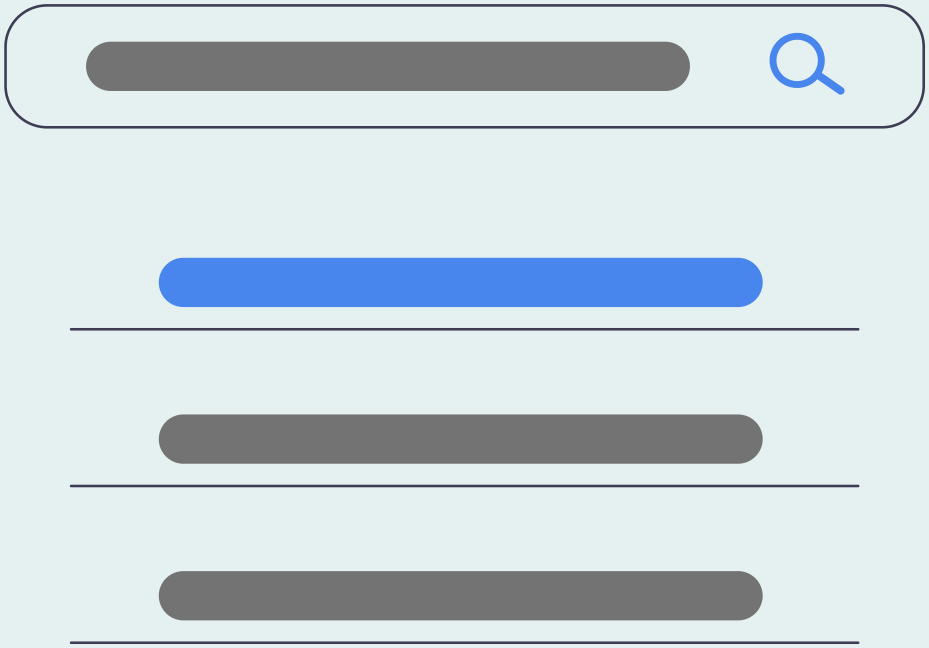
KAN-CBUSO2000U Master's Thesis

The effect of search engine ranking elements on the user experience (UX)

Uncovering how Google's algorithmic choices affect users behind the screens



STUDENT
Lærke Saura Birk
Student number: 133944
MSc in BA & E-Business
17th of May 2021
lasa19ad@student.cbs.dk

SUPERVISOR
Tina Kretschel
PhD Fellow
Department of Marketing
tkr.marktg@cbs.dk

Page count: 79
Character count: 172.858

CBS
COPENHAGEN

Purpose - Algorithms define how users look for information online. Search engines, notably Google, have marketed their services by fiercely defending an algorithmic ideology that sustains their business model. A consequence of this underlying algorithmic logic is the unquestioned everyday use of search engine rankings. These have become a convenient tool for users to unproblematically look for information on the web as they reduce complexity in a cluttered online sphere. Experts have focused on investigating how these rankings appear regarding content and form; and come with suggestions on how they can be improved through evaluative research. Despite that, the effect of these rankings on the user experience (UX) is underinvestigated. In this work, stance bias and featured snippets are explored as ranking elements that potentially affect the measured UX on Google.

Methodology - On the researcher's own initiative, data on different subjects' UX was collected with a survey-based 2x2 full factorial vignette experiment (N=136) to assess three hypotheses through a quantitative data analysis. Participants were presented 3 hypothetical search situations and synthetically created Google rankings. These were assessed through a measurement instrument that calculated a participant's overall UX mean score with the use of a 5-point Likert scale. The experimental data was analyzed with the Python programming language performing a multivariate parametric test (two-way ANOVA) to compare experimental group means. The adjusted R^2 is reported to show the model fit, and the partial η^2 quantifies effect sizes.

Findings - The difference among the mean UX scores in the different experimental groups showed that participants treated with stance bias on rankings had a significantly lower UX than participants who did not experience any stance bias ($p = 0.000$, coef. = -0.6216), reporting a significant simple effect. No simple effect was reported for participants exposed to featured snippets ($p = 0.506$), and interaction effects were neither identified for participants facing featured snippets in combination with stance bias ($p = 0.415$).

