/alzheimers/guide/20061101/who-gets-alzheimers-genes-hold-key

(b) UI design as presented in [76].

Ergebnisse **1 - 9** von ungefähr **1.643** für **Morbus Bechterew Radon**. (**0,26** Sekunden)

Subjektive Informationen

Folgen des **Radon**-Heilstollens? [Archiv] - DVMB-Forum
Im letzten **Bechterew**-Journal war ein Bericht über die Wirkungsweise ... Ich wollte es für den Anfang einmal mit dem **Radon**-Stollen ...
www.**bechterew**forum.de/archive/index.php/t-6243.html - 48k -

Bitte Hilfe: LVA und **Radon**-Kur in Gastein - med1
Zur Krankheit **Morbus Bechterew** ... Anraten meines Rheumatologen eine Kur beantragt ... Mein Arzt sagt ich soll wegen dem **Radon** da hin.
www.med1.de/Forum/Rheumatologie/**Morbus-Bechterew**/427723/ - 46k -

Radon Bad Gastein - DasXundheitsBoard
Bad Gastein Erfahrungsberichte / Konzepte und Methoden, zur Behandlung des **Morbus Bechterew**, die von der Schulmedizin abweichen. **Radon** ...
www.jacomet.ch/forum/**MorbusBechterew**/themen/board/viewtopic.php?t=208 - 65k -

Kommerzielle Informationen

Stollentherapie Bad Kreuznach: **Morbus-Bechterew** ...
Hin zu weniger Schmerzen und mehr Beweglichkeit. Diese Anwendung im **Radon**stollen richtet sich speziell an **Morbus-Bechterew**-Erkrankte. ...
www.vitaparc-rlp.de/**Radon**anwendung/index.php?pauschalenID=24 - 32k -

Kurhotel Helenenburg, Bad Gastein, Kur, Gasteiner **Radon**stollen ...
Rund 3000 Patienten mit **Morbus Bechterew** kommen jährlich zur Stollenkur. ... ein erhöhter natürlicher **Radon**gehalt spielen perfekt zusammen. ...
www.helenenburg.at/**radon**/heilstollen.php - 26k -

Johannesbad - Gesundheitszentrum Bärenhof - **Morbus Bechterew**
Das Edelgas **Radon** regt durch seinen kurzzeitigen Strahlenreiz .... ist das Gesundheitszentrum Bärenhof zu einem Zentrum für **Morbus Bechterew** geworden. ...
www.johannesbad.de/Seiten_**Morbus-Bechterew**.aspx - 18k -

Objektive Informationen

Wirkung der **Radon**therapie beim **Morbus Bechterew**
Home-Page vom ärztemagazin - ... bestätigen die positive Wirkung der **Radon**behandlung ... regt die Selbstheilungskräfte des Körpers an ...
www.aerztemagazin.at/v4/archiv/artikel.asp?id=17688/**Radon**therapie.html - 43k -

Rheuma Service Center - Wie wirksam ist die **Radon** ...
Die Hinweise mehren sich, dass eine **Radon**-Therapie ... Bei den **Morbus-Bechterew**-Patienten ohne Therapie stiegen die ...
www.rheumawelt.de/home_aktuelles/**MorbusBechterew**/DisplayContent-1,113500.do - 33k -

rheuma-online: **Radon** - Gibt es Nebenwirkungen?
Bei der Frage nach möglichen unerwünschten Folgen einer **Radon**-Heilstollenkur zur Behandlung des **Morbus Bechterew** muß zwischen den ...
www.rheuma-online.de/therapie/natuerliche-therapien/**radon**/gibt-es-nebenwirkungen.html - 35k -

(c) UI design as presented in [39].

El Calentamiento Global
Editorial ANAYA. 2005

El efecto invernadero de origen humano
Universidad de Oslo. 2005

Iniciativas ambientales internacionales y nacionales
Ministerio de Medio Ambiente. 2006

¿Puede cambiar el clima debido a causas naturales?
Universitat de Barcelona. 2006

Nueva tecnología que reduce las emisiones de los gases de efecto invernadero
REPSOL-YPF, 2006

Las consecuencias negativas de un efecto invernadero más intenso
El Mundo. 2005

Un clima cálido ofrece nuevas oportunidades
EL PAÍS. 2006

(d) UI design as presented in [75].

Kadoodle
UK politics "David Cameron" OR "Ed Miliband"
Search Result Bias Reports
Miliband
Cameron
Miliband
Cameron

Bias Alert
The current page of search rankings you are viewing appears to be biased in favor of **David Cameron**
Biased search rankings can influence your attitudes, beliefs, and behaviors

Page 1 of results

Ed Miliband is a blancmange in a hurricane - Telegraph
www.bbc.co.uk
Ed Miliband is a blancmange in a hurricane Labour's leader is weak, indecisive, lacks clarity, and has turned his party into a vacuum...

Regressive, sarcastic and pious - welcome to Britain...
www.theguardian.co.uk
They will want five years of socialism in the UK, followed by another chance at breaking away. The risk is that by then the people of England will be so ...

(e) UI design as presented in [23]

Answer 1: The three largest cities in southeastern Australia are Sydney (capital of New South Wales), Melbourne (capital of Victoria) and Adelaide (capital of South Australia).t has an area of 424,400 sq km and is located in Western Australia. The second largest desert is the Great Sand Desert in Western Australia), with an area of 284,993 sq km. The third largest desert in Australia is the Tanami Desert which extends over part of Western Australia and the Northern Territory.

Largest Cities in Australia - Ranked by Population and Area in ...
https://www.swedishnomad.com/largest-cities-in-australia/
Jump to Sydney - Sydney is the largest city in Australia ranked by population, and it's also one of the most popular tourist destinations as well as university cities.

List of cities in Australia by population - Wikipedia
https://en.wikipedia.org/wiki/List_of_cities_in_Australia_by_population
Jump to largest local government areas by population - This list of Australian cities by population provides ...... 17, City of Sydney, 208,374, 169,505, Sydney, 8. Increase, 18, City of Liverpool, 204,326, 180,143, Sydney, 9. Decrease ...

Biggest Cities Australia - GeoNames
https://www.geonames.org/AU/largest-cities-in-australia.html
Name, Population, Latitude/Longitude. 1, Sydney · wikipedia article , New South Wales, 4,627,345, -33.868 / 151.207. 2, Melbourne · wikipedia article , Victoria ...

Population of Cities in Australia (2019)
http://worldpopulationreview.com/countries/australia-population/cities/
The largest city in Australia is Sydney, with a population of people. Sydney. Australia's most populous metropolis boasts the world's largest natural harbor. Melbourne. Melbourne has been named as the world's most liveable city five times in a row. Brisbane. Perth. Adelaide. Gold Coast. Canberra. Newcastle.

The 50 largest cities and towns in Australia, by population ...

(f) UI design as presented in [85]

Enter your query here: piracy at sea [Search]
You can use AND, OR and NOT operators in your query.

All these query terms: □ □ □ □ □
Any of these query terms: □ □ □ □ □
None of these query terms: □ □ □ □ □ [Search]

(g) UI design as presented in [21]

some structural and aerelastic considerations of high speed flight
some structural and aerelastic considerations of high speed flight . the dominating factors in structural design of high-speed aircraft are thermal and aeroelastic in origin . the subject matter is concerned largely with a discussion of these factors and their ...

two and three-dimensional unsteady lift problems in high speed flight
two and three-dimensional unsteady lift problems in high speed flight . the problem of transient lift on two- and three-dimensional wings flying at high speeds is discussed as a boundary-value problem for the classical wave equation . kirchhoffs formula is ...

(h) UI design as presented in [72]

Figure 2.1: Various user interfaces presented in related work designed to mitigate numerous types of negative effects in web search.

ships between multiple sources"; the idea of search result explanations in SERPs was later investigated by Ramos and Eickhoff [72], finding that a user interface with query term contribution bars per search result (4. in Figure 2.1h) resulted in increased transparency and search efficiency.

Novin and Meyers [63] also propose more explicit identification of conflicting information by exposing the user to other perspectives of a controversial topic. Related to this is the idea suggested by Draws et al. [18] to nudge users to interact with more search results to address SEME. To improve source evaluations in SERPs, Kammerer and Gerjets [40] suggest to "add social information". A work by Muralidharan et al. [62] suggests the use of social annotations in web search is limited as social annotations often go unnoticed, whereas Fernquist and Chi [25] find more promising yet opposing results stating that social annotations are seen 60% of the time when placed on top of the snippet.

All in all, a wide variety of literature has explored alternative user interfaces. Albeit the commonality of several works suggesting to come up with methods that nudge search engine users to explore more search results, the evidence that available literature in this direction presents is thin and would justify further research efforts.

2.5. Time

In our daily lives we all experience time constraints while increasingly turning to the use of search engines. This section investigates the literature revolving around time in web search.

The importance of time in web search is confirmed by Crescenzi et al. [12] finding that participants of their crowdsourced user study who perceived to be under time pressure experienced lower search satisfaction and higher task difficulty. This work was later extended by Crescenzi et al. [14] with a comparable experimental setup, finding numerous significant effects of the time constraints on, amongst others, time pressure, task difficulty, and search performance satisfaction. Also, they looked at the influence of delays in terms of query submit delays and document download delays which caused participants to think the system was slower only when the delays were present in each second task performed. In terms of these delays, Arapakis et al. [3] found that, when adding query submit delays, they are quicker noticed by users of a fast search engine and that the user's belief in the search site helping in completing the search task decreased as the added latency value rose. Maxwell and Azzopardi [52] also experimented with delays and found them to affect the behavior of participants of their laboratory study in terms of time spent within documents when faced with query and document download delays. In a user study by Liu et al. [48] varying the presence of a time constraint, participants in the no time constraint condition self-rated significantly higher pre-search confidence, better post-search performance, higher post-search familiarity with the topic, more gained knowledge, increased estimated required time for the search task and fewer negative moods. No significant difference was found in pre-search and post-search task difficulty. The importance of the effects of time constraints in practice is made clear in a clinical decision-making setting in a study by Van der Vegt et al. [79]: the gained accuracy from being allowed to use a medical search system for clinical decisions reduced from 32% to only 6% as time pressure increased; thus significantly reducing the efficacy of a medical search engine. Opposing to much existing work, Crescenzi et al. [15] state that the traditional notion of an explicitly required amount of search effort is not realistic as it is seldom known in advance. Instead, they gave participants of their user study six tasks making recommendations to a friend in which they had the freedom to decide personally how much information they would include while varying the presence of a 5 minute time constraint per task. Interestingly and opposing to various previous work, no significant differences were found in terms of search behavior or perceived search difficulty between time constraint conditions; as for the recommendations, participants in the time constraint condition made less specific, faster, and shorter recommendations. As for search strategy, Lui and Wei [47] showed that when presented with time constraints, searchers move from an "economic" search strategy to a more "cautious" one which is reflected by the fact that fewer results per query were viewed. Mishra et al. [61] investigated time-critical search and showed per crowdsourced survey that 12% of search engine users who experienced an emergency in the past year turned to the web to look for information, showing that time in web search is an important notion in multiple aspects.

Time in web search is not only relevant in terms of time restrictions, but also in terms of the ranking process of a search engine. Lin et al. [46] propose methods for time-sensitive rankings by determining the focused time for a web page using implicit and explicit temporal expressions present in search results, which is then included in a re-ranking process together with textual relevance to create a time-aware ranking. Arguing that current popular ranking techniques lack a temporal aspect, Yu et al. [89] propose techniques to include the notion of time in academic publication search. Yet, only a relatively limited amount of literature is available on the temporality facet in web search and the effects on the end-user are rarely evaluated. This is recognized by Alonso et al. [41] arguing that “information retrieval applications do not take full advantage of all the temporal information”.

On the whole, this chapter has looked at the effects of web search on users and the experimental techniques used to establish these. Also, the limited amount of literature concerning time and user interfaces would rationalize research endeavors addressing this knowledge gap.

3 Experimental Setup

As the previous chapter concluded, the current body of literature revolving around time and user interfaces in web search allows for dilatation. Therefore, this chapter details the experimental setup investigating how SERP interface and time constraints influence task performance, user behavior, and user experience in a crowd-sourced 4 (SERP interfaces) × 4 (time constraints) between-subjects factorial design study. Approval for this experimental setup has been given by the TU Delft Human Research Ethics Committee (approval no. #1557).

3.1. User Study

Participants of the user study are tasked with identifying arguments in favor or against the topic described below.

Scenario. As for the search task of the user study, participants are asked to imagine they are a journalist working for a newspaper. At the last minute, their boss replaces a colleague reporting on an international discussion forum on DNA cloning. It is the task of participants to gather arguments supporting and opposing the topic in question to prepare themselves for the event. Hence, the search outcome of the task will be a list of arguments both favoring and opposing the topic, which will be used as a measure of task performance as described later.

Topics. Requirements a suitable topic should comply with are:

- Topics should be sufficiently complex: participants should have limited prior knowledge about it to prevent them from drawing up too many arguments beforehand.
- Topics should be debatable and controversial: to allow for finding sufficient arguments it is desirable there is sufficient content regarding the topic on the web.
- Topics should be engaging: to encourage the participants, the topic must be interesting to work on.

A variety of topics meeting these requirements are nuclear energy, genetic modification, DNA cloning, economic monopolies, and weather modification. The topic of DNA cloning was chosen because of its performance in pilot runs. While drawing up the requirements of the topic and designing the search task, desired characteristics of exploratory search tasks by Kules and Capra [44] were adhered to (e.g. search tasks should "Indicate uncertainty, ambiguity in information need and/or need for discovery" or "Suggest a knowledge acquisition, comparison, or discovery task").

Participants. Participants for the user study will be recruited using the online participant recruitment tool Prolific[1]. Participants will be rewarded at a rate of £7.50/h for successfully completing the task. The required sample size is calculated using a power analysis for an ANCOVA using G*Power [24] with an

[1] https://www.prolific.co

effect size f = 0.25 (indicating a moderate effect), significance threshold α = 0.05 / 21 = 0.00238 (due to testing 21 hypotheses), and a statistical power of $(1-\beta)$ = 0.8. The sample size required for each hypothesis was determined using each hypothesis' respective number of groups, degrees of freedom, and covariates resulting in a required sample size of 431 participants. Due to various reasons for exclusions as explained in Section 4.1, 523 participants are recruited. Participants will be prescreened based on three restrictions:

1. Participants must be older than 18 years
2. Participants must be fluent in English.
3. Participants use a desktop to complete the study.

The first two criteria can be assessed using Prolific's screening functionality. A logging framework (see Section 3.2) logs the device used by participants and is used to enforce the third criterion. Participants who do not meet these criteria will not be able to participate in the study. The age restriction is present for legal reasons. The language requirement ensures that participants will be able to understand the nuances of the English language when reading and assessing documents for arguments. The desktop requirement is motivated by the fact that, due to the larger viewport, there is more room to design experimental interfaces. Also, "more leisure content is searched on mobile device, in contrast to more study and work related searches on PC" [45], which suggests that search tasks of the nature used in this experimental setup would predominantly be completed on a desktop.
As the completion time and hence the reward for the study will vary due to different time constraints, a series of four studies with one time constraint per study will be launched on Prolific. Participants who took part in a preceding study using one time constraint will be excluded in studies using subsequent time constraints.

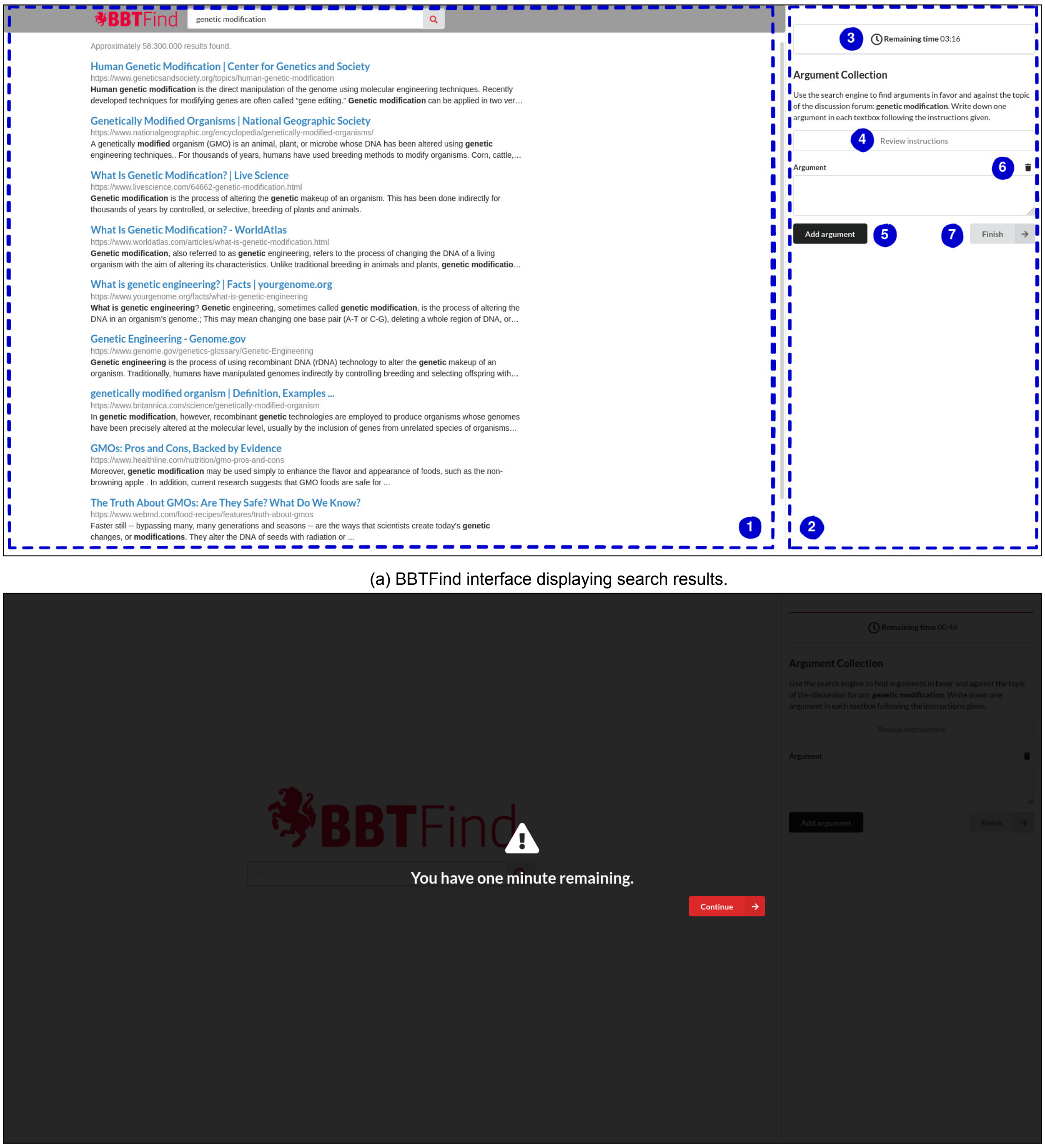
Quality control. Next to the participant restrictions and rewards, other quality control measures will be in place. As Chandler and Kapelner [10] found, output quality increases when crowd workers are aware of what the task will be used for, which is why the true purpose of the task will be outlined to the participants. To further assure quality, participants make use of a training phase in which they are allowed to explore the experimental user interface and are presented with attention checks in the pre-task and post-task questionnaires asking the participants to pick a specific option using a radio button. Due to the cognitive load of the task and to prevent distraction, no attention checks are shown during the search task itself. Of course, the aforementioned measures do not guarantee complete success. Therefore, all arguments submitted by the participants will also be manually checked to identify participants stopping the experiment early, outliers, bogus arguments, empty arguments, arguments submitted more than once, etc.

3.2. Search Platform

This section introduces the search platform and its underlying architecture that is used in the user study.

3.2.1. Interface

During the search task, participants will make use of the custom-designed search platform called *BBTFind*. A screenshot of BBTFind is shown in Figure 3.1a. The circled numbers in this subsection refer to the respective elements of the interface in Figure 3.1a. The interface is divided into the *search frame* ① on the left and the *experiment frame* ② on the right. BBTFind places the search results in the interfaces according to the experimental condition the participant is in, as explained in the following section. A timer ③ shows the remaining time at the top of the experiment frame. An alert in the form of a pop-up will be given when there is one minute remaining (see Figure 3.1b). Also, a button to review the instructions ④ is present in case a participant would like to do so. Furthermore, the goal of the participant's search task is repeated succinctly and participants are given the opportunity to look for arguments favoring or opposing the topic. Text boxes for arguments can be added ⑤ or removed ⑥ using the buttons in the interface. No autocompletion is used to prevent inducing any biases. Participants can interact with the search engine as they normally would on a commercial search engine like *Google*. The experiment frame shows a finish button (7) to allow participants to finish early if they feel like they have collected enough information or further search does not yield additional arguments.



BBTFind genetic modification

Approximately 58.300.000 results found.

Human Genetic Modification | Center for Genetics and Society
https://www.geneticsandsociety.org/topics/human-genetic-modification
Human genetic modification is the direct manipulation of the genome using molecular engineering techniques. Recently developed techniques for modifying genes are often called "gene editing." **Genetic modification** can be applied in two ver...

Genetically Modified Organisms | National Geographic Society
https://www.nationalgeographic.org/encyclopedia/genetically-modified-organisms/
A genetically **modified** organism (GMO) is an animal, plant, or microbe whose DNA has been altered using **genetic** engineering techniques.. For thousands of years, humans have used breeding methods to modify organisms. Corn, cattle,...

What Is Genetic Modification? | Live Science
https://www.livescience.com/64662-genetic-modification.html
Genetic modification is the process of altering the **genetic** makeup of an organism. This has been done indirectly for thousands of years by controlled, or selective, breeding of plants and animals.

What Is Genetic Modification? - WorldAtlas
https://www.worldatlas.com/articles/what-is-genetic-modification.html
Genetic modification, also referred to as **genetic** engineering, refers to the process of changing the DNA of a living organism with the aim of altering its characteristics. Unlike traditional breeding in animals and plants, **genetic modificatio**...

What is genetic engineering? | Facts | yourgenome.org
https://www.yourgenome.org/facts/what-is-genetic-engineering
What is genetic engineering? **Genetic** engineering, sometimes called **genetic modification**, is the process of altering the DNA in an organism's genome.; This may mean changing one base pair (A-T or C-G), deleting a whole region of DNA, or...

Genetic Engineering - Genome.gov
https://www.genome.gov/genetics-glossary/Genetic-Engineering
Genetic engineering is the process of using recombinant DNA (rDNA) technology to alter the **genetic** makeup of an organism. Traditionally, humans have manipulated genomes indirectly by controlling breeding and selecting offspring with...

genetically modified organism | Definition, Examples ...
https://www.britannica.com/science/genetically-modified-organism
In **genetic modification**, however, recombinant **genetic** technologies are employed to produce organisms whose genomes have been precisely altered at the molecular level, usually by the inclusion of genes from unrelated species of organisms...

GMOs: Pros and Cons, Backed by Evidence
https://www.healthline.com/nutrition/gmo-pros-and-cons
Moreover, **genetic modification** may be used simply to enhance the flavor and appearance of foods, such as the non-browning apple . In addition, current research suggests that GMO foods are safe for ...

The Truth About GMOs: Are They Safe? What Do We Know?
https://www.webmd.com/food-recipes/features/truth-about-gmos
Faster still -- bypassing many, many generations and seasons -- are the ways that scientists create today's **genetic** changes, or **modifications**. They alter the DNA of seeds with radiation or ...

Remaining time 03:16

Argument Collection

Use the search engine to find arguments in favor and against the topic of the discussion forum: **genetic modification**. Write down one argument in each textbox following the instructions given.

Argument

(a) BBTFind interface displaying search results.

Remaining time 00:46

Argument Collection

Use the search engine to find arguments in favor and against the topic of the discussion forum: **genetic modification**. Write down one argument in each textbox following the instructions given.

Argument

You have one minute remaining.

(b) BBTFind interface displaying the popup indicating one minute remaining.

Figure 3.1: Screenshots of the BBTFind interface.

3.2.2. Implementation

The actual search is performed using the Bing Web Search API[2]. An overview of the settings in the query parameters used to retrieve results from the Bing search API is found in Table 3.1, whose descriptions are based on the Bing Web Search API Documentation[3]. While making use of the search engine, search behavior is recorded using LogUI [54], a logging framework for web-based experiments. Steps have been taken to prevent participants from copying and pasting arguments directly into the text boxes to encourage active involvement and prevent a lack of compliance with the study instructions. The collected arguments, logged search behavior, and answers to the questionnaires are submitted to

[2] https://www.microsoft.com/en-us/bing/apis/bing-web-search-api
[3] https://docs.microsoft.com/en-us/bing/search-apis/bing-web-search/reference/query-parameters

Parameter	Description / Motivation
`q`	The query as entered by the participant.
`count`	Number of results to return. In the case of BBTFind: 9 results.
`offset`	Number of results to skip, used with count to simulate the effect of pagination. In the case of BBTFind dependent on the requested page.
`textDecorations`	Enables text decorations highlighting important words in the snippet, improving the look of the user interfaces.
`textFormat`	Format of the markers that make up the text decorations, set to HTML.
`mkt`	Market where the results should come from. Set to `en-US` because (i) it ensures consistency among search results and (ii) likely the biggest share of the participants will be from the US.
`responseFilter`	Set to Webpages to prevent any other types of result such videos or news being returned.

Table 3.1: Values of the query parameters used to retrieve search results from the Bing Web Search API.

a back-end that subsequently stores them in a MongoDB database for later statistical analysis. The MongoDB database, LogUI back-end, and the BBTFind back-end are hosted in the SURFsara Research Cloud[4]. The BBTFind back-end built using Node.js exposes its static content and endpoints through express[5]. Please find an architecture overview of the components used in the implementation in Figure 3.2. The code of the BBTFind platform will be published in a project repository hosted by the Open Science Foundation[6]

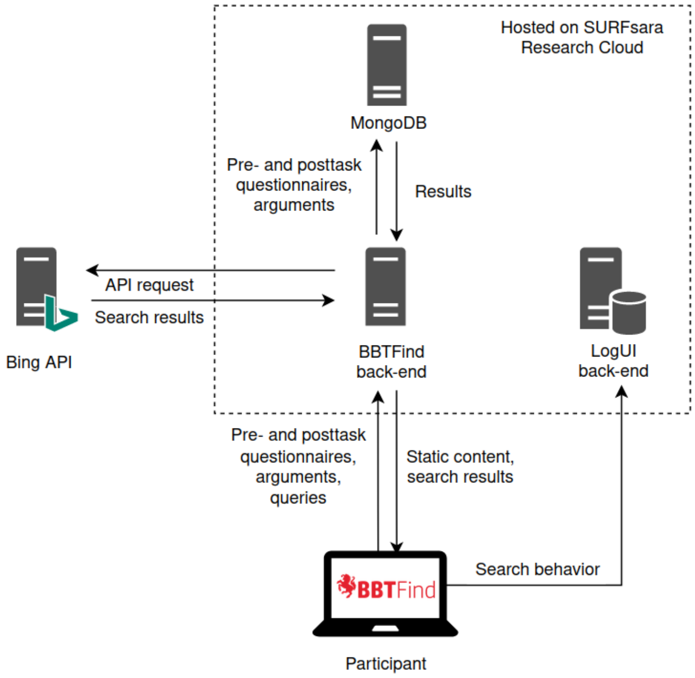

Hosted on SURFsara Research Cloud
MongoDB
Pre- and posttask questionnaires, arguments
Results
API request
Search results
Bing API
BBTFind back-end
LogUI back-end
Pre- and posttask questionnaires, arguments, queries
Static content, search results
Search behavior
BBTFind
Participant

Figure 3.2: Architecture overview of the BBTFind implementation.

[4] https://www.surf.nl/en/surf-research-cloud-collaboration-portal-for-research

[5] https://expressjs.com/

[6] https://osf.io/3wx42/

3.3. Variables

This section outlines the independent, dependent, moderating, and descriptive variables used.

3.3.1. Independent Variables

The independent variables form the experimental conditions of this work and concern the user interface and the time constraint. We now further elaborate on both independent variables.

- SERP interface. Next to each user interface listed below, a reference name to refer to the respective SERP interface in future sections and a figure containing a screenshot can be found. The screenshots of each experimental SERP interfaces are placed in Appendix A due to space considerations. Four user interface variations are used:

  1. List view (see Figure A.1, referred to as `list-view`). Motivation: De facto standard used ubiquitously by search engines ('ten blue links').
  2. Grid view (see Figure A.2, referred to as `grid-view`). Motivation: Kammerer and Gerjets have shown that a grid interface helps reduce the effects of position in a SERP [40] and makes searchers select more trustworthy search results and divides attention to search results more equally [38]. Therefore, it would be interesting to find out whether these effects remain preserved in the time-pressured search task the participants are presented with and have any effect on task performance.
  3. Snippet absence view (see Figure A.3, referred to as `sa-view`). Like the list view, but without a snippet. Motivation: existing literature reports a wide range of effects of snippets. Clarke et al. [11] found that short or missing snippets have a negative impact on click-through rates. Cutrell and Guan [16] found that adding information to snippets in SERPs resulted in increased task performance in informational tasks, a finding not replicated by Maxwell et al. [56] using a different search task, who find that longer snippets are not better in terms of performance. The disagreement justifies the inclusion of this SERP interface in the experimental setup.
  4. Interrupted linear scanning pattern view (see Figure A.4, referred to as `ilsp-view`). SERP interface in which the snippet is placed to the right of other data (cf. [16]). Motivation: while this will interrupt the well-known linear scanning pattern adopted by many searchers [36], this design focuses the users' attention to the metadata of the search results (title and URL) instead of the complete search result including the snippet, as suggested by Cutrell and Guan [16] as a solution to the problem of long snippets being problematic for navigational search tasks.

  Following prior work using grid interfaces [38, 40], the grid interface will show nine search results per SERP. For consistency and fair comparison, all other experimental interfaces will also show nine search results per SERP.

- Time constraint. To identify proper time constraints for this task, inspiration is drawn from similar works with comparable information needs without time constraints. To complete similar tasks, average session lengths in related work are five minutes [28], 8.66 minutes [49], and 5.3 minutes [86]. Therefore, we restrict the available time to two, five, and eight minutes and no time constraint, with the no time constraint acting as a baseline.

Hence, there will be 4 (SERP interfaces) × 4 (time constraints) = 16 experimental conditions. The experimental condition with the list interface and no time constraint is considered the control condition.

3.3.2. Dependent Variables

Three main categories of dependent variables are analyzed:

- Task performance metrics: required for answering **RQ1** and **RQ3**. These metrics aim to assess the quality of the arguments which were identified by the participants. In this assessment, techniques developed by Wilson and Wilson [84] to measure the depth of learning based on Bloom's Taxonomy [8] are used. Albeit being designed for written summaries, we believe these metrics are suitable to evaluate arguments with since they are evaluated ("coded") per fact or sentence, which can also be done in the case of arguments. The measures are adapted to fit in the context of extracted arguments where necessary as outlined below.

1. D-Qual: Quality of the argument. Assessed per argument using Table 3.2 and averaged to create one final value per participant.

Rating	Description
0	Facts within one argument are irrelevant to the subject; facts hold no useful information or advice.
1	Facts are generalized to the overall subject matter; facts hold little useful information or advice.
2	Facts fulfill the required information need and are useful.
3	A level of technical detail is given via at least one key term associated with the technology of the subject; statistics are given.

Table 3.2: Quality of Arguments (D-Qual, adapted from [84]).

2. D-Intrp: Interpretation of data into arguments. Assessed per argument using Table 3.3 and averaged to create one final value per participant.

Rating	Description
0	Facts contained within one argument with no association.
1	Association of two useful or detailed facts: “$A \rightarrow B$”
2	Association of multiple useful or detailed facts: “$A + B \rightarrow C$;” “$A \rightarrow B \rightarrow C$;” “$A \rightarrow B \therefore C$”

Table 3.3: Interpretation of data into arguments (D-Intrp, adapted from [84]).

3. T-Depth: Level of topic focus. T-Depth judges subtopic coverage of the arguments on a scale of 0 to 3. Assessed per argument. Then, the arguments are counted per subtopic and averaged to create one final value per participant. The subtopics used are:
   - Benefits of cloning (e.g. solving fertility issues, prevention of species going extinct)
   - Safety considerations (e.g. low success rates, accuracy)
   - Ethical considerations (e.g. interference with nature, limited genetic uniqueness)
   - Drawbacks of cloning (not being safety consideration or ethical considerations; e.g. costs, probability of faster aging)

   These subcategories have come about by looking at various arguments presented on the web, during which these subcategories were identified as most common themes.
4. F-Argument: Number of arguments. Analogous to F-Fact in [84], but with arguments. Assessed per participant as opposed to the first metrics.

Next to manual metrics, automatic measures of task performance were also considered and sought after. Such measures can be found in the field of computational linguistics. Unfortunately, a search for a suitable metric ended without success for several reasons. Most state-of-the-art papers like [70], [31], and [78] report seemingly promising results, but only publish the data set on which their classifiers are trained, not the classifiers themselves. While completely replicating the classifiers as described in their respective papers is likely possible, this is out of the scope of this thesis. Works like [20] and [32] that do publish their classifiers have shown unsatisfactory performance after a closer thorough inspection (i.e. orders of magnitude difference between learned and original value).

- Search behavior. Required for answering **RQ2** and **RQ3**. Search behavior is logged using the LogUI framework [54]. The considered metrics, chosen because they were most prevalent in related work thus allowing best for (future) comparison of relations, are:
  - Query rate: number of queries issued per minute.

  - Query length: average length of the issued queries in words.
  - Search results: depth to which individuals will click on results in the ranked list.
  - SERP dwell time: time spent on the SERP per minute (i.e., the total time the participant actively focused on the SERP interface in their browser, not reading a search result).

  For exploratory purposes, the initial viewport size and viewport resizing events are also logged. The device type used (desktop, tablet, mobile phone) is logged to ensure people used a desktop to complete the experiment.

- User experience. Required for answering **RQ4**. During the search task, participants are exposed to unseen SERP interfaces. To measure user experience, the User Experience Scale - Short Form (UES-SF) by O'Brien et al. [64] consisting of 12 questions to be rated on a 5-point Likert scale is used. The scale can be further subdivided into four subscales: focused attention, perceived usability, aesthetic appeal, and reward.

3.3.3. Moderator Variable

To understand to what extent affinity for technology interaction (ATI) moderates the relationship between time constraint and task performance, the ATI scale by Franke et al. [27] is used. This is required for **RQ5**. The ATI scale measures to which extent users like to actively approach new technological systems using a questionnaire consisting of 9 items rated on a 6-point Likert scale without subscales.

3.3.4. Descriptive and Exploratory Variables

The following variables are collected or calculated for the purpose of providing descriptive and exploratory analysis:

- Gender. Collected in pre-task questionnaire. Options: Male, Female, Other, Don't want to tell.
- Age. Collected in pre-task questionnaire.
- Highest level of education completed. Collected in pre-task questionnaire. Options: No formal qualifications, secondary education, high school, technical/community college, undergraduate degree, graduate degree, doctorate degree, don't know (conform Prolific's options).
- Web search experience. Collected in pre-task questionnaire. Options: use the web to search more than once per day, once per day, once per three days, once per week, once per month, or less. Participants are requested to pick the option that applies to them the best.
- Prior knowledge. Collected in pre-task questionnaire. Participants are asked to rate their prior knowledge of the subject on a 5-point Likert scale (-2: no prior knowledge, 2: very knowledgeable).
- Topical interest. Collected in pre-task questionnaire. Participants are asked to rate their topical interest in the topic on a 5-point Likert scale (-2: very uninterested, 2: very interested).
- Task definition. Collected in pre-task questionnaire. Participants are asked to rate how clearly the task is defined on a 5-point Likert scale (-2: unclear, 2: very clear).
- Perception of time pressure. Collected in the post-task questionnaire. Participants are asked to rate whether they felt they had enough time to complete the task on a 7-point Likert scale (-3: way too little time, 3: way too much time.
- Initial viewport size and viewport resizing events.

Next to the aforementioned variables, we also consider calculating the following post-hoc descriptive variable. Since it requires the participant's full set of arguments and examined documents it is computed upon the completion of the experiment.

- The extent to which participants relied on a single web page containing readily available arguments. Determined afterward by computing the textual similarity between the search results visited and submitted arguments.

3.4. Procedure

The procedure will consist of several phases (see Figure 3.3 for an overview). Note that once a participant ends up in the introduction phase, he has already undergone Prolific's screening requirements.

Introduction. Participants are welcomed and informed about the purpose of the study. They are informed of how their data will be used and are asked for consent using the informed consent statement in Appendix B. Also, the participants are given the instructions necessary to complete the search task successfully. The scenario, information need, and expected search outcome are introduced to the participants as follows:

Imagine you are a journalist working for a newspaper. A colleague planning to attend and report on an international discussion forum about a controversial topic has called in sick last minute. Being asked by your superior, it is now your responsibility to substitute him at the discussion forum. Unfortunately, you have only very limited time to prepare yourself for the topic you are unfamiliar with. To allow for decent and thorough reporting, you decide to use the limited time available to familiarize yourself with the topic by searching for arguments opposing or supporting the topic using a search engine. You stop your search when you feel like you have collected enough arguments or the time is up.