Implication - Users register a decrease in their UX when exposed to stances on rankings, whereas seeing featured snippets does not affect the UX. The study outcome suggests that users are willing to engage in research tasks when performing informational queries instead of seeing results on rankings that offer a particular editorial positioning on a topic. This finding encourages managers to reassess their search engine optimization (SEO) strategy to enhance the UX. SEO strategists can attract users to their content on SE rankings by removing stances from their pages' snippets. This research also challenges some pre-established beliefs in academia, such as claiming that users rely blindly on rankings or their idealized view towards Google. Future research can contextualize these findings with qualitative data and build on the regression model presented in this work.

Keywords: Google, user experience (UX), rankings, search engine bias, featured snippet

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANOVA - Analysis of variance
CB - Confirmation bias
CFA - Confirmatory factor analysis
CIS - Communicated information structure
FS - Featured snippet
H1 - Hypothesis 1
H2 - Hypothesis 2
H3 - Hypothesis 3
HCI - Human-computer interaction
IR - Information retrieval
IS - Information systems
Min - Minimum value
Max - Maximum value
OS - Outcome satisfaction
RQ - Research question
SASI - School Assignment Satisfaction Index
SB - Stance bias
SE - Search engine
SEB - Search engine bias
SEA - Search engine advertising
SEM - Search engine marketing
SEO - Search engine optimization
SERP - Search engine results page
T1 - Treatment 1
T2 - Treatment 2
T3 - Treatment 3
TES - Task effectiveness
TEY - Task efficiency
UI - User interface
UX - User experience
VB - Visual brand
WPR - Whole page relevance

1. Introduction

"Emily, I am going to Google you!" yelled Richard Gilmore, grandfather and one of the main characters in the American TV series Gilmore Girls, while enjoying some time on his laptop. He was not looking for any type of information in particular. Still, he implied he had found the pleasure in simply typing queries in the search bar, hoping for something interesting to come up about his wife, Emily. The described scene was aired in 2003 (Sherman-Palladino, Palladino & Moore, 2003), and the characters lived in an age where films were rented in video clubs, people were localized through pagers, and voice messages were constantly left on answering machines. Many of these tools and practices would seem analog and outdated today. Yet, Google is still known as a "reference experience" that has set "expectations for billions of users" (Rosenberg, 2018, p. 29).

Search engines (SEs) allow users to access vast amounts of information from the palm of their hands; and performing searches online has become an embedded practice for citizens all over the world (Ziakis et al., 2019; Harvey & Pointon, 2017). Whenever users feel the urge to input queries into a search bar, there is a particular SE that stands out for being their preferred choice: Google. Using Google for information-seeking purposes has become a habit in people's daily routines (Liu & Li, 2016), and the SE has become an apparatus of social order that dictates contemporary information dynamics, which go unquestioned (Bilić, 2016). Google has achieved this privileged position by promoting an "algorithmic objectivity" that has led to the SE to be "mythologized" among SE users (Gillespie, 2014; p. 14). Google has therefore earned a considerable spot on the SE market, accumulating 92,05% of the market share by February 2021 (StatsCounter, 2021), and has completely wiped out the competition due to its superiority in usefulness (Bilić, 2016).

For the time being, SEs also serve as information gatekeepers in the online sphere by reducing complexity and providing orientation in a myriad of information (Latzer et al., 2016; Steiner et al., 2020; Röhle, 2009). One of the core SE features reinforcing this gatekeeping logic is SE rankings (Röhle, 2009), which are responsible for ordering sites and making choices for the user, contributing to significantly simplifying decision-making processes (e.g., Beer, 2016; Latzer et al., 2016; Steiner et al., 2020). The role of rankings is fundamental on contemporary SEs, as they contribute to avoiding having humans making individualized choices regarding the relevance of websites (Gao & Shah, 2020; Goldman, 2005). Applying this ranking logic also means that SEs such as Google decide which information providers get a privileged position on people's screens, as SEs must make editorial choices. In other words, Google makes operational decisions on how to present website data (Goldman, 2005).