Training phase. Once aware of all instructions, participants enter the training phase to familiarize themselves with the search interface. This training phase will follow the same experimental condition as the participant is assigned to for the real experiment. The time constraint is not yet enforced, as the goal of the training phase is to facilitate familiarization with the user interface. After all, when performing a time-constrained search that is not under an experimental condition, a participant is likely to turn to a search engine they are already familiar with. The participants are given a sample topic to encourage search behavior and familiarize themselves with the nature of the search task. The result of the training task will not be incorporated in the final result.

Pre-task questionnaire. Participants are presented with the real topic and the pre-task questionnaire with the variables as discussed in section 3.3.4. The answers will be used to verify the absence of confounding effects and for exploratory and descriptive purposes.

Main task. The main task is performed using the BBTFind platform. The arguments are collected using the mock search system according to the experimental condition the participant was assigned to. To ensure that arguments are captured by participants in situ, participants are not allowed to add, alter or remove arguments after the available time has expired. There will be an alert in the form of a popup when there is one minute remaining. Once time is up, participants are taken automatically to the post-task questionnaire.

Post-task questionnaire. Upon (automatic) submission of the main task, the participant is taken to the post-task questionnaire. These contain the questionnaires for assessing the Affinity for Technology Interaction [27] and the User Engagement [64]. Also, the participant will be asked about their perception of time pressure.

Completion. Once the post-task questionnaire has been completed successfully, the participants are thanked for their participation and taken back to Prolific to receive their reward.

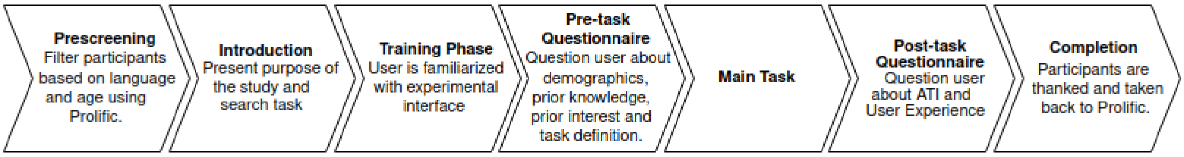

Prescreening
Filter participants based on language and age using Prolific.
Introduction
Present purpose of the study and search task
Training Phase
User is familiarized with experimental interface
Pre-task Questionnaire
Question user about demographics, prior knowledge, prior interest and task definition.
Main Task
Post-task Questionnaire
Question user about ATI and User Experience
Completion
Participants are thanked and taken back to Prolific.

Figure 3.3: Overview of the procedure participants of the study will be subjected to.

To test the experimental setup, a pilot run with 32 participants has been conducted. The results have been used for minor clarifications in the task instructions, testing analysis scripts (see Section 4.2) and approximating the average completion time.

Next to the time taken for the main task as restricted by the time constraint, participants are expected to take approximately 6 minutes to complete the remaining parts of the procedure as derived from the pilot run. This would result in the expected task completion times and corresponding rewards at a rate of £7.50/h as outlined in Table 3.4.

Time constraint	**Expected task completion time**	**Reward**
2 minutes	2 + 6 = 8 minutes	£1.00 ($1.39)
5 minutes	5 + 6 = 11 minutes	£1.38 ($1.92)
8 minutes	8 + 6 = 14 minutes	£1.75 ($2.44)
None (rewarded for 10 minutes based on pilot run)	10 + 6 = 16 minutes	£2.00 ($2.78)

Table 3.4: Expected task completion times and corresponding rewards.

4 Data Preparation

Understanding the experimental setup, we are now ready to run the experiment. Before evaluation of the retrieved submissions, data preparation will need to take place. After all, the collected data will need to be manually inspected for quality control purposes, assessed by humans in terms of task performance, or transformed from raw logs before they form the metrics suitable for analysis. This chapter describes the steps taken before the results are put forward.

4.1. Data Cleaning

Despite all quality control measures, there is no guarantee that nothing will go wrong. Therefore, all submissions are inspected manually as an additional quality control measure. Here, we describe the number of submissions that were excluded and for what reason.

- Failed attention checks, 5 submissions excluded. Submissions were excluded when one attention check was failed and there was at least one other reason to assume unreliability such as no or limited search behavior. No participant failed both attention checks.

- No desktop used, 11 submissions excluded. Despite a requirement for participating being the use of a desktop, 11 participants still used a mobile device to complete the study while participants were clearly informed, on multiple occasions, that a desktop ought to be used.

- Invalid submission, 2 submissions excluded. These participants entered a wrong completion code on Prolific (method used to verify completion of the study), submitted no arguments, and used less than 4 minutes in total to finish their submission.

- Low effort responses, 8 submissions excluded. We individually outline the reasons for exclusion due to a low effort response:

  - No arguments submitted while using only 27 seconds for the main task.
  - Submitting 2 arguments in 9 seconds and using the training topic instead of the main topic in the main task.
  - Using the training topic instead of the main topic in the main task (2 participants).
  - Submitting the nonsensical arguments of “DNA”, “Cloning sheep”, and “Human dna” while using only 44 seconds.
  - Submitting arguments completely unrelated to the main topic (spike in Covid-19 cases due to football matches attracting too many fans)
  - Stating that arguments could not be copied and pasted into the text boxes for arguments while it was specified in the instructions that this was made impossible.
  - Submitting “lack undesirable cara” as the only argument and using 42 seconds.

- Technical issues, 11 submissions excluded. For 9 participants, no search behavior was logged. For 1 participant, no search results were shown according to the user. For 1 participant, no search behavior was logged and no search results were shown according to the participant. Since it is unclear whether the technical issues arose due to the participant or a technical shortcoming in the BBTFind platform, these participants were still rewarded despite their submissions not being usable.

In total, 37 submissions had to be excluded. Due to the relatively large number of exclusions, the number of submissions in 4 experimental conditions fell below the desired number of submissions per experimental condition as calculated by the power analysis. Therefore, additional participants were recruited. In total, 523 participants completed the study. Note that this number of participants is greater than mentioned in the preregistration. With 37 submissions being excluded this resulted in 486 valid submissions.

Manual inspection of the data also revealed partially invalid submissions. To remedy this, the following data editing was done where possible:

- Removed empty arguments from 213 participants. This occurred due to the arguments in the interface being submitted with an empty text box. The submissions are still valid as the remaining arguments are ordinary arguments.
- Removed 20 incomplete arguments. The arguments were removed only when it was the last argument submitted by the participant and no rational meaning could be deduced. This was likely the cause of time being over and participants' arguments being submitted automatically. Removed arguments are: "outcome not fully", "could prevent g", "he", "hel", "Can M", "Is i", "there are three", "Elimina", "DNA cloning is the process of", "Scientists use", "We can m", "Allows you to create exactly", "Cloning could prove hel", "H", "could prevent g", "can cause a further di", "Human clone prevent organ", "he", "helps make cop", "it can be considered u", "It can lead to el", and "some clones".
- Separated arguments from 11 participants who entered multiple arguments in one text box. Several participants entered all arguments in one text box or used one text box for all arguments in favor and one text box for all arguments against the topic. These were separated only when it was clear these meant to be separate arguments through the use of bulleted lists, enumerated lists, commas, or newlines.
- Removed 1 duplicate argument.

Because only arguments have been removed that were evidently beneath contempt, no data has been edited such that data validity is jeopardized; in fact, this data editing improves the quality of the data. Once all the data has been cleaned, it is ready for preprocessing.

4.2. Data Preprocessing

This section discusses the processing required to go from the raw user-submitted data to the desired variables presented in Section 3.3. The required processing per category is as follows:

- Task performance metrics. The assessment of D-Qual, D-Intrp, and T-Depth using the rubrics presented in Section 3.3.2 requires human involvement. Therefore, a random sample of 277 arguments (approximately 10% of all arguments) was assessed by 3 raters (all graduating Computer Science students) familiar with the topic for assessment. This resulted in a Fleiss' κ of 0.748, 0.334, and 0.448 for T-Depth, D-Qual, and D-Intrp respectively. With a coding task of such iterative, subjective nature, a relatively low inter-rater agreement is expected. In contrast, 3 judges in [84] introducing these metrics achieved a Fleiss' κ of 0.64 and 0.58 for D-Qual and D-Intrp respectively using a different experimental setup while going "through three major iterations of refining our measurements" until "'substantial agreement'" was reached.
- The logged search interaction behavior of the participants stored in the back-end requires processing to calculate the metrics per participant.

  - To calculate the query rate, the number of queries a user issued are counted and divided by the time taken to complete the search task.
  - To calculate the average query length, the words in the issued queries are counted and divided by the number of queries issued.
  - To find the deepest rank of the clicked search results, all ranks of the clicked search results need to be examined.
  - To calculate the SERP dwell time per minute, the amount of time spent on the SERP is calculated using the viewport focus change events as logged by the logging framework. Once the total time spent on the SERP is calculated, it is divided by the time taken to complete the search task.

- As for the pre-task and post-task questionnaires, the final values and values of the sub-scales need to be calculated from the responses to the questions that are stored in the database.
- As for reliance on a single web page containing a list of readily available arguments, a bag-of-words model of the textual content of every search result clicked and the arguments submitted are created, subsequently calculating the cosine similarity therebetween. Then, the maximum similarity between any clicked search result and the arguments is taken as the final similarity. BeautifulSoup[1] is used for extracting the text of the clicked search results and scikit-learn [9] is used to create the bag of words model as well as calculating the cosine similarity.
- To visualize initial viewport sizes and resizing events, we made use of Matplotlib [34].

As for the data collected, JSON exports from the MongoDB database containing the submitted arguments and answers to the pre-task and post-task questionnaires and logs of the logging framework containing the search behavior are used as inputs; preprocessed data are outputted as CSV files for convenient importing in the next phase. The dataset, containing raw and processed data, as well as the analysis scripts used to pre-process the data and the implementation of the BBTFind platform will be published in a project repository hosted by the Open Science Foundation[2].

4.3. Hypothesis Testing

Once all the desired data is available, statistical evaluation can begin. The data will be analyzed using SPSS 26. The hypotheses will be evaluated while applying Holm-Bonferroni correction [33], using a target α-value of 0.05. Analyses of Covariance (ANCOVAs) are used for hypothesis testing. When significant results are present, post hoc analyses will be used to find significantly different means. The earlier presented hypothesis will be evaluated as presented in Table 4.1.

[1] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[2] https://osf.io/3wx42/

	H1a-H1d	**H2a-H2d**	**H3a-H3h**	**H4**	**H5a-H5d**
Statistical test	4x one-way ANCOVA	4x one-way ANCOVA	8x two-way ANCOVA	1x one-way ANCOVA	4x one-way ANCOVA
Independent variable(s)	Time constraint	Time constraint	Time constraint & SERP interface	SERP interface	Time constraint
Dependent variable(s)	Task performance metrics	Search behavior metrics	Task performance metrics & search behavior metrics	User experience	Task performance metrics
Covariates	SERP interface, ATI, prior knowledge, topical interest, web search experience, perception of time pressure	SERP interface, prior knowledge, topical interest, web search experience, perception of time pressure	Prior knowledge, topical interest, web search experience, perception of time pressure	Time constraint, prior knowledge, topical interest, web search experience, perception of time pressure	SERP interface, ATI

Table 4.1: Overview of the statistical tests and variables used to test the hypotheses.

5 Results

In this section, descriptive statistics and the results of the hypothesis tests are presented. Also, we discuss exploratory findings.

5.1. Descriptive Statistics

To test the hypotheses posed in the introduction, we use the 486 valid submissions received from the participants. The submissions were divided over the experimental conditions as presented in Table 5.1. Due to a mistake in typing, more participants than required have been recruited in the (grid-view, 5 minutes)-condition. Other than that, the participants are well-balanced over the experimental conditions.

	list-view	grid-view	ilsp-view	sa-view	**Total**
2 min	29	27	30	31	**117**
5 min	49	31	29	29	**138**
8 min	29	28	28	29	**114**
No time constraint	29	29	30	29	**117**
Total	**136**	**115**	**117**	**118**	**486**

Table 5.1: Number of valid submissions per experimental condition.

The gender distribution is somewhat skewed towards males (male: 59.9%, female: 39.3%, other: 0,8%). Regarding education, the highest levels of education completed most prevalent in the participant sample are high school (37.7%), graduate degree (26.7%), and technical/community college (24.5%). See Table 5.2 for all other education levels. 92.2% of the participants indicated using a search engine to search the web at least once per day, thus our participants are practiced searchers. The mean age of the participants was 25.65 years (SD = 9.05) with the youngest participants being 18 and the oldest participant 72. See Figure 5.1 for the age distribution of the participants. Prior knowledge on the topic varied but was moderately low (mean = -0.55, SD = 1.20, scale: [-2, 2]) as desired. Prior interest in the topic was sufficient (mean = 0.42, SD = 1.15, scale: [-2, 2]). The scores for task definition revealed that participants had an adequate understanding of the search task (mean = 1.61, SD = 0.65, scale: [-2, 2]). Based on the prior knowledge and prior interest, we argue that the search task adheres to the desired characteristics of exploratory search tasks [44] well and that participants understood what was asked of them.

Next, we analyze descriptive statistics per dependent variable.

Education	*N*	%
No formal qualifications	2	0.4%
Secondary education	18	3.7%
High school	183	37.7%
Technical/community college	27	5.6%
Undergraduate degree	119	24.5%
Graduate degree	130	26.7%
Doctorate degree	6	1.2%
Don’t know	1	0.2%

Table 5.2: Highest completed education levels of all participants.

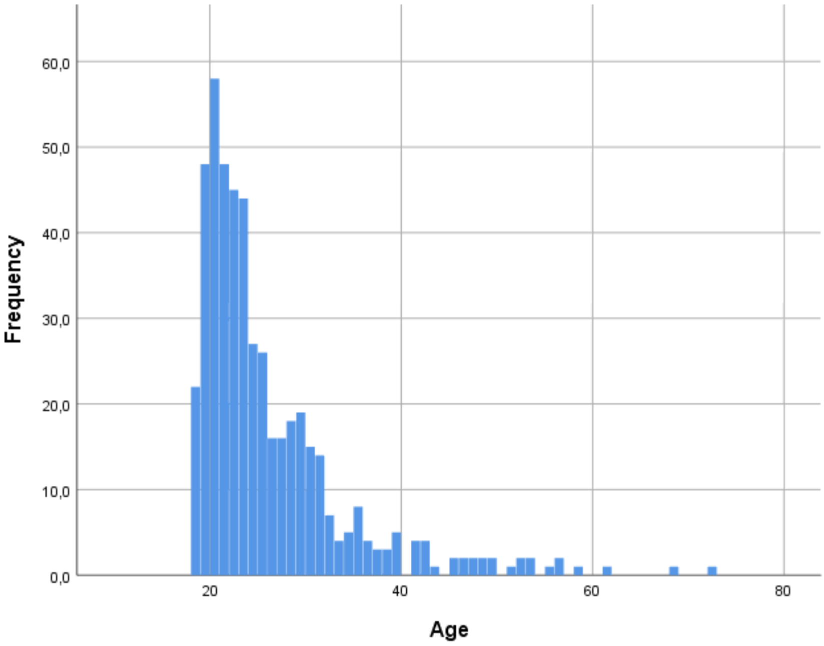


Figure 5.1: Age distribution of the participants.

5.1.1. Task Performance

The participants of our user study submitted 2,763 arguments in total. On average, a participant submitted 5.69 arguments (F-Argument). Five participants submitted 0 arguments (search behavior and questionnaires showed that these participants did take the task seriously) and the greatest number of arguments submitted was 32 arguments by a participant in the no time constraint condition. The submitted arguments were of varying quality. From short and to the point arguments such as

“Help with infertility problems.”

and

“cure for diseases”

to long and detailed arguments such as

“Con: Genetically cloned animals are very likely to either die very soon, even before reaching adulthood or to be born with deficiencies caused by the lack of technology to support DNA cloning, which is still imperfect. Actually, çess (sic) than 10% of the animals born from DNA cloning reach adulthood.”

and

> *“Might help in organ replacement and cure diseases (in this case the clone is not a whole person; there are tecnologies (sic) in which the embryo only grows a specific type of cell. As far as I known (sic) this is a growing field and not perfect at all.”*

The mean argument length is 10.24 words (SD = 9.07) words or 62.36 characters (SD = 54.06). Decomposing the length of arguments over the four time constraints shows that the available time is an important factor in argument length: mean argument length in words of the 2 minutes, 5 minutes, 8 minutes, and no time constraint condition are 6.13 (SD = 5.15), 9.06 (SD = 6.75), 10.35 (SD = 9.09), and 13.20 words (SD = 11.29) respectively.

Argument	**D-Qual**	**D-Intrp**	**T-Depth**
It can save animals from possibly extinction, or even species who were already extinct.	2	0	1
The negative of DNA cloning is that is can lead to in-breeding. This is because the same genotypes are reproducing.	3	1	4
One of the best advantages of DNA cloning is, it helps infertile couples to reproduce	1	0	1
Cons: DNA cloning present a lot of ethical and religious dilemmas	2	0	3
Reproductive cloning is controversial and may cause a lot of problems, since it creates two identical organism.	2	2	3

Table 5.3: Example assessments of D-Qual, D-Intrp, and T-Depth.

The mean quality of the arguments (D-Qual) is 1.56 (SD = 0.75; scale: [0, 3]) and the mean score for interpretation of data into arguments (D-Intrp) is 0.20 (SD = 0.47; scale [0, 2]). As for the level of topic focus (T-Depth), the best-covered subtopic was *benefits of cloning* (45.9%) followed by *ethical considerations* (23.0%), *safety considerations* (11.3%), and *drawbacks of cloning* (8.1%). 322 arguments (11.6%) were not assigned to any subtopic due to being irrelevant to the topic (e.g. arguments such as *“DNA cloning is a molecular biology technique that copies a piece of DNA.”* that contain explanations of the topic). The mean level of topic focus of the participants is 1.07 (SD = 0.67; scale: [0, 3]). Examples of how arguments were assessed in terms of D-Qual, D-Intrp, and T-Depth can be found in Table 5.3. The mean values of the task performance metrics per experimental condition can be found in Table 5.4.

		list-view	grid-view	ilsp-view	sa-view
2 minutes	D-Qual	1.16 (±0.68)	1.36 (±0.75)	1.66 (±0.77)	1.25 (±0.83)
	D-Intrp	0.05 (±0.28)	0.06 (±0.17)	0.18 (±0.36)	0.01 (±0.04)
	T-Depth	0.66 (±0.45)	0.67 (±0.55)	0.61 (±0.44)	0.71 (±0.54)
	F-Argument	3.72 (±2.03)	3.30 (±1.94)	2.80 (±1.73)	3.26 (±1.93)
5 minutes	D-Qual	1.56 (±0.56)	1.41 (±0.51)	1.50 (±0.68)	1.83 (±0.58)
	D-Intrp	0.15 (±0.21)	0.24 (±0.42)	0.17 (±0.23)	0.38 (±0.49)
	T-Depth	1.16 (±0.53)	1.07 (±0.61)	1.03 (±0.51)	0.95 (±0.54)
	F-Argument	6.20 (±3.14)	5.65 (±3.51)	5.24 (±2.39)	4.48 (±3.07)
8 minutes	D-Qual	1.42 (±0.62)	1.46 (±0.56)	1.58 (±0.63)	1.61 (±0.57)
	D-Intrp	0.39 (±0.47)	0.26 (±0.33)	0.27 (±0.33)	0.26 (±0.34)
	T-Depth	1.19 (±0.80)	1.45 (±0.76)	1.21 (±0.74)	1.28 (±0.73)
	F-Argument	7.34 (±4.87)	7.96 (±4.39)	6.50 (±3.36)	6.59 (±3.29)
no time constraint	D-Qual	1.63 (±0.65)	1.64 (±0.49)	1.56 (±0.60)	1.64 (±0.52)
	D-Intrp	0.36 (±0.48)	0.27 (±0.37)	0.46 (±0.57)	0.23 (±0.44)
	T-Depth	1.28 (±0.75)	1.39 (±0.72)	1.05 (±0.53)	1.36 (±0.65)
	F-Argument	7.10 (±6.06)	4.74 (±7.52)	5.53 (±2.46)	7.52 (±3.72)

Table 5.4: Mean values of D-Qual, D-Intrp, T-Depth, and F-Argument per experimental condition.
± indicates standard deviation.

5.1.2. Search Behavior

To find the arguments, participants issued a mean number of 1.98 queries (SD = 1.81). The most issued query is “*DNA cloning*”. Other examples of issued queries include “*dna cloning pros and cons*”, “*DNA cloning advantages*”, and “*dna cloning side effects*”. All participants entered at least 1 query. Outliers were 1 participant entering 20 queries and 1 participant entering 24 queries. Figure 5.2 depicts a histogram showing the distribution of the number of issued queries per participant in total. Taking into account the time that participants used, this resulted in a mean query rate of 0.47 queries issued per minute (SD = 0.41). Regarding query length, the issued queries had a mean length of 3.25 words (SD = 1.21) or 18.51 characters (SD = 6.82). On average, participants visited 2.78 SERPs (SD = 2.67). Regarding the SERPs visited, 27.6% of participants visited a second SERP in their search session. The mean values of the search behavior metrics per experimental condition can be found in Table 5.5.

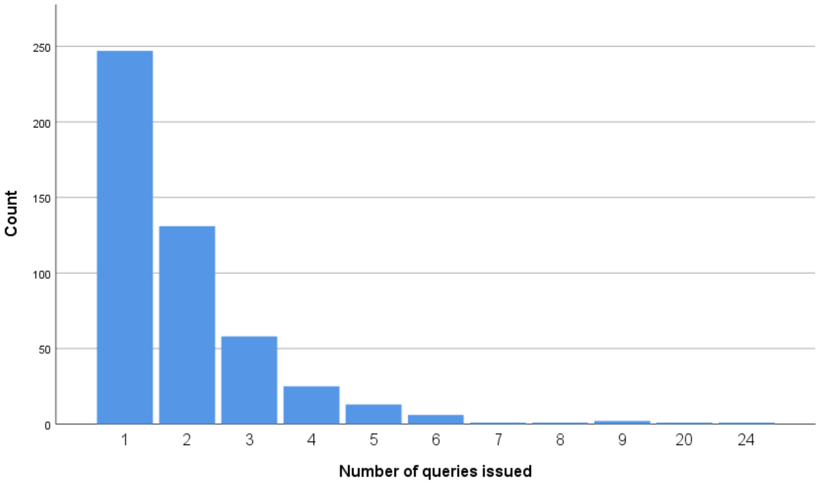


Figure 5.2: Distribution of the number of issued queries per participant.

Due to a technical fault, no clicks and SERP dwell time has been registered for 78 participants (16.05%). To prevent making judgments based on unreliable data, the descriptive statistics presented in the remainder of this subsection only take into account participants for whom SERP dwell time or clicks on search results have been correctly logged with certainty.

To find the results, participants clicked 2.56 links on average (SD = 2.31). Taking into account the issued queries, this comes down to 1.32 clicks per query on average. The average rank of clicked search results was 3.53 (SD = 6.71). Some participants went to great lengths to find relevant search results clicking results in rank 76, 115, and 175; the average deepest result clicked per participant however was 4.20 (SD = 7.13). Figure 5.3 shows the distribution of the ranks of the clicked search results per SERP interface. Only the clicks on the first nine search results are considered in this Figure since every SERP shows nine search results and searchers mostly only inspect the first SERP [16, 36]. Due to the ambiguous ways of interpreting the ranks of the search results in the `grid-view`, statistics regarding the deepest click in the `grid-view` should be interpreted with caution (in the case of BBTFind, the search results were laid out from left to right, then from the top down). Therefore, it is not depicted in Figure 5.3. The click distribution follows an expected distribution reproducing the well-known position bias [36]. Per minute, participants spent on average 40.95 seconds on the SERP (SD = 10.51).

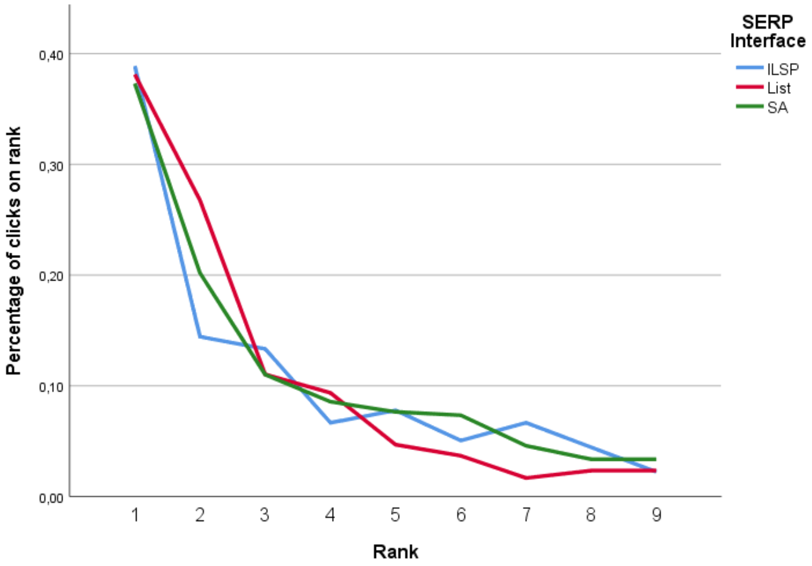


Figure 5.3: Distribution of the ranks of the clicked search results. Only the clicks on the first nine search results are included.

		list-view	grid-view	ilsp-view	sa-view
2 minutes	Query rate	0.78 (±0.56)	0.82 (±0.52)	0.95 (±0.50)	0.67 (±0.29)
	Query length	3.46 (±1.31)	3.22 (±1.03)	3.03 (±0.96)	3.35 (±1.78)
5 minutes	Query rate	0.42 (±0.27)	0.38 (±0.19)	0.49 (±0.24)	0.42 (±0.32)
	Query length	3.57 (±1.16)	3.47 (±1.25)	3.13 (±1.24)	3.00 (±1.03)
8 minutes	Query rate	0.29 (±0.20)	0.23 (±0.11)	0.58 (±0.58)	0.38 (±0.45)
	Query length	3.03 (±1.04)	2.96 (±1.10)	3.44 (±1.58)	3.28 (±1.13)
no time constraint	Query rate	0.30 (±0.28)	0.28 (±0.24)	0.36 (±0.33)	0.24 (±0.14)
	Query length	3.24 (±1.09)	3.14 (±1.26)	3.12 (±1.06)	3.27 (±1.09)

Table 5.5: Mean values for query rate and query length in words per experimental condition. ± indicates standard deviation.

5.1.3. User Experience

The UES-SF [64] questionnaire was used to capture user experience towards the experimental SERP interfaces. Note that all user experience statistics were gathered on a [1, 5]-scale. The mean user experience of all experimental conditions was 3.53 (SD = 0.54). When split out over the SERP interfaces, the traditional `list-view` had the best experience with a user experience of 3.61 (SD = 0.51) followed by the `grid-view` with a user experience of 3.54 (SD = 0.55), the `sa-view` with a user experience of 3.49 (SD = 0.53), and the `ilsp-view` with a user experience of 3.47 (SD = 0.58). Overall, it seems users were moderately satisfied with the experimental SERP interfaces.

The subscales rated the highest are *perceived usability* (mean = 3.94, SD = 0.74), followed by *reward* (mean = 3.71, SD = 0.70), *focused attention* (mean = 3.25, SD = 0.75), and *aesthetic appeal* (mean = 3.24, SD = 0.87). The user experience per experimental condition can be found in Table 5.6.

	list-view	grid-view	ilsp-view	sa-view
2 minutes	3.66 (±0.32)	3.49 (±0.63)	3.47 (±0.69)	3.40 (±0.56)
5 minutes	3.60 (±0.45)	3.51 (±0.59)	3.55 (±0.46)	3.53 (±0.53)
8 minutes	3.73 (±0.51)	3.60 (±0.47)	3.32 (±0.45)	3.66 (±0.49)
No time constraint	3.44 (±0.71)	3.58 (±0.52)	3.54 (±0.67)	3.39 (±0.52)
Total	**3.61 (±0.51)**	**3.54 (±0.55)**	**3.49 (±0.53)**	**3.47 (±0.58)**

Table 5.6: Mean values for user experience per experimental condition. ± indicates standard deviation. All values are on a [1, 5]-scale.

5.2. Hypothesis Tests

In Table 5.7 we report the outcomes of our hypothesis tests. Due to the aforementioned technical fault, we do not evaluate H2c, H2d, H3g, and H3h regarding click behavior and SERP dwell time here; instead, these hypotheses are considered as exploratory findings in Section 5.3.1. Additionally, we also report the Holm-Bonferroni corrected α-value corresponding to the hypothesis (using a target α-value of 0.05) and whether the null hypothesis is rejected. Due to considering 4 hypotheses as exploratory, we correct for 17 hypotheses.

Regarding Research Question 1, a one-way ANCOVA investigating the effect of the time constraints on the level of topic focus (T-Depth) revealed a significant effect ($F(3, 476) = 9.853$, $p < 0.001$, partial $\eta^2 = 0.058$; **H1a**). Post hoc analysis revealed statistically significant differences at the $p < 0.001$ level between the 2 minutes time constraint (mean = 0.766) and 8 minutes time constraint (mean = 1.260). A one-way ANCOVA was run to determine the effect of time constraints on quality of arguments (D-Qual), finding no significant effect ($F(3, 476) = 2.159$, $p = 0.092$, partial $\eta^2 = 0.013$; **H1b**). The effect of time constraints on interpretation of data into arguments (D-Intrp) was found to be statistically significant ($F(3, 476) = 10.46$, $p < 0.001$, partial $\eta^2 = 0.062$; **H1c**) using a one-way ANCOVA with significantly different groups at the $p < 0.001$ level being the 2 minute time constraint (mean = 0.037) in relation to the 8 minute time constraint (mean = 0.301) and the no time constraint condition (mean = 0.374). Also, a statistically significant relationship between time constraints and number of arguments (F-Argument) was found ($F(3, 476) = 11.82$, $p < 0.001$, partial $\eta^2 = 0.069$; **H1d**) with the 2 minute time constraint (mean = 3.924) being statistically significantly different to the 8 minute time constraint (mean = 6.974) and no time constraint condition (mean = 6.337) at the $p < 0.001$ level.

As for Research Question 2, a one-way ANCOVA revealed a statistically signifcant effect of time constraints on query rate ($F(3, 477) = 34.62$, $p < 0.001$, partial $\eta^2 = 0.179$; **H2a**). Significant differences at the $p < 0.001$ level existed between the 2 minute time constraint (mean = 0.829) in relation to the 5 minute time constraint (mean = 0.437), 8 minute time constraint (mean = 0.360), and no time constraint (mean = 0.267) condition. No statistically significant effects of time constraints on average length of queries was found ($F(3, 477) = 0.71$, $p = 0.545$, partial $\eta^2 = 0.004$; **H2b**).

Concerning Research Question 3, using two-way ANCOVAs no statistically significant interaction effect between time constraint and user interface on level of topic focus (T-Depth; $F(9, 466) = 0.648$, $p = 0.756$, partial $\eta^2 = 0.012$; **H3a**), quality of arguments (D-Qual; $F(9, 466) = 1.608$, $p = 0.110$, partial $\eta^2 = 0.030$; **H3b**), interpretation of data into arguments (D-Intrp; $F(9, 466) = 1.653$, $p = 0.098$, partial $\eta^2 = 0.031$; **H3c**), number of arguments (F-Argument; $F(9, 466) = 0.627$, $p = 0.775$, partial $\eta^2 = 0.012$; **H3d**), query rate ($F(9, 466) = 1.268$, $p = 0.252$, partial $\eta^2 = 0.024$; **H3e**), and average length of queries ($F(9, 466) = 0.942$, $p = 0.488$, partial $\eta^2 = 0.018$; **H3f**) was found.

With respect to Research Question 4, a one-way ANCOVA revealed no significant effect of SERP interface on user experience ($F(3, 477) = 1.925$, $p = 0.125$, partial $\eta^2 = 0.012$; **H4**).

In relation to Research Question 5, the covariate in the one-way ANCOVAs, affinity for technology interaction, was statistically insignificant to the level of topic focus (T-Depth; $F(1, 480) = 0.359$, $p = 0.549$, partial $\eta^2 = 0.001$; **H5a**), quality of arguments (D-Qual; $F(1, 480) = 0.682$, $p = 0.409$, partial $\eta^2 = 0.001$; **H5b**), interpretation of data into arguments (D-Intrp; $F(1, 480) = 2.185$, $p = 0.140$, partial $\eta^2 = 0.005$; **H5c**), and number of arguments (F-Argument; $F(3, 480) = 0.095$, $p = 0.758$, partial $\eta^2 = 0.000$; **H5d**).

Thus, the null hypotheses regarding hypotheses **H1a**, **H1c**, **H1d**, and **H2a** are rejected.

No.	Hypothesis	F	p	partial η^2	α	Rejected?
H1a	**Stricter time constraints reduce the level of topic focus (T-Depth)**	9.85	< 0.001	0.058	0.00294	**Yes**
H1b	Stricter time constraints reduce the quality of arguments (D-Qual)	2.16	0.092	0.013	0.00385	No
H1c	**Stricter time constraints reduce interpretation of data into arguments (D-Intrp)**	10.46	< 0.001	0.062	0.00313	**Yes**
H1d	**Stricter time constraints reduce the number of arguments extracted (F-Argument)**	11.82	< 0.001	0.069	0.00333	**Yes**
H2a	**Stricter time constraints increase the query rate (number of queries issued per minute)**	34.62	< 0.001	0.179	0.00357	**Yes**
H2b	Stricter time constraints decrease the average length of queries	0.71	0.545	0.004	0.00100	No
H2c	Stricter time constraints increase the depth to which individuals will click on results in the ranked list	Considered in exploratory findings				
H2d	Stricter time constraints increase the time spent on the SERP on average, at the cost of time spent reading web pages	Considered in exploratory findings				
H3a	SERP interface moderates level of topic focus (T-Depth)	0.65	0.756	0.012	0.01667	No
H3b	SERP interface moderates quality of arguments (D-Qual)	1.61	0.110	0.030	0.00455	No
H3c	SERP interface moderates interpretation of data into arguments (D-Intrp)	1.65	0.098	0.031	0.00417	No
H3d	SERP interface moderates number of arguments extracted (F-Argument)	0.63	0.775	0.012	0.05000	No
H3e	SERP interface moderates query rate.	1.27	0.252	0.024	0.00625	No
H3f	SERP interface moderates average length of queries.	0.94	0.488	0.018	0.00833	No
H3g	SERP interface moderates deepest rank of the clicked search results.	Considered in exploratory findings				
H3h	SERP interface moderates the time spent on the SERP on average, at the cost of time spent reading web pages	Considered in exploratory findings				
H4	Different elements and their presentation on the SERP interface affects user experience.	1.93	0.125	0.012	0.00500	No
H5a	Affinity for Technology Interaction moderates the relationship between time constraints and level of topic focus (T-Depth)	0.359	0.549	0.001	0.00125	No
H5b	Affinity for Technology Interaction moderates the relationship between time constraints and quality of arguments (D-Qual)	0.682	0.409	0.001	0.00714	No
H5c	Affinity for Technology Interaction moderates the relationship between time constraints and interpretation of data into arguments (D-Intrp)	2.185	0.140	0.005	0.00556	No
H5d	Affinity for Technology Interaction moderates the relationship between time constraints and the number of arguments extracted (F-Argument)	0.095	0.758	0.000	0.02500	No

Table 5.7: Outcomes of the conducted ANCOVAs together with the corresponding Holm-Bonferroni corrected α-value and an evaluation of whether null hypothesis is rejected. Statistically significant hypotheses are bolded. For hypotheses H5a-H5d the values corresponding to the covariate (ATI) are reported.

5.3. Exploratory Findings

This section aims to dive deeper into the results and present exploratory findings that may have been of influence in the web search process other than the variables already discussed. The hypotheses in this section were not pre-registered. Therefore, the statistical tests in this section are run with a target α-value of 0.05. Bonferroni correction is applied when testing multiple hypotheses.

5.3.1. Click Behavior and Dwell Time

Due to the aforementioned technical fault occurring in the logging of click behavior and dwell time, we consider these metrics as an exploratory variable. Again, in this subsection, we use only the data of participants for whom dwell time and click behavior has been properly logged. Table 5.8 shows the mean depth to which individuals clicked on results in the ranked list, SERP dwell time per minute, and the number of valid measurements for these measures per experimental condition.

		list-view	grid-view	ilsp-view	sa-view
2 min	Deepest click	2.35 (±2.23)	4.19 (±4.09)	2.43 (±1.99)	3.00 (±2.00)
	SERP dwell time	41.24 (±11.92)	42.95 (±9.63)	47.06 (±6.88)	37.32 (±9.69)
	# measurements	23	21	7	30
5 min	Deepest click	4.35 (±3.29)	8.60 (±10.90)	4.75 (±6.32)	9.79 (±15.63)
	SERP dwell time	39.80 (±8.14)	43.55 (±9.39)	48.68 (±10.10)	36.90 (±10.06)
	# measurements	48	25	20	29
8 min	Deepest click	4.92 (±8.54)	8.85 (±11.47)	5.00 (±6.99)	5.25 (±6.25)
	SERP dwell time	42.47 (±9.63)	40.33 (±9.42)	48.86 (±10.70)	37.47 (±9.72)
	# measurements	26	27	17	28
no time constraint	Deepest click	7.96 (±11.26)	17.76 (±36.94)	7.09 (±12.04)	8.18 (±7.35)
	SERP dwell time	38.06 (±9.69)	40.71 (±11.67)	48.95 (±10.90)	34.22 (±8.52)
	# measurements	27	29	23	28

Table 5.8: Mean values for the rank of the deepest click, SERP dwell time per minute, and the number of valid measurements per experimental condition. ± indicates standard deviation.