Rankings impact users' beliefs and, in consequence, they induce cognitive biases (Gao & Shah, 2020). An example of how biases can occur on SE rankings is favoring a particular stance (i.e., results can be in favor or against a specific topic). Some examples of how different viewpoints are highlighted on search results were proposed by Gezici et al. (2021) with issues such as abortion, medical marijuana, or gay marriage. Google has a clear idea on how these topics should be presented to different users using signals such as "previous search history, quality of content on websites, recommendations from users' social networks, etc." In that sense, Google provides relevance by leveraging existing data and predicting which viewpoints should be favored to show a particular user belief-consistent information (Bilić, 2016; Kayhan, 2015, p. 1).

Ranked presentation of content on the search engine results page (SERP) also affects users' behavior (e.g., Ong et al., 2017; Ghose, Goldfarb & Han, 2013; Joachims et al., 2017) as there is a tendency to click on the top results on rankings (e.g., Joachims et al., 2017; Baeza-Yates, 2018). Users trust the SE's ability, notably Google, to rank results neutrally (Pan et al., 2007; Schultheiß & Lewandowski, 2021). Still, they are mostly unaware of the consequences of SE rankings, especially when exploring topics where diverse or opposite views are possible (Gezici et al., 2021). For instance, search results polarized towards a particular political option can change the voting preferences of undecided voters by 20% (Epstein & Robertson, 2015). Therefore, SE rankings and their underlying algorithm are known to have huge societal impacts that go unnoticed by users (Beer, 2017).

In parallel, rich elements on SERPs such as featured snippets (i.e., short summaries of a search result shown directly on SERPs) have emerged in the past years to improve rankings' presentation. Many studies have focused on understanding how featured snippets appear, and they have been evaluated based on their ability to adequately summarize a document (Ageev et al., 2013; Strzelecki & Rutecka, 2020). However, it remains unclear how different snippet layouts affect the user. Still, SEs such as Google have devoted themselves to updating their algorithm and the result presentation to enhance the user experience (UX) by adding new layout elements similar to featured snippets, such as other rich snippets (Marcos et al., 2015; Bilić, 2016).

It is precisely this underlying algorithmic philosophy that defines Google's business model. The algorithm enabling the current ranking logic is being fiercely defended by Google's engineers and marketed to end-users by perpetuating an algorithmic ideology (Mager, 2012). Even though the academic community has

shown concern towards the outcomes of SEs (e.g., biases, unfairness, or low quality of information), Google's strategy is to stay away from conflict and make the algorithm responsible for any controversial outcome (Mager, 2012; Bilić, 2016; Modave et al., 2014).

This behavior from Google's side becomes particularly eye-opening if a thought is given on Google's intentions, as the SE primarily concentrates on providing relevance to internet users (Bilić, 2016). As Mager (2012) noted, Google is interested in keeping users satisfied to provide an optimal service to their real customers: Advertisers. In that sense, Google uses data as a currency. Before a user starts inputting a query in the search bar, the Californian company has already begun collecting data about a user's desires and intentions, which is later sold for advertising purposes. To collect data, Google needs users, and users become the "goldmine" that enables success. Therefore, ensuring that the people behind the screens get an excellent experience and are willing to stay loyal to the brand is key to sustaining Google's current business model (Mager, 2012, p. 776). Not only experts have found out that Google takes the UX on their services seriously. They have reported so themselves, mentioning that they have "added a variety of user experience criteria" as ranking factors (*Evaluating Page Experience for a Better Web*, 2020, n.p.).

This current research builds on previous premises established in the literature and hypotheses that nothing on rankings is coincidental. Based on previous literary discussions, it can be questioned whether elements such as bias and rich snippets become a side effect of Google trying to craft a relevant experience (Bilić, 2016) and, in consequence, a positive UX. Despite that, this question remains unanswered by academia. Studies from the fields of sociology and information systems (IS) have evaluated the type of content and layout of rankings with mainly descriptive, exploratory, and evaluative research. Thus, no link has been established between the configuration of SE rankings and Google's intentions of offering a SE that enhances the UX. Previous studies have focused on highlighting how rankings bring across consequences for the user, such as bias or low information quality (e.g., Modave et al., 2014; Pan et al., 2007; Lewandowski, 2017). Moreover, IS research aims have been very evaluative, and authors have tried to create solutions to problems such as biases and unfairness (e.g., Gao & Shah, 2020). Hence, state-of-the-art literature leaves room for explanatory research. Instead of following the trend of orienting investigations towards solving a problem (e.g., bias on rankings), it becomes appropriate to establish a relationship between the current ranking configuration and the UX to understand whether phenomena labelled as "troubling" by academia might be positive for the user behind the screen (Haim et al., 2018, p. 339).