Regarding the SERP interface, what stands out in Table 5.8 in terms of SERP dwell time is the difference between sa-view and the other experimental interfaces; dwell time in the sa-view is relatively low. A Kruskal-Wallis H test confirmed that median SERP dwell time was statistically different between SERP interfaces ($H(3) = 56.779$, $p < 0.001$). With statistical significance being accepted at the $p < 0.0083$ level (Bonferroni corrected due to testing 6 hypotheses), statistically significant different medians existed between the sa-view in relation to the grid-view and ilsp-view as well as between the ilsp-view in relation to the list-view and the grid-view. Hence, the SERP dwell time is affected by the SERP interface used. A partial explanation for this phenomenon may be the absence of a snippet in the sa-view, causing people to spend more time on the search results as minimal information can be extracted from the SERP. A Kruskal-Wallis H test ran to determine whether differences in the deepest clicked ranks in the experimental SERP interfaces were significant ($H(3) = 14.097$, $p = 0.003$), but revealed no statistically significant pairwise comparisons.

As for the time constraints, a Kruskal-Wallis H test was run to determine whether a difference in median depth to which an individual will click on results in the ranked list existed and turned out statistically significant ($H(3) = 20.090$, $p < 0.001$). Accepting significance at the $p < 0.0083$ level due to testing multiple hypotheses, post hoc tests revealed statistically significantly different click depths between the 2 minute time constraint and 5 minute time constraint as well as the 2 minute time constraint and no time constraint condition. Hence, click depth is reduced by stricter time constraints. A Kruskal-Wallis H test examining the median differences in SERP dwell time in the time constraint conditions did not result in significance ($H(3) = 1.278$, $p = 0.734$).

Thus, as time constraints are tightened, participants' click depth is reduced. The 2 significant tests regarding SERP dwell time should be interpreted with carefulness as time constraints could not be corrected for while likely playing a role in the analysis.

5.3.2. Early Stoppers

While there was a time constraint present, participants were also given the option to stop their search session early if they felt like they had collected enough information. This subsection aims to analyze early stopping behavior using the time participants had left over. Dividing the time participants had left over in four quartiles, we define a participant to be an *early stopper* if the time left over after completing the submission is in the quartile with the participants having the most time left over. Conversely, we call participants who are not early stoppers *non-early stoppers*. Using this definition, that means a participant is an early stopper when he has more than 27.93 seconds left over. We present the number of participants who stopped early per time constraint in Table 5.9. The no time constraint condition is excluded: due to the absence of a time constraint participants could not have time left over.

Time Constraint		*N*	% of time constraint
2 minutes	Early stopper	3	2.56%
	Non-early stopper	114	97.44%
5 minutes	Early stopper	38	27.54%
	Non-early stopper	100	72.46%
8 minutes	Early stopper	51	44.74%
	Non-early stopper	63	55.26%
Total	Early stopper	92	24.93%
	Non-early stopper	277	75.07%

Table 5.9: Number and percentages of participants stopping early per time constraint.

What can be clearly seen in Table 5.9 is the general pattern of participants stopping early less often in stricter time constraints. Overall, the experimental setup managed to induce time pressure well. Given this notion, we set our sight on task performance, aiming to find out whether it is influenced by early stopping behavior. Table 5.10 presents the mean task performance metrics separated across early and non-early stoppers.

Early stopper?	D-Qual	D-Intrp	T-Depth	F-Argument
Early stopper	1.58 (±0.61)	0.28 (±0.42)	1.00 (±0.62)	5.24 (±3.07)
Non-early stopper	1.48 (±0.66)	0.17 (±0.31)	1.02 (±0.66)	5.38 (±3.57)

Table 5.10: Mean task performance for early and non-early stoppers. ± indicates standard deviation.

A Mann-Whitney U test revealed no differences in T-Depth ($U = 12,436$, $z = -0.140$, $p = 0.888$), D-Qual ($U = 13,766.5$, $z = 1.397$, $p = 0.162$), and F-Argument ($U = 12,642$, $z = 0.097$, $p = 0.923$) between early and non-early stoppers, but did reveal a statistically significant difference with regard to D-Intrp ($U = 14,589$, $z = 2.627$, $p = 0.009$). Thus, except for D-Intrp, the early stopping possibility for participants did not notably influence task performance.

5.3.3. Perception of Time Pressure

To dive deeper into the perception of time pressure the participants had, we asked them to specify to what extent they felt they had enough time to complete the task. Low values indicated way too little time, high values indicated way too much time. Next to the average time pressure experienced by participants in every time constraint as presented in Table 5.11, we also distinguish between early and non-early stoppers. As in the previous section, we do not distinguish between early and non-early stoppers for the no time constraint condition.

Time Constraint		Mean Perception of Time Pressure
2 minutes	Early stopper	-0.67 (±2.08)
	Non-early stopper	-1.96 (±1.23)
	Total	**-1.92 (±1.26)**
5 minutes	Early stopper	1.21 (±1.47)
	Non-early stopper	-0.95 (±1.35)
	Total	**-0.36 (±1.69)**
8 minutes	Early stopper	1.14 (±1.31)
	Non-early stopper	-0.29 (±1.31)
	Total	**0.35 (±1.49)**
no time constraint	**Total**	**1.83 (±1.39)**

Table 5.11: Mean perception of time pressure in the time constraint condition. ± indicates standard deviation.

What can be clearly seen is that, as time constraints tighten, more time pressure is experienced. A one-way ANCOVA with time constraint as independent variable, time pressure as dependent variable, and SERP interface as covariate confirmed these differences in terms of statistical significance ($F(3, 481) = 130.984$, $p < 0.001$, partial $\eta^2 = 0.450$). Accepting significance at the $p < 0.0083$ level due to testing multiple hypotheses, the post hoc analysis revealed that all time constraints differed significantly from each other. Thus, participants' perception of time pressure is explained to a great extent by the time constraint they are in. Also, in the line of expectation, it stands out that early stoppers constantly experienced less time pressure as compared to non-early stoppers. On top of that, in the opportunity for participants to leave any remarks regarding issues occurring during the experiment, various participants left a comment stating that they found two minutes to be too short to make a meaningful submission (e.g. "*2 minutes was a short time to do a fast research and add some arguements (sic).*" and "*I did not have enough time.*").

Furthermore, we were wondering whether any of the experimental SERP interfaces influenced time perception. A one-way ANCOVA with time constraint as independent variable, time pressure as dependent variable, and time constraint as a covariate revealed no significant difference ($F(3, 479) = 0.456$, $p = 0.713$, partial $\eta^2 = 0.003$). Thus, our experimental SERP interfaces were not able to reduce nor impose a perception of time pressure.

5.3.4. Reliance on Single Web Pages

In their search tasks, participants may visit search results that contain readily available arguments in favor or against the topic and use those in their submitted arguments. In this subsection, we investigate whether reliance on single web pages correlates with task performance. After all, relying on a single web page is a little burdensome option; yet, we wonder to what extent such behavior took place and influenced our task. The mean reliance on single web pages per time constraint is presented in Table 5.12.

Time Constraint	Reliance on a Single Web Page
2 minutes	0.17 (±0.18)
5 minutes	0.33 (±0.21)
8 minutes	0.34 (±0.24)
no time constraint	0.42 (±0.24)

Table 5.12: Mean reliance on lists per time constraint condition in terms of cosine similarity. ± indicates standard deviation.

Clearly, reliance on lists rose as participants were given more time. Intuitively, one would expect the opposite: being highly time-pressured provokes an increased reliance on lists containing readily available arguments. We could not come up with any reasonable justifications for this relationship at first glance and, hence, contemplated further analysis. In the search for an explanation, interest in the relationship between reliance on single web pages containing readily available arguments and time pressure was sparked.

The previous section has established that with stricter time constraints comes a greater perception of time pressure. This finding was complemented by a discovery in related work that time constraints may affect users' search confidence and performance [48]. Thus, we looked at the relation between reliance on a single web page and perception of time pressure, with Spearman's correlation coefficient revealing, interestingly, a weak but positive relation therebetween ($r_s(484) = 0.30, p < 0.001$). This suggests that the degree of time pressure experienced by participants may play a minor role in the relationship between time constraints and reliance on a single web page.

5.3.5. Viewport and Viewport Resizing

As an exploratory measure, we captured the size of the viewport and viewport resizing events. The initial viewport sizes of all participants are portrayed in Figure 5.4.

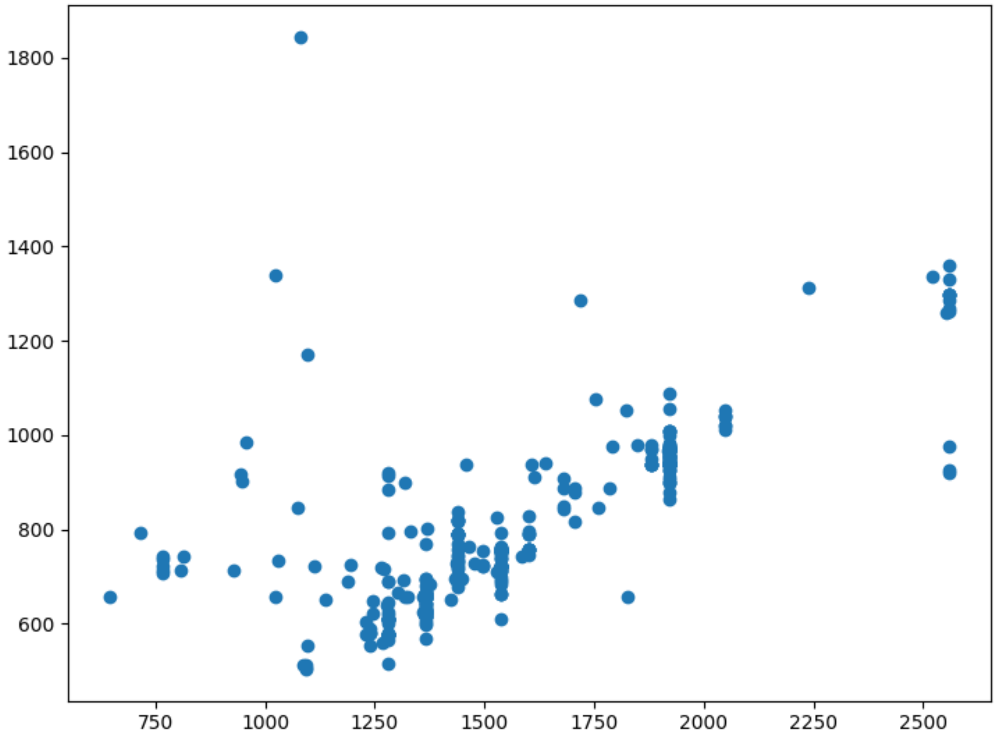


Figure 5.4: Initial viewport sizes of all participants. Every dot represents one initial viewport resolution

The initial viewport sizes reveal that a large share of the participants made use of a monitor wide enough to be able to view the experimental SERP interface optimally. Manual inspection showed that, even using the smallest viewport size logged, no scaling issues appeared; participants have likely scrolled more.

Next to initial viewport sizes, we also consider the viewport resizing events of participants. All viewport resizing events are shown in Figure 5.5.

In total, 38 participants (7.82%) performed at least one resizing event. There appears to be a trend towards scaling the viewport's width. We conjecture this behavior is caused by participants dividing their screen, opening the actual search result next to the experimental SERP interface. Again, even at the smallest resized viewports, no scaling issues arose upon manual inspection.

While there is various prior work looking into the effects of various SERP elements on user experience [4, 21, 26] and search behavior [69, 85], limited works consider the effect of the viewport size on task performance. Therefore, we look at the correlation between the area of the initial viewport (that is, initial width multiplied by the initial height) and the task performance metrics. Spearman's correlation coefficient did not reach significance for a relationship between initial viewport area and T-Depth ($r_s(484) = 0.039$, $p = 0.396$), D-Qual ($r_s(484) = -0.001$, $p = 0.980$), D-Intrp ($r_s(484) = -0.440$, $p = 0.333$), and F-Argument ($r_s(484) = 0.069$, $p = 0.128$). Clearly, the area of the initial viewport is not related to task performance.

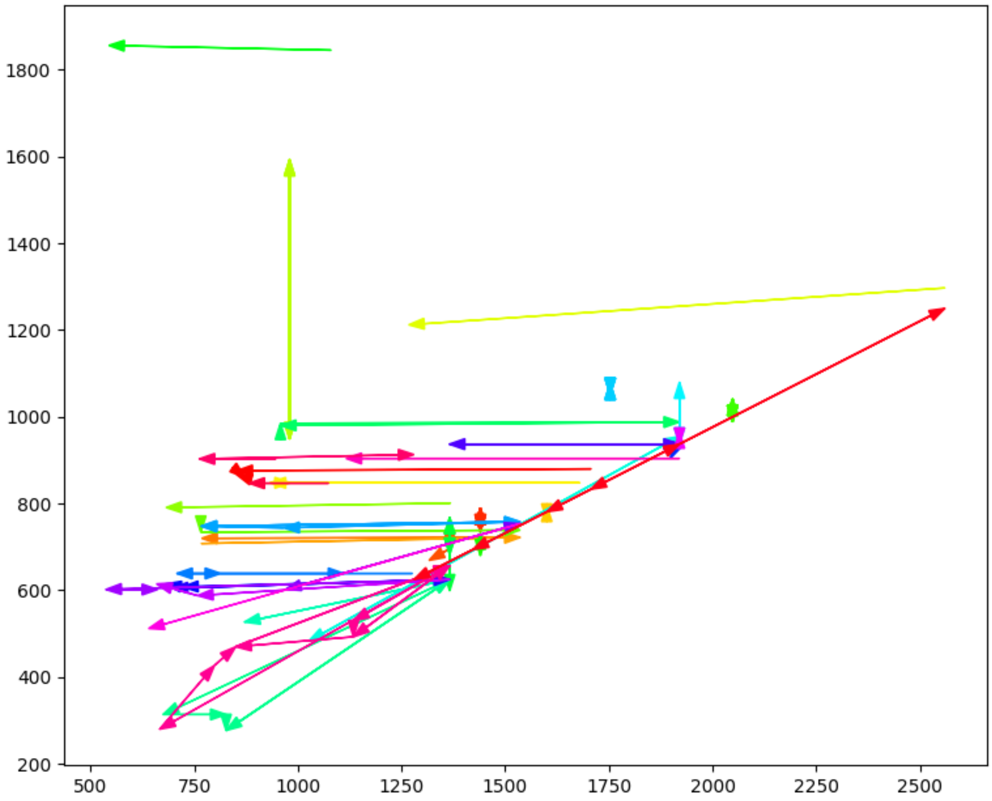


Figure 5.5: Viewport resize traces. Each arrow represents one resizing events, each colored sequence of arrows represent the resizing events by one participant.

5.3.6. Role of Topical Interest

Intuitively, it can be argued that one who has a great topical interest in a particular topic will go to greater lengths in terms of a task related to that topic. Previous work has incorporated topical interests in user modeling of web search behavior to improve the performance of the created models [59, 60]. In this subsection, we review the role of topical interest by looking at the correlation between topical interest and task performance. Accepting significance at the $p < 0.0125$ level due to testing 4 hypotheses, using Spearman's correlation coefficient no significant association between topical interest and T-Depth ($r_s(484) = -0.031$, $p = 0.494$) and F-Argument ($r_s(484) = -0.028$, $p = 0.539$) was found, but statistically significantly very weak associations were found for D-Qual ($r_s(484) = 0.114$, $p = 0.012$) and D-Intrp ($r_s(486) = 0.127$, $p = 0.005$). The weakness of the significant relationships suggests that the role of topical interest in task performance was minimal in our experimental setup.

Therefore, we wonder, despite the minimal overall role of task performance, whether people with a greater topical interest have at least put in more effort in their search task. As a measure of effort, we employ the time used to complete the search task and query rate. Regarding query rate, Spearman's correlation coefficient found topical interest to be very weakly associated with query rate ($r_s(454) = -0.100$, $p = 0.028$). As for time used to complete the search task, a Mann-Whitney U test to determine whether median topical interest in early and non-early stoppers was different failed to reach statistical significance ($U = 14,281.5$, $z = 1.814$, $p = 0.070$). In summary, the exploratory findings do not prove the presence or absence of an indisputable association of topical interest in relation to task performance or effort.

5.3.7. Role of Prior Knowledge

We have already seen that the mean prior knowledge in this task is varying but moderately low. While prior knowledge was included only as a covariate in our hypothesis tests, this subsection focuses on the role of prior knowledge itself during the search task.

Firstly, we delve into how prior knowledge shapes task performance. This question is relevant since related work has shown that prior knowledge of participants has impacted learning gains [28, 73]. Accordingly, we inspect how prior knowledge correlates with task performance using Spearman's correlation coefficient. Accepting significance at the $p < 0.0125$ level due to testing 4 hypotheses, this resulted in no evidence for a relationship between prior knowledge and T-Depth ($r_s(484) = 0.023$, $p = 0.613$), D-Qual ($r_s(484) = 0.090$, $p = 0.048$), D-Intrp ($r_s(484) = 0.096$, $p = 0.035$), and F-Argument ($r_s(484) = 0.011$, $p = 0.815$). Hence, unlike related work, these results could not prove prior knowledge to correlate with any of the task performance metrics.

Secondly, we consider the consequences of prior knowledge on user experience. Instinctively, having a lot of prior knowledge would reduce the burden of finding the required knowledge for completing the search task. To validate this idea, we use Spearman's correlation coefficient to test the relationship between prior knowledge and user experience, revealing a very weak positive but significant effect ($r_s(484) = 0.11$, $p = 0.01$). Hence, prior knowledge is to some extent associated with user experience.

5.3.8. Role of Education

Until now, education has only been used as a descriptive variable in previous analyses. This subsection therefore seeks to make concrete the role of education in this work. Due to the limited number of participants in other educational levels, we only consider participants with *high school*, *undergraduate degree*, and *graduate degree* as the highest completed educational level in the analyses in this subsection.

We begin with the relation between education and task performance. Previous work has suggested educational background affects verification attitudes [87], further triggering interest in this relationship. We analyze the relationship between education and task performance using Kruskal-Wallis H tests. This analysis revealed no statistically significantly differences in terms of T-Depth ($H(2) = 1.337$, $p = 0.513$), D-Qual ($H(2) = 5.001$, $p = 0.082$), D-Intrp ($H(2) = 3.390$, $p = 0.184$), and F-Argument ($H(2) = 0.059$, $p = 0.971$) for all considered educational levels. Interestingly, despite the intuitive expectation that the varying educational levels would explain variance in terms of task performance, this was not the case in this work.

Furthermore, we were interested in the relation between education and ATI. While a correlation between educational level and affinity for technology interaction may seem obvious; we found no evidence for a statistically significant correlation between education and ATI using Spearman's correlation coefficient ($r_s(430) = -0.07$, $p = 0.13$) in line with [27].

5.3.9. Queries

Since query behavior has been considered as a variable in an abundant number of works regarding web search (e.g. [68, 88]), this subsection aims to further explore how query behavior played a role in our work.

Beginning with a broad perspective, we examine closer the relation between average query length and task performance. Accepting significance at the $p < 0.0125$ level due to testing 4 hypotheses, evidence was found in favor of a relationship between average query length and all task performance metrics, T-Depth ($r_s(484) = 0.363$, $p < 0.001$), D-Qual ($r_s(484) = 0.152$, $p = 0.001$), D-Intrp ($r_s(484) = -0.114$, $p = 0.012$), and F-Argument ($r_s(484) = 0.206$, $p < 0.001$) using Spearman's correlation coefficient. As participants progress through a search task, they will gain knowledge [28, 73]. Considering only participants entering at least two queries, this seems to be reflected in terms of the mean length in words of the first and last query, 2.62 (SD = 1.13) and 3.52 (SD = 1.02) words respectively. As a result of this finding, we wonder whether perhaps the length of the first and last query correlates with task performance and use Spearman's correlation coefficient to test this notion. The results are shown in Table 5.13.

Query		$r_s(237)$	p
First query	D-Qual	0.110	0.090
	D-Intrp	0.008	0.899
	T-Depth	0.122	0.060
	F-Argument	0.079	0.222
Last query	D-Qual	0.051	0.429
	D-Intrp	-0.077	0.234
	T-Depth	0.070	0.283
	F-Argument	0.020	0.758

Table 5.13: Spearman's correlation coefficients between the length of the first and last query and task performance. Only participants entering at least two queries were included in the analyses.

No statistically significant relationships have been found between the length of the first and last query in relation to any of the task performance metrics. Thus, while the average query length is correlated with task performance, this could not be related to a learning effect during the task as existed in other related works.

Also, query behavior is known to correlate with various aspects of user experience (e.g. mood [86] and user engagement [93]). For that reason, we are curious whether such a relationship also exists in our sample. However, we found no evidence for a relationship between average query length and user experience using Spearman's correlation coefficient ($r_s(237) = -0.10, p = 0.11$). In light of the previous paragraph, no statistically significant association was found between user experience and length of the first query ($r_s(237) = 0.002$, $p = 0.98$) or length of the last query ($r_s(237) = 0.49$, $p = 0.45$). Hence, opposing to other work, we were not able to establish a relationship between querying behavior and user engagement.

6 Discussion

This chapter discusses the main findings of our experiment and reviews the limitations that have emerged throughout this study.

6.1. Interpretation and Implications

We studied how time constraints and SERP interfaces influence web search behavior, task performance, and user experience; and in what way different SERP interfaces are susceptible to the effects of time constraints. To that extent, we conducted a 4 × 4 crowd-sourced user study.

Regarding the effects of time constraints on task performance, as considered in **RQ1**, we found that stricter time constraints reduced level of topic focus (T-Depth, **H1a**), interpretation of data into arguments (D-Intrp, **H1c**), and the number of arguments extracted (F-Argument, **H1d**). These findings are in line with similar works reporting reduced task performance related to the presence of a time constraint [13]. Moreover, in line with [79], we find significant differences in task performance between different lengths of time constraints; though, it must be mentioned that significant differences arising from post hoc analyses were mainly present in the two most extreme time constraints conditions. Nevertheless, these results show that, as time constraints tighten, search engines are decreasingly efficient in assisting the user in terms of task performance and imply that the effects of time constraints should be considered in the design of search systems.

As for the impact of time constraints on search behavior, as regarded in **RQ2**, the results indicate that participants in stricter time constraints issued queries at a higher rate (**H2a**). Although the increased query rate fits with findings in previous research [13, 15, 47], other influences in behavioral metrics could not be established.

Considering the susceptibility of SERP interfaces to the effects of time constraints, as regarded in **RQ3**, no SERP interface has been found to interact with task performance metrics or web search behavior. While being unaware of related literature examining this interaction, previous work has shown that SERP interfaces impact task performance [16, 37, 38] and search behavior [11, 39, 40]. Anyhow, while this work may not have found significant effects, it has laid out the foundation for future work into how search result presentation may assist time-constrained searchers.

The influence of SERP interfaces on user experiences was considered in **RQ4**. Contrary to the hypothesis, the SERP interface did not affect user experience in this study. This is an interesting finding since various elements and their presentation on the SERP have been found to impact user experience related measures such as informativeness [55], satisfaction [4], and difficulty [42]. However, the absence of statistically significant results is not inherently bad: it implies the experimental SERP interfaces used in this work neither improved nor worsened user experience.

Lastly, **RQ5** questioned the extent to which ATI moderates the relationship between time constraints and task performance. No significant relations have been found contrary to our conjecture based on correlations of ATI with characteristics such as technology usage and learning success [27], indicating no proof that ATI serves as a moderating variable for task performance.

Exploratory findings suggest that the increased query rate as time constraints increase comes at the cost of reduced depth to which searchers click in the search results, an effect that should be taken into account in the development of search systems to better support searchers performing exploratory search tasks. Another interesting exploratory finding was made in terms of early stopping behavior of participants. Participants making use of the option to stop their search early had only one significant difference in terms of task performance; apparently illustrating that participants may well indicate themselves when there is little or nothing more to learn.

6.2. Limitations

Like many studies, this study has some limitations despite its careful preparations and attention devoted to its design.

Firstly, the generalizability of this study is limited by the fact that only one topic was used in the search task. This has perhaps best expressed itself in variables such as topical interest and prior knowledge collected in the pre-task questionnaire that are greatly specific to the topic. Given the constraint in terms of funding, a trade-off between sample size and experimental conditions was to be made. Reducing the number of experimental conditions would diminish the benefit of this work, whereas reducing the sample size per condition would decrease statistical power. In the end, the decision to design the study as is was based on the identified knowledge gaps in related literature, maximizing the potentially added value. To minimize the impact of using one topic as well as possible, the variables specific to the topic have been included as covariates in the hypothesis tests.

Secondly, a caveat must be placed on task performance assessments. Although the task performance metrics by Wilson and Wilson were designed “using grounded theory” [84], subjectivity is on the lurk in the application of these rubrics used to judge task performance. Alternatives such as automatic measures were examined, but manual inspection has shown unsatisfactory performance. As a second alternative objective measure such as argument length were considered, but these are easily confounded by factors like time constraint. A more qualitative measure analyzing whether a deeper understanding of the topic was attained was preferred over a quantitative measure, which is why ultimately Wilson and Wilson’s task performance metrics were used. To give meaning to the degree of subjectivity present in the assessments, inter-rater reliability statistics on a random sample of arguments that has been evaluated by three raters are presented.

Lastly, as a more technical limitation, using a proxy to serve search results such that all participants would see exactly the same page would be ideal. However, using a proxy would increase response times, possibly affecting user experience. Also, no existing solutions were satisfactory in terms of cost or ease-of-use. Hence, the presence of a proxy did not outweigh its drawbacks. Presumably, the absence of a proxy mainly manifested itself in personalization differences in terms of ads; content-wise, most important to this work, no significant changes due to the absence of proxy are expected. As an alternative, server-side web page rendering solutions like Prerenderer[1] and Rendertron[2] were shortly considered, but unfortunately did not suffice either as these remove all JavaScript, with the undesired consequence that (for example) cookie consent popups cannot be closed.

[1] https://prerender.io/

[2] https://github.com/GoogleChrome/rendertron

7 Conclusion and Future Work

This chapter presents the conclusion and identifies directions for future work.

7.1. Conclusion

In this work, a user study into the effects of time constraints and SERP interfaces on task performance, user behavior, and user experience was presented. Participants were tasked with finding arguments in favor or against a controversial topic under a time constraint while using a mock search system. To evaluate task performance, we used qualitative measures to examine whether participants have acquired a deeper understanding of the topic. The results of the user study show that stricter time constraints reduce task performance. Additionally, tighter time constraints affect web search behavior in terms of increased query rate. An exploratory finding suggests this comes at the cost of click depth, which increases as time constraints loosen. Using various SERP interfaces, the susceptibility to the effects of time constraints was investigated. While no susceptibility has been found, a foundation for future work in this research direction has been laid out. Also, the various SERP interfaces used have not influenced user experience, implying that user experience was neither improved nor worsened. We did not find affinity for technology interaction to serve as a moderating variable in the relationship between time constraints and task performance. Exploratory findings have shown that participants who stopped the task early because they believed they had collected enough arguments had comparable task performance scores to those who used all the time available. Our results have strengthened the existing literature revolving around time constraints and SERP interfaces and made contributions to the knowledge gap on the interplay between time constraints and SERP interfaces.

7.2. Future Work

The following directions for future studies are presented based on this work.

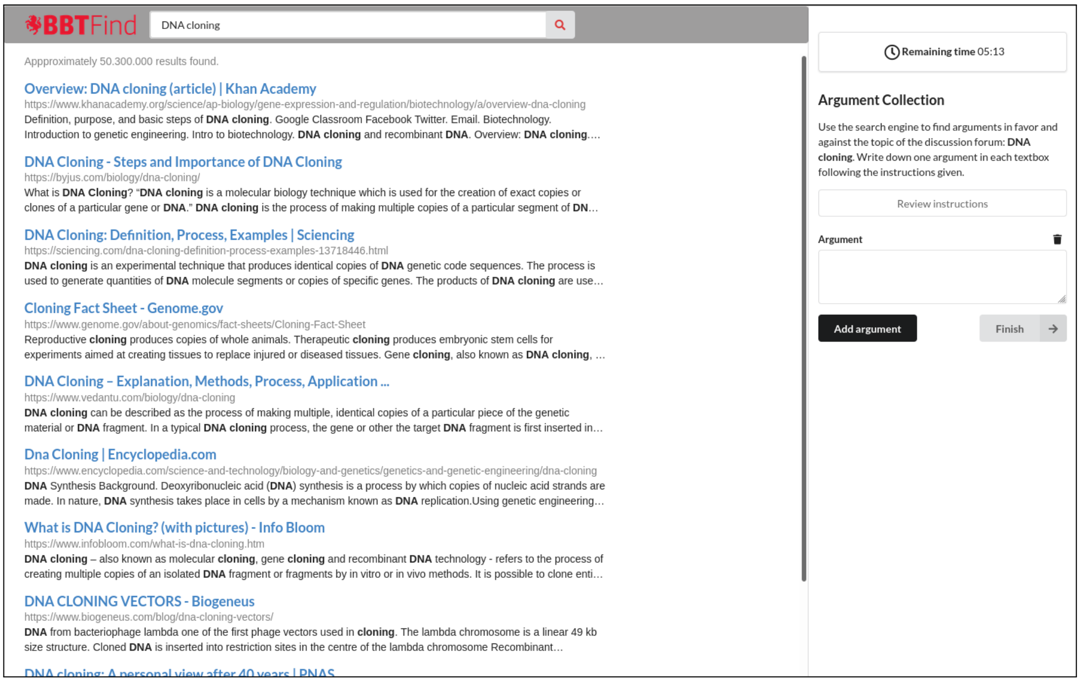
- Further research is needed to establish the generalizability of this work. As was already acknowledged as a limitation, the fact that only one topic was used is a shortcoming of this research. Therefore, future work could confirm and generalize the findings presented in this work. The experimental setup used in this work may serve as a foundation for this future work.

- The absence of effects on user experience as a result of SERP interfaces used in this work warrants further investigation. Future work is required to determine whether this effect is also found in other settings and whether different elements of the SERP interface can be used in support of searchers in other directions of the information retrieval field.

- Additionally, considering that differences in task performance were found mainly in the most extreme time constraints, further research efforts could be devoted to examining the sensitivity between time constraints and performance (related) measures.

A

Experimental Interfaces



DNA cloning

Appproximately 50.300.000 results found.

Overview: DNA cloning (article) | Khan Academy
https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/overview-dna-cloning
Definition, purpose, and basic steps of **DNA cloning**. Google Classroom Facebook Twitter. Email. Biotechnology. Introduction to genetic engineering. Intro to biotechnology. **DNA cloning** and recombinant **DNA**. Overview: **DNA cloning**....

DNA Cloning - Steps and Importance of DNA Cloning
https://byjus.com/biology/dna-cloning/
What is **DNA Cloning**? "**DNA cloning** is a molecular biology technique which is used for the creation of exact copies or clones of a particular gene or **DNA**." **DNA cloning** is the process of making multiple copies of a particular segment of **DN**…

DNA Cloning: Definition, Process, Examples | Sciencing
https://sciencing.com/dna-cloning-definition-process-examples-13718446.html
DNA cloning is an experimental technique that produces identical copies of **DNA** genetic code sequences. The process is used to generate quantities of **DNA** molecule segments or copies of specific genes. The products of **DNA cloning** are use…

Cloning Fact Sheet - Genome.gov
https://www.genome.gov/about-genomics/fact-sheets/Cloning-Fact-Sheet
Reproductive **cloning** produces copies of whole animals. Therapeutic **cloning** produces embryonic stem cells for experiments aimed at creating tissues to replace injured or diseased tissues. Gene **cloning**, also known as **DNA cloning**, …

DNA Cloning – Explanation, Methods, Process, Application ...
https://www.vedantu.com/biology/dna-cloning
DNA cloning can be described as the process of making multiple, identical copies of a particular piece of the genetic material or **DNA** fragment. In a typical **DNA cloning** process, the gene or other the target **DNA** fragment is first inserted in…

Dna Cloning | Encyclopedia.com
https://www.encyclopedia.com/science-and-technology/biology-and-genetics/genetics-and-genetic-engineering/dna-cloning
DNA Synthesis Background. Deoxyribonucleic acid (**DNA**) synthesis is a process by which copies of nucleic acid strands are made. In nature, **DNA** synthesis takes place in cells by a mechanism known as **DNA** replication.Using genetic engineering…

What is DNA Cloning? (with pictures) - Info Bloom
https://www.infobloom.com/what-is-dna-cloning.htm
DNA cloning – also known as molecular **cloning**, gene **cloning** and recombinant **DNA** technology - refers to the process of creating multiple copies of an isolated **DNA** fragment or fragments by in vitro or in vivo methods. It is possible to clone enti…

DNA CLONING VECTORS - Biogeneus
https://www.biogeneus.com/blog/dna-cloning-vectors/
DNA from bacteriophage lambda one of the first phage vectors used in **cloning**. The lambda chromosome is a linear 49 kb size structure. Cloned **DNA** is inserted into restriction sites in the centre of the lambda chromosome Recombinant…

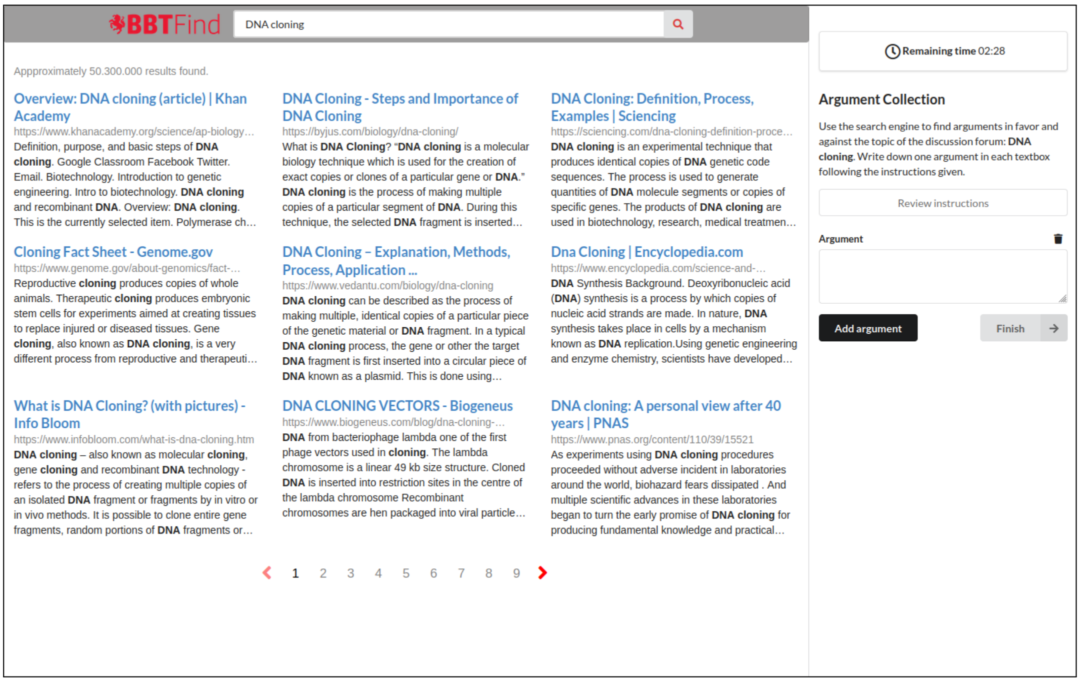
DNA cloning: A personal view after 40 years | PNAS

Argument Collection

Use the search engine to find arguments in favor and against the topic of the discussion forum: **DNA cloning.** Write down one argument in each textbox following the instructions given.

Argument

Figure A.1: SERP interface showing the list view.



Appproximately 50.300.000 results found.