Building on the premises exposed in this introduction, a relationship between the ranking configuration and the UX is likely to be found. Hence, the objective of this study is further exploring this link to fill a gap in the existing literature. Establishing such a relationship provides the academic field with updated understandings on how rankings impact users browsing the web and whether bias and rich snippets carry societal consequences. A deeper and explanatory understanding between rankings and users also becomes an object of guidance in IS research to reassess problem formulations in evaluative solution-oriented studies. Getting a deeper understanding on the UX on today's Google rankings is also attractive for managers in search engine optimization (SEO) eager to uncover what defines a positive UX on Google. Such an understanding would provide guidance to apply more user-centric SEO practices to attract visitors to websites.

To satisfy the academic, social and managerial interests mentioned, this study will attempt to understand how particularly bias and rich snippets have an effect on rankings by answering the following research question (RQ): *How does the presence of stance bias and featured snippets on rankings affect the user experience on search engines?*, focusing on Google as a use case. The scope of this research focuses on investigating stance bias and featured snippets as ranking elements, given that these allow to differentiate between the content and form of rankings. Furthermore, they can be easily defined as constructs and quantified in quantitative research settings. The data to answer this RQ is gathered through a 2x2 full factorial between survey-based experiment, which is used to assess three research hypotheses established using state-of-the-art theories and with an abductive approach to theory development.

Hence, the following pages are oriented towards designing an experiment that answers the RQ. The state-of-the-art theoretical underpinnings necessary for conducting this research are provided in the next chapter (2). This section emphasizes the underlying characteristics of today's commercial SEs, and constructs are defined to be later appropriately used during experimentation. Chapter 3 outlines the research design, where three hypotheses are formed based on existing theories. The subsequent section of this work (chapter 4) describes the methodology for data collection and analysis, focusing on detailing the survey-based experiment and justifying the suitability of this method for the research objectives. The validity, reliability, and steps for analyzing the data are also provided to offer a transparent overview of the experimental outcome. The subsequent results chapter (5) shows the findings in the data analysis. The discussion chapter (6) contextualizes the results for academic and managerial readers by including insights on how the experimental outcomes are used for both theory-building and assessing previous theoretical findings. A limitations section is also included for transparency and provides a better understanding of

this study's outcomes. Further research directions are also suggested. In the concluding chapter (7), the findings are put into a larger perspective to outline the most relevant implications of this thesis.

Figure 1. *Outline of the thesis structure.*

2. Theoretical underpinnings

In the previous chapter, Google was presented as a SE with an unquestionable market dominance, capable of defining users' online information habits. Google, and by extension, SEs, should be understood under the concept of information retrieval systems. These have been extensively explored in information systems (IS) works since 1953 (Ellis, 1989) and have evolved to the SEs used today.

This chapter is divided into five sections that outline state-of-the-art theoretical findings on (1) the context of use of SEs, (2) the role of rankings and their influence, (3) the construct of user experience (UX) and how to interpret it in the domain of SEs, (4) biases on SEs, and (5) the characteristics of user interfaces on SEs.

2.1 Search engines (SE)

Search engines (SEs) allow users to access vast amounts of information within a few clicks or from the palm of their hands (e.g., Ziakis et al., 2019; Harvey & Pointon, 2017). They have become one of the technologies with the most considerable societal impact in the 21st century as they contribute to simplifying decision-making processes both for businesses and individuals thanks to the underlying algorithm (e.g., Beer, 2016; Latzer et al., 2016; Steiner et al., 2020). They can do so thanks to their online gatekeeping logic: They select the information relevant to the user's search queries, giving quick and simple access to the online sphere (Latzer et al., 2016; Steiner et al., 2020). Also, Google and other SEs are primarily advertised as being available to a large set of users independently of age, gender, or other demographic factors (Mehrotra et al., 2017). As a result, both the production of content on the Internet and the use of SEs is increasing (Palos-Sanchez & Saura, 2018). Therefore, today's SEs are constantly changing with frequent algorithmic updates, turning them into products that are in a permanent state of evolution (Ziakis, 2019).