Overview: DNA cloning (article) | Khan Academy
https://www.khanacademy.org/science/ap-biology…
Definition, purpose, and basic steps of **DNA cloning**. Google Classroom Facebook Twitter. Email. Biotechnology. Introduction to genetic engineering. Intro to biotechnology. **DNA cloning** and recombinant **DNA**. Overview: **DNA cloning**. This is the currently selected item. Polymerase ch…

DNA Cloning - Steps and Importance of DNA Cloning
https://byjus.com/biology/dna-cloning/
What is **DNA Cloning**? "**DNA cloning** is a molecular biology technique which is used for the creation of exact copies or clones of a particular gene or **DNA**." **DNA cloning** is the process of making multiple copies of a particular segment of **DNA**. During this technique, the selected **DNA** fragment is inserted…

DNA Cloning: Definition, Process, Examples | Sciencing
https://sciencing.com/dna-cloning-definition-proce…
DNA cloning is an experimental technique that produces identical copies of **DNA** genetic code sequences. The process is used to generate quantities of **DNA** molecule segments or copies of specific genes. The products of **DNA cloning** are used in biotechnology, research, medical treatmen…

Cloning Fact Sheet - Genome.gov
https://www.genome.gov/about-genomics/fact-…
Reproductive **cloning** produces copies of whole animals. Therapeutic **cloning** produces embryonic stem cells for experiments aimed at creating tissues to replace injured or diseased tissues. Gene **cloning**, also known as **DNA cloning**, is a very different process from reproductive and therapeuti…

DNA Cloning – Explanation, Methods, Process, Application ...
https://www.vedantu.com/biology/dna-cloning
DNA cloning can be described as the process of making multiple, identical copies of a particular piece of the genetic material or **DNA** fragment. In a typical **DNA cloning** process, the gene or other the target **DNA** fragment is first inserted into a circular piece of **DNA** known as a plasmid. This is done using…

Dna Cloning | Encyclopedia.com
https://www.encyclopedia.com/science-and-…
DNA Synthesis Background. Deoxyribonucleic acid (**DNA**) synthesis is a process by which copies of nucleic acid strands are made. In nature, **DNA** synthesis takes place in cells by a mechanism known as **DNA** replication.Using genetic engineering and enzyme chemistry, scientists have developed…

What is DNA Cloning? (with pictures) - Info Bloom
https://www.infobloom.com/what-is-dna-cloning.htm
DNA cloning – also known as molecular **cloning**, gene **cloning** and recombinant **DNA** technology - refers to the process of creating multiple copies of an isolated **DNA** fragment or fragments by in vitro or in vivo methods. It is possible to clone entire gene fragments, random portions of **DNA** fragments or…

DNA CLONING VECTORS - Biogeneus
https://www.biogeneus.com/blog/dna-cloning-…
DNA from bacteriophage lambda one of the first phage vectors used in **cloning**. The lambda chromosome is a linear 49 kb size structure. Cloned **DNA** is inserted into restriction sites in the centre of the lambda chromosome Recombinant chromosomes are hen packaged into viral particle…

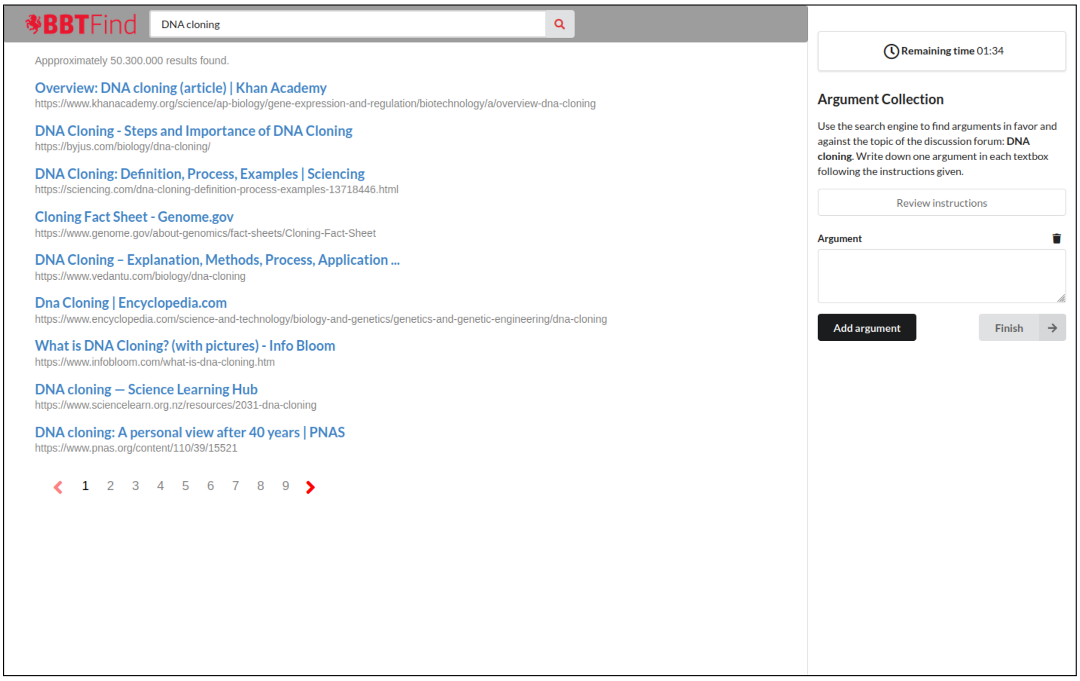
DNA cloning: A personal view after 40 years | PNAS
https://www.pnas.org/content/110/39/15521
As experiments using **DNA cloning** procedures proceeded without adverse incident in laboratories around the world, biohazard fears dissipated . And multiple scientific advances in these laboratories began to turn the early promise of **DNA cloning** for producing fundamental knowledge and practical…

Remaining time 02:28

Argument Collection

Use the search engine to find arguments in favor and against the topic of the discussion forum: **DNA cloning**. Write down one argument in each textbox following the instructions given.

Figure A.2: SERP interface showing the grid view.



Appproximately 50.300.000 results found.

Overview: DNA cloning (article) | Khan Academy
https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/overview-dna-cloning

DNA Cloning - Steps and Importance of DNA Cloning
https://byjus.com/biology/dna-cloning/

DNA Cloning: Definition, Process, Examples | Sciencing
https://sciencing.com/dna-cloning-definition-process-examples-13718446.html

Cloning Fact Sheet - Genome.gov
https://www.genome.gov/about-genomics/fact-sheets/Cloning-Fact-Sheet

DNA Cloning – Explanation, Methods, Process, Application ...
https://www.vedantu.com/biology/dna-cloning

Dna Cloning | Encyclopedia.com
https://www.encyclopedia.com/science-and-technology/biology-and-genetics/genetics-and-genetic-engineering/dna-cloning

What is DNA Cloning? (with pictures) - Info Bloom
https://www.infobloom.com/what-is-dna-cloning.htm

DNA cloning — Science Learning Hub
https://www.sciencelearn.org.nz/resources/2031-dna-cloning

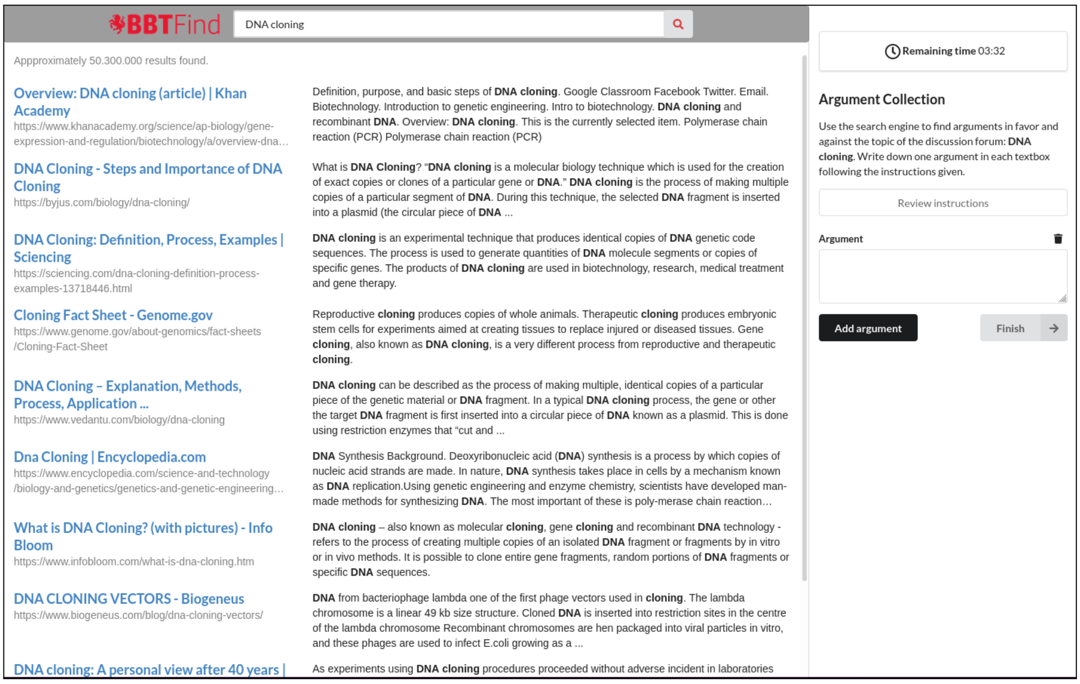
DNA cloning: A personal view after 40 years | PNAS
https://www.pnas.org/content/110/39/15521

Remaining time 01:34

Argument Collection

Use the search engine to find arguments in favor and against the topic of the discussion forum: **DNA cloning**. Write down one argument in each textbox following the instructions given.

Figure A.3: SERP interface showing the snippet absence view.



DNA cloning

Appproximately 50.300.000 results found.

Overview: DNA cloning (article) | Khan Academy
https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/overview-dna…

Definition, purpose, and basic steps of **DNA cloning**. Google Classroom Facebook Twitter. Email. Biotechnology. Introduction to genetic engineering. Intro to biotechnology. **DNA cloning** and recombinant **DNA**. Overview: **DNA cloning**. This is the currently selected item. Polymerase chain reaction (PCR) Polymerase chain reaction (PCR)

DNA Cloning - Steps and Importance of DNA Cloning
https://byjus.com/biology/dna-cloning/

What is **DNA Cloning**? "**DNA cloning** is a molecular biology technique which is used for the creation of exact copies or clones of a particular gene or **DNA**." **DNA cloning** is the process of making multiple copies of a particular segment of **DNA**. During this technique, the selected **DNA** fragment is inserted into a plasmid (the circular piece of **DNA** ...

DNA Cloning: Definition, Process, Examples | Sciencing
https://sciencing.com/dna-cloning-definition-process-examples-13718446.html

DNA cloning is an experimental technique that produces identical copies of **DNA** genetic code sequences. The process is used to generate quantities of **DNA** molecule segments or copies of specific genes. The products of **DNA cloning** are used in biotechnology, research, medical treatment and gene therapy.

Cloning Fact Sheet - Genome.gov
https://www.genome.gov/about-genomics/fact-sheets/Cloning-Fact-Sheet

Reproductive **cloning** produces copies of whole animals. Therapeutic **cloning** produces embryonic stem cells for experiments aimed at creating tissues to replace injured or diseased tissues. Gene **cloning**, also known as **DNA cloning**, is a very different process from reproductive and therapeutic **cloning**.

DNA Cloning – Explanation, Methods, Process, Application ...
https://www.vedantu.com/biology/dna-cloning

DNA cloning can be described as the process of making multiple, identical copies of a particular piece of the genetic material or **DNA** fragment. In a typical **DNA cloning** process, the gene or other the target **DNA** fragment is first inserted into a circular piece of **DNA** known as a plasmid. This is done using restriction enzymes that "cut and ...

Dna Cloning | Encyclopedia.com
https://www.encyclopedia.com/science-and-technology/biology-and-genetics/genetics-and-genetic-engineering…

DNA Synthesis Background. Deoxyribonucleic acid (**DNA**) synthesis is a process by which copies of nucleic acid strands are made. In nature, **DNA** synthesis takes place in cells by a mechanism known as **DNA** replication.Using genetic engineering and enzyme chemistry, scientists have developed man-made methods for synthesizing **DNA**. The most important of these is poly-merase chain reaction…

What is DNA Cloning? (with pictures) - Info Bloom
https://www.infobloom.com/what-is-dna-cloning.htm

DNA cloning – also known as molecular **cloning**, gene **cloning** and recombinant **DNA** technology - refers to the process of creating multiple copies of an isolated **DNA** fragment or fragments by in vitro or in vivo methods. It is possible to clone entire gene fragments, random portions of **DNA** fragments or specific **DNA** sequences.

DNA CLONING VECTORS - Biogeneus
https://www.biogeneus.com/blog/dna-cloning-vectors/

DNA from bacteriophage lambda one of the first phage vectors used in **cloning**. The lambda chromosome is a linear 49 kb size structure. Cloned **DNA** is inserted into restriction sites in the centre of the lambda chromosome Recombinant chromosomes are hen packaged into viral particles in vitro, and these phages are used to infect E.coli growing as a ...

DNA cloning: A personal view after 40 years |

As experiments using **DNA cloning** procedures proceeded without adverse incident in laboratories

Remaining time 03:32

Argument Collection

Use the search engine to find arguments in favor and against the topic of the discussion forum: **DNA cloning**. Write down one argument in each textbox following the instructions given.

Figure A.4: SERP interface showing the interrupted linear scanning pattern view.

B
Informed Consent Statement

You are being invited to participate in a research study titled Study about Web Search Behavior. This study is being done by Mike Beijen from the TU Delft.

The purpose of this research study is to investigate the influence of time constraints and user interfaces on web search behavior, task performance and user experience pertaining to certain presented scenarios, and will take you approximately X[1] minutes to complete. The data will be used for analyzing how users interact with a search engine. You will be paid for your participation at the posted rate provided that you complete the whole study, including demographic questions.

Your participation in this study is entirely voluntary and you can withdraw at any time.

We believe there are no known risks associated with this research study; however, as with any online related activity the risk of a breach (exposure of data to an unauthorized person) is always possible. To the best of our ability your answers regarding personal data in this study will remain confidential. We will minimize any risks by storing your data on TU Delft Project Storage, in compliance with GDPR regulations. Anonymised data may later be made openly available for other research. No names are collected. As a participant, you can request data corresponding to your participation in our study at any time for the duration of the research project by sending an e-mail to Mike Beijen

If you have any questions or requests, please send an e-mail to Mike Beijen (m.f.beijen@student.tudelft.nl).

[tickbox] By ticking the box you give your informed consent, indicate you are aware how your data is stored including the risks associated therewith, and agree to participate in accordance with the above conditions.

[1] The duration of the experiment depended on the time condition the user was in. The expected completion times shown were 8, 11, 14, and 16 minutes for the 2, 5, 8, and no time constraint conditions respectively.

Bibliography

[1] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. Estimating position bias without intrusive interventions. In J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman, editors, *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 474–482. ACM, 2019. doi: 10.1145/3289600.3291017. URL https://doi.org/10.1145/3289600.3291017.

[2] Ioannis Arapakis and Luis A. Leiva. Learning efficient representations of mouse movements to predict user attention. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1309–1318. ACM, 2020. doi: 10.1145/3397271.3401031. URL https://doi.org/10.1145/3397271.3401031.

[3] Ioannis Arapakis, Xiao Bai, and Berkant Barla Cambazoglu. Impact of response latency on user behavior in web search. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 103–112. ACM, 2014. doi: 10.1145/2600428.2609627. URL https://doi.org/10.1145/2600428.2609627.

[4] Ioannis Arapakis, Luis A. Leiva, and Berkant Barla Cambazoglu. Know your onions: Understanding the user experience with the knowledge module in web search. In James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu, editors, *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1695–1698. ACM, 2015. doi: 10.1145/2806416.2806591. URL https://doi.org/10.1145/2806416.2806591.

[5] Leif Azzopardi and Guido Zuccon. An analysis of the cost and benefit of search interactions. In Ben Carterette, Hui Fang, Mounia Lalmas, and Jian-Yun Nie, editors, *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval, ICTIR 2016, Newark, DE, USA, September 12- 6, 2016*, pages 59–68. ACM, 2016. doi: 10.1145/2970398.2970412. URL https://doi.org/10.1145/2970398.2970412.

[6] Natã Miccael Barbosa and Monchu Chen. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 543. ACM, 2019. doi: 10.1145/3290605.3300773. URL https://doi.org/10.1145/3290605.3300773.

[7] Nilavra Bhattacharya and Jacek Gwizdka. Measuring learning during search: Differences in interactions, eye-gaze, and semantic similarity to expert knowledge. In Leif Azzopardi, Martin Halvey, Ian Ruthven, Hideo Joho, Vanessa Murdock, and Pernilla Qvarfordt, editors, *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*, pages 63–71. ACM, 2019. doi: 10.1145/3295750.3298926. URL https://doi.org/10.1145/3295750.3298926.

[8] B. S. Bloom, M. B. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain*. Longmans Green, New York, 1956.

[9] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[10] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *CoRR*, abs/1210.0962, 2012. URL http://arxiv.org/abs/1210.0962.

[11] Charles L. A. Clarke, Eugene Agichtein, Susan T. Dumais, and Ryen W. White. The influence of caption features on clickthrough patterns in web search. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 135–142. ACM, 2007. doi: 10.1145/1277741.1277767. URL https://doi.org/10.1145/1277741.1277767.

[12] Anita Crescenzi, Robert Capra, and Jaime Arguello. Time pressure, user satisfaction and task difficulty. In *Beyond the Cloud: Rethinking Information Boundaries - Proceedings of the 76th ASIS&T Annual Meeting, ASIST 2013, Montreal, Canada, November 1-5, 2013*, volume 50 of *Proceedings of the Association for Information Science and Technology*, pages 1–4. Wiley, 2013. doi: 10.1002/meet.14505001121. URL https://doi.org/10.1002/meet.14505001121.

[13] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Time pressure and system delays in information search. In Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto, editors, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 767–770. ACM, 2015. doi: 10.1145/2766462.2767817. URL https://doi.org/10.1145/2766462.2767817.

[14] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Impacts of time constraints and system delays on user experience. In Diane Kelly, Robert Capra, Nicholas J. Belkin, Jaime Teevan, and Pertti Vakkari, editors, *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016*, pages 141–150. ACM, 2016. doi: 10.1145/2854946.2854976. URL https://doi.org/10.1145/2854946.2854976.

[15] Anita Crescenzi, Rob Capra, Bogeum Choi, and Yuan Li. Adaptation in information search and decision-making under time constraints. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, page 95–105, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380553. doi: 10.1145/3406522.3446030. URL https://doi.org/10.1145/3406522.3446030.

[16] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In Mary Beth Rosson and David J. Gilmore, editors, *Proceedings of the 2007 Conference on Human Factors in Computing Systems, CHI 2007, San Jose, California, USA, April 28 - May 3, 2007*, pages 407–416. ACM, 2007. doi: 10.1145/1240624.1240690. URL https://doi.org/10.1145/1240624.1240690.

[17] Djellel Eddine Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek, editors, *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 135–143. ACM, 2018. doi: 10.1145/3159652.3159661. URL https://doi.org/10.1145/3159652.3159661.

[18] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, SIGIR '21, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3404835.3462851. URL https://doi.org/10.1145/3404835.3462851.

[19] Susan T. Dumais, Edward Cutrell, and Hao Chen. Optimizing search by showing results in context. In Julie A. Jacko and Andrew Sears, editors, *Proceedings of the CHI 2001 Conference on Human Factors in Computing Systems, Seattle, WA, USA, March 31 - April 5, 2001*, pages 277–284. ACM, 2001. doi: 10.1145/365024.365116. URL https://doi.org/10.1145/365024.365116.

[20] Lorik Dumani and Ralf Schenkel. Quality-aware ranking of arguments. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 335–344. ACM, 2020. doi: 10.1145/3340531.3411960. URL https://doi.org/10.1145/3340531.3411960.

[21] Ashlee Edwards, Diane Kelly, and Leif Azzopardi. The impact of query interface design on stress, workload and performance. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, volume 9022 of *Lecture Notes in Computer Science*, pages 691–702, 2015. doi: 10.1007/978-3-319-16354-3_76. URL https://doi.org/10.1007/978-3-319-16354-3_76.

[22] Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1419828112. URL https://www.pnas.org/content/112/33/E4512.

[23] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. Suppressing the search engine manipulation effect (SEME). *Proc. ACM Hum. Comput. Interact.*, 1(CSCW):42:1–42:22, 2017. doi: 10.1145/3134677. URL https://doi.org/10.1145/3134677.

[24] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Buchner AG. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods Instruments & Computers*, 39:175–191, 05 2007.

[25] Jennifer Fernquist and Ed H. Chi. Perception and understanding of social annotations in web search. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 403–412, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488424. URL https://doi.org/10.1145/2488388.2488424.

[26] Olivia Foulds, Leif Azzopardi, and Martin Halvey. Investigating the influence of ads on user search performance, behaviour, and experience during information seeking. In Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith, editors, *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*, pages 107–117. ACM, 2021. doi: 10.1145/3406522.3446024. URL https://doi.org/10.1145/3406522.3446024.

[27] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *Int. J. Hum. Comput. Interact.*, 35(6):456–467, 2019. doi: 10.1080/10447318.2018.1456150. URL https://doi.org/10.1080/10447318.2018.1456150.

[28] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing knowledge gain of users in informational search sessions on the web. In Chirag Shah, Nicholas J. Belkin, Katriina Byström, Jeff Huang, and Falk Scholer, editors, *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*, pages 2–11. ACM, 2018. doi: 10.1145/3176349.3176381. URL https://doi.org/10.1145/3176349.3176381.

[29] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. Evaluation metrics for measuring bias in search engine results. *Information Retrieval Journal*, 24, 04 2021. doi: 10.1007/s10791-020-09386-w.

[30] Amira Ghenai, Mark D. Smucker, and Charles L. A. Clarke. A think-aloud study to understand factors affecting online health search. In Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska, editors, *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*, pages 273–282. ACM, 2020. doi: 10.1145/3343413.3377961. URL https://doi.org/10.1145/3343413.3377961.

[31] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7805–7813. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6285.

[32] Ivan Habernal and Iryna Gurevych. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany, 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1150.

[33] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[34] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

[35] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. Domain bias in web search. In Eytan Adar, Jaime Teevan, Eugene Agichtein, and Yoelle Maarek, editors, *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, pages 413–422. ACM, 2012. doi: 10.1145/2124295.2124345. URL https://doi.org/10.1145/2124295.2124345.

[36] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum*, 51(1):4–11, 2017. doi: 10.1145/3130332.3130334. URL https://doi.org/10.1145/3130332.3130334.

[37] Hideo Joho and Joemon M. Jose. A comparative study of the effectiveness of search result presentation on the web. In Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings*, volume 3936 of *Lecture Notes in Computer Science*, pages 302–313. Springer, 2006. doi: 10.1007/11735106_27. URL https://doi.org/10.1007/11735106_27.

[38] Yvonne Kammerer and Peter Gerjets. How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In Carlos Hitoshi Morimoto, Howell O. Istance, Aulikki Hyrskykari, and Qiang Ji, editors, *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ETRA 2010, Austin, Texas, USA, March 22-24, 2010*, pages 299–306. ACM, 2010. doi: 10.1145/1743666.1743736. URL https://doi.org/10.1145/1743666.1743736.

[39] Yvonne Kammerer and Peter Gerjets. Effects of search interface and internet-specific epistemic beliefs on source evaluations during web search for medical information: an eye-tracking study. *Behav. Inf. Technol.*, 31(1):83–97, 2012. doi: 10.1080/0144929X.2011.599040. URL https://doi.org/10.1080/0144929X.2011.599040.

[40] Yvonne Kammerer and Peter Gerjets. The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *Int. J. Hum. Comput. Interact.*, 30(3):177–191, 2014. doi: 10.1080/10447318.2013.846790. URL https://doi.org/10.1080/10447318.2013.846790.

[41] Nattiya Kanhabua and Avishek Anand. Temporal information retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1235–1238, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340694. doi: 10.1145/2911451.2914805. URL https://doi.org/10.1145/2911451.2914805.

[42] Diane Kelly and Leif Azzopardi. How many results per page?: A study of SERP size, search behavior and user experience. In Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto, editors, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 183–192. ACM, 2015. doi: 10.1145/2766462.2767732. URL https://doi.org/10.1145/2766462.2767732.

[43] Silvia Knobloch-Westerwick and Jingbo Meng. Looking the other way. *Commun. Res.*, 36(3): 426–448, 2009. doi: 10.1177/0093650209333030. URL https://doi.org/10.1177/0093650209333030.

[44] Bill Kules and Robert Capra. Creating exploratory tasks for a faceted search interface. *Proc. of HCIR 2008*, pages 18–21, 2008.

[45] Shuang Li, Xiang Lan, Yuezhi Zhou, and Yaoxue Zhang. Exploring and understanding web search behavior with human activities. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017, San Francisco, CA, USA, August 4-8, 2017*, pages 1–8. IEEE, 2017. doi: 10.1109/UIC-ATC.2017.8397516. URL https://doi.org/10.1109/UIC-ATC.2017.8397516.

[46] Sheng Lin, Peiquan Jin, Xujian Zhao, and Lihua Yue. Exploiting temporal information in web search. *Expert Systems with Applications*, 41(2):331 – 341, 2014. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2013.07.048. URL http://www.sciencedirect.com/science/article/pii/S0957417413005356.

[47] Chang Liu and Yiming Wei. The impacts of time constraint on users' search strategy during search process. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology*, ASIST '16, USA, 2016. American Society for Information Science.

[48] Chang Liu, Fan Yang, Yu Zhao, Qin Jiang, and Lu Zhang. What does time constraint mean to information searchers? In David Elsweiler, Bernd Ludwig, Leif Azzopardi, and Max L. Wilson, editors, *Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26-29, 2014*, pages 227–230. ACM, 2014. doi: 10.1145/2637002.2637029. URL https://doi.org/10.1145/2637002.2637029.

[49] Jiqun Liu, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. Task, information seeking intentions, and user behavior: Toward A multi-level understanding of web search. In Leif Azzopardi, Martin Halvey, Ian Ruthven, Hideo Joho, Vanessa Murdock, and Pernilla Qvarfordt, editors, *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR 2019, Glasgow, Scotland, UK, March 10-14, 2019*, pages 123–132. ACM, 2019. doi: 10.1145/3295750.3298922. URL https://doi.org/10.1145/3295750.3298922.

[50] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. Influence of vertical result in web search examination. In Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier A. Ribeiro-Neto, editors, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 193–202. ACM, 2015. doi: 10.1145/2766462.2767714. URL https://doi.org/10.1145/2766462.2767714.

[51] Mari-Carmen Marcos, Ferran Gavin, and Ioannis Arapakis. Effect of snippets on user experience in web search. In Pere Ponsa and Daniel Guasch, editors, *Proceedings of the XVI International Conference on Human Computer Interaction, Interacción 2015, Vilanova i la Geltrú, Spain, September 7-9, 2015*, pages 47:1–47:8. ACM, 2015. doi: 10.1145/2829875.2829916. URL https://doi.org/10.1145/2829875.2829916.

[52] David Maxwell and Leif Azzopardi. Stuck in traffic: how temporal delays affect search behaviour. In David Elsweiler, Bernd Ludwig, Leif Azzopardi, and Max L. Wilson, editors, *Fifth Information Interaction in Context Symposium, IIiX '14, Regensburg, Germany, August 26-29, 2014*, pages 155–164. ACM, 2014. doi: 10.1145/2637002.2637021. URL https://doi.org/10.1145/2637002.2637021.

[53] David Maxwell and Leif Azzopardi. Information scent, searching and stopping - modelling SERP level stopping behaviour. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 210–222. Springer, 2018. doi: 10.1007/978-3-319-76941-7_16. URL https://doi.org/10.1007/978-3-319-76941-7_16.

[54] David Maxwell and Claudia Hauff. LogUI: Contemporary Logging Infrastructure for Web-Based Experiments. In *Advances in Information Retrieval (Proc. ECIR)*, pages 525–530, 2021.

[55] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. Searching and stopping: An analysis of stopping rules and strategies. In James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu, editors, *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 313–322. ACM, 2015. doi: 10.1145/2806416.2806476. URL https://doi.org/10.1145/2806416.2806476.

[56] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. A study of snippet length and informativeness: Behaviour, performance and user experience. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 135–144. ACM, 2017. doi: 10.1145/3077136.3080824. URL https://doi.org/10.1145/3077136.3080824.

[57] Dana McKay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. We are the change that we seek: Information interactions during a change of viewpoint. In Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska, editors, *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*, pages 173–182. ACM, 2020. doi: 10.1145/3343413.3377975. URL https://doi.org/10.1145/3343413.3377975.

[58] Alan Medlar, Jing Li, and Dorota Glowacka. Query suggestions as summarization in exploratory search. In Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith, editors, *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*, pages 119–128. ACM, 2021. doi: 10.1145/3406522.3446020. URL https://doi.org/10.1145/3406522.3446020.

[59] Rishabh Mehrotra and Emine Yilmaz. Terms, topics & tasks: Enhanced user modelling for better personalization. In James Allan, W. Bruce Croft, Arjen P. de Vries, and Chengxiang Zhai, editors, *Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ICTIR 2015, Northampton, Massachusetts, USA, September 27-30, 2015*, pages 131–140. ACM, 2015. doi: 10.1145/2808194.2809467. URL https://doi.org/10.1145/2808194.2809467.

[60] Mark R. Meiss, Bruno Gonçalves, José J. Ramasco, Alessandro Flammini, and Filippo Menczer. Agents, bookmarks and clicks: a topical model of web navigation. In Mark H. Chignell and Elaine G. Toms, editors, *HT'10, Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, Toronto, Ontario, Canada, June 13-16, 2010*, pages 229–234. ACM, 2010. doi: 10.1145/1810617.1810658. URL https://doi.org/10.1145/1810617.1810658.

[61] Nina Mishra, Ryen W. White, Samuel Ieong, and Eric Horvitz. Time-critical search. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 747–756. ACM, 2014. doi: 10.1145/2600428.2609613. URL https://doi.org/10.1145/2600428.2609613.

[62] Aditi S. Muralidharan, Zoltán Gyöngyi, and Ed Chi. Social annotations in web search. In Joseph A. Konstan, Ed H. Chi, and Kristina Höök, editors, *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*, pages 1085–1094. ACM, 2012. doi: 10.1145/2207676.2208554. URL https://doi.org/10.1145/2207676.2208554.

[63] Alamir Novin and Eric M. Meyers. Making sense of conflicting science information: Exploring bias in the search engine result page. In Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel, editors, *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, pages 175–184. ACM, 2017. doi: 10.1145/3020165.3020185. URL https://doi.org/10.1145/3020165.3020185.

[64] Heather L. O'Brien, Paul A. Cairns, and Mark Hall. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *Int. J. Hum. Comput. Stud.*, 112:28–39, 2018. doi: 10.1016/j.ijhcs.2018.01.004. URL https://doi.org/10.1016/j.ijhcs.2018.01.004.

[65] Maeve O'Brien and Mark Keane. Modeling result-list searching in the world wide web: The role of relevance topologies and trust bias. In *Proceedings of the 28st Annual Meeting of the Cognitive Science Society (CogSci 2006)*, CogSci '21, Austin, TX, USA, 01 2006. Cognitive Science Society.

[66] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura A. Granka. In google we trust: Users' decisions on rank, position, and relevance. *J. Comput. Mediat. Commun.*, 12(3):801–823, 2007. doi: 10.1111/j.1083-6101.2007.00351.x. URL https://doi.org/10.1111/j.1083-6101.2007.00351.x.

[67] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L. A. Clarke. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz, editors, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, pages 209–216. ACM, 2017. doi: 10.1145/3121050.3121074. URL https://doi.org/10.1145/3121050.3121074.

[68] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa. Analyzing the effects of document's opinion and credibility on search behaviors and belief dynamics. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1653–1662. ACM, 2019. doi: 10.1145/3357384.3357886. URL https://doi.org/10.1145/3357384.3357886.

[69] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa. Analyzing the effects of "people also ask" on search behaviors and beliefs. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 101–110, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370981. doi: 10.1145/3372923.3404786. URL https://doi.org/10.1145/3372923.3404786.

[70] Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. Argument search: Assessing argument relevance. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1117–1120. ACM, 2019. doi: 10.1145/3331184.3331327. URL https://doi.org/10.1145/3331184.3331327.

[71] Filip Radlinski and Thorsten Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. *CoRR*, abs/cs/0605037, 2006. URL http://arxiv.org/abs/cs/0605037.

[72] Jerome Ramos and Carsten Eickhoff. Search result explanations improve efficiency and trust. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1597–1600. ACM, 2020. doi: 10.1145/3397271.3401279. URL https://doi.org/10.1145/3397271.3401279.

[73] Nirmal Roy, Felipe Moraes, and Claudia Hauff. Exploring users' learning gains within search sessions. In Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska, editors, *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*, pages 432–436. ACM, 2020. doi: 10.1145/3343413.3378012. URL https://doi.org/10.1145/3343413.3378012.

[74] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. Note the highlight: Incorporating active reading tools in a search as learning environment. In Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith, editors, *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*, pages 229–238. ACM, 2021. doi: 10.1145/3406522.3446025. URL https://doi.org/10.1145/3406522.3446025.

[75] Ladislao Salmerón, Laura Gil, Ivar Bråten, and Helge I. Strømsø. Comprehension effects of signalling relationships between documents in search engines. *Comput. Hum. Behav.*, 26(3):419–426, 2010. doi: 10.1016/j.chb.2009.11.013. URL https://doi.org/10.1016/j.chb.2009.11.013.

[76] Julia Schwarz and Meredith Ringel Morris. Augmenting web pages and search results to support credibility assessment. In Desney S. Tan, Saleema Amershi, Bo Begole, Wendy A. Kellogg, and Manas Tungare, editors, *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011*, pages 1245–1254. ACM, 2011. doi: 10.1145/1978942.1979127. URL https://doi.org/10.1145/1978942.1979127.

[77] Christina Schwind, Jürgen Buder, Ulrike Cress, and Friedrich W. Hesse. Preference-inconsistent recommendations: An effective approach for reducing confirmation bias and stimulating divergent thinking? *Comput. Educ.*, 58(2):787–796, 2012. doi: 10.1016/j.compedu.2011.10.003. URL https://doi.org/10.1016/j.compedu.2011.10.003.

[78] Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. Automatic argument quality assessment - new datasets and methods. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5624–5634. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1564. URL https://doi.org/10.18653/v1/D19-1564.

[79] Anton van der Vegt, Guido Zuccon, Bevan Koopman, and Anthony Deacon. How searching under time pressure impacts clinical decision making. *Journal of the Medical Library Association : JMLA*, 108(4):564–573, Oct 2020. ISSN 1558-9439. doi: 10.5195/jmla.2020.915. URL https://pubmed.ncbi.nlm.nih.gov/33013213.

[80] Liwen Vaughan and Mike Thelwall. Search engine coverage bias: evidence and possible causes. *Inf. Process. Manag.*, 40(4):693–707, 2004. doi: 10.1016/S0306-4573(03)00063-3. URL https://doi.org/10.1016/S0306-4573(03)00063-3.

[81] Yue Wang, Dawei Yin, Luo Jie, Pengyuan Wang, Makoto Yamada, Yi Chang, and Qiaozhu Mei. Beyond ranking: Optimizing whole-page presentation. In Paul N. Bennett, Vanja Josifovski, Jennifer Neville, and Filip Radlinski, editors, *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 103–112. ACM, 2016. doi: 10.1145/2835776.2835824. URL https://doi.org/10.1145/2835776.2835824.

[82] Ryen White. Beliefs and biases in web search. In Gareth J. F. Jones, Paraic Sheridan, Diane Kelly, Maarten de Rijke, and Tetsuya Sakai, editors, *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*, pages 3–12. ACM, 2013. doi: 10.1145/2484028.2484053. URL https://doi.org/10.1145/2484028.2484053.

[83] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In Ricardo Baeza-Yates, Paolo Boldi, Berthier A. Ribeiro-Neto, and Berkant Barla Cambazoglu, editors, *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, pages 132–141. ACM, 2009. doi: 10.1145/1498759.1498819. URL https://doi.org/10.1145/1498759.1498819.

[84] Mathew J. Wilson and Max L. Wilson. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *J. Assoc. Inf. Sci. Technol.*, 64(2):291–306, 2013. doi: 10.1002/asi.22758. URL https://doi.org/10.1002/asi.22758.

[85] Zhijing Wu, Mark Sanderson, B. Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. Providing direct answers in search results: A study of user behavior. In Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1635–1644. ACM, 2020. doi: 10.1145/3340531.3412017. URL https://doi.org/10.1145/3340531.3412017.

[86] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. Revealing the role of user moods in struggling search tasks. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1249–1252. ACM, 2019. doi: 10.1145/3331184.3331353. URL https://doi.org/10.1145/3331184.3331353.

[87] Takehiro Yamamoto, Yusuke Yamamoto, and Sumio Fujita. Exploring people's attitudes and behaviors toward careful information seeking in web search. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang, editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 963–972. ACM, 2018. doi: 10.1145/3269206.3271799. URL https://doi.org/10.1145/3269206.3271799.

[88] Yusuke Yamamoto and Takehiro Yamamoto. Query priming for promoting critical thinking in web search. In Chirag Shah, Nicholas J. Belkin, Katriina Byström, Jeff Huang, and Falk Scholer, editors, *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*, pages 12–21. ACM, 2018. doi: 10.1145/3176349.3176377. URL https://doi.org/10.1145/3176349.3176377.

[89] Philip S. Yu, Xin Li, and Bing Liu. On the temporal dimension of search. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *Proceedings of the 13th international conference on World Wide Web - Alternate Track Papers & Posters, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 448–449. ACM, 2004. doi: 10.1145/1013367.1013519. URL https://doi.org/10.1145/1013367.1013519.

[90] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 1011–1018. ACM, 2010. doi: 10.1145/1772690.1772793. URL https://doi.org/10.1145/1772690.1772793.

[91] Yao Zhang and Chang Liu. Users' knowledge use and change during information searching process: A perspective of vocabulary usage. In Ruhua Huang, Dan Wu, Gary Marchionini, Daqing He, Sally Jo Cunningham, and Preben Hansen, editors, *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*, pages 47–56. ACM, 2020. doi: 10.1145/3383583.3398532. URL https://doi.org/10.1145/3383583.3398532.

[92] Yuchen Zhang, Weizhu Chen, Dong Wang, and Qiang Yang. User-click modeling for understanding and predicting search-behavior. In Chid Apté, Joydeep Ghosh, and Padhraic Smyth, editors, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 1388–1396. ACM, 2011. doi: 10.1145/2020408.2020613. URL https://doi.org/10.1145/2020408.2020613.

[93] Mengdie Zhuang, Gianluca Demartini, and Elaine G. Toms. Understanding engagement through search behaviour. In Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li, editors, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1957–1966. ACM, 2017. doi: 10.1145/3132847.3132978. URL https://doi.org/10.1145/3132847.3132978.