2.1.1 Search engines in the context of information retrieval systems

Contemporary SEs are part of a broader concept: Information retrieval (IR) systems. SEs are, at the core, IR systems (Baeza-Yates & Ribiero-Neto, 1999) that are built to facilitate human-computer interaction (HCI) (Spink & Saracevic, 1998). This subsection focuses on exploring the definition of IR, and how the concept has been explored in previous literature to provide a deeper understanding of the underlying logic of today's SEs.

2.1.1.1 Definition of IR and IR systems

Aggarwal (2019, p. 259) defines IR as a "process of satisfying user information needs" which he/she has expressed through a textual query. Building on these definitions, Spink & Saracevic (1998, p. 250) state that IR is an "interactive process" involving HCI. Hence, based on the previous definitions, IR is a process requiring HCI, started by a user inputting a query to satisfy an information need.

An IR system has been defined as a synonym of a SE by Baeza-Yates & Ribiero-Neto (1999) and describes it as the system that can process the user's queries to retrieve relevant information from a database. Given these definitions, throughout this study, an IR system will be understood as the system that facilitates the activity of IR by connecting a user to a database system.

2.1.1.2 The architecture of an IR system

IR systems cannot be understood without considering a database's presence. IR systems are built to grant access to databases that include the relevant documents to the user's information need, given a particular search query (Baeza-Yates & Ribiero-Neto, 1999). According to Silberschatz, Korth & Sudarshan (2020), in the context of the web, "each web page can be considered to be a document." These documents contain or have keywords associated with them and act as the reason why it is often mentioned that SEs do keyword searches. Keyword-based information retrieval can also be used to retrieve other types of data than web pages, such as images or audio (Grossman & Frieder, 2012, p. xv).

Web crawlers

An IR system relies on a database containing data to encounter websites. The tool responsible for filling the database is known as a web crawler, a program that downloads web pages given one or more seed URLs. The web crawler extracts the hyperlinks in these URLs and recursively continues downloading the web pages within these hyperlinks (Najork, 2009).

Indexing

Out of these collected web pages, an IR system indexes the web pages as documents. Indices are crucial for efficient database handling and, in consequence, well-functioning IR systems (Silberschatz, Korth & Sudarshan, 2020). Indexing is a process happening before the retrieval process, and it consists of collecting documents to store in the form of text to ensure efficient retrieval. Hence, it allows a system to quickly match a user query with the documents stored in a database (Huang & Zhang, 2009). In non-text documents such as pictures or geographic locations, metadata is used for resource discovery. In other words, documents carry "self-descriptive data" (Dempsey & Heery, 1998, p. 146).

Rankings

Another characteristic of an IR system is its ranking mechanism, used to define how SE algorithms sort documents. Each IR system handles keywords differently, depending on how it is modeled. Modeling is the process of producing a ranking function responsible for assigning scores to documents regarding a given query. The process consists of (1) building a logical framework for representing documents and queries; and (2) defining a ranking function that can compute a rank to each document depending on the issued query (Baeza-Yates & Ribiero-Neto, 1999). Hence, many IR systems can estimate the relevance of the documents to a query "so that the documents can be shown in order of estimated relevance" by using "information on keyword occurrences, as well as hyperlink information" (Silberschatz, Korth & Sudarshan, 2020, p. 383).

2.1.2 Task complexity during the use of SE

Users satisfy their information needs using SEs. To do so, users must engage in search tasks. For years, it has been understood that success in information seeking depends on the user's task and the problems encountered during the search process (Byström & Järvelin, 1995). It is known that the more the task complexity increases, the more time is spent using a SE by issuing more queries, clicking on more search results, or clicking on more URLs (Wu et al, 2012; Wu et al., 2014, Pan et al., 2007).

Furthermore, not all users interact with a system in the same way, and a single user can also show behavioral variability when interacting with a SE depending on contextual circumstances (Carterette, Kanoulas & Yilmaz, 2012). Therefore, the use of SEs can vary deeply from user to user, who react differently to a variety of search tasks.

Tasks also have different natures, and Vuong et al. (2019, p. 1250-1251) categorize them with three factors: “Task goals”, “individual intentions”, and the “substance domain” of the tasks.

Task goals

Understanding task goals is equivalent to answering why the user has decided to search for something, and SEs developers must take task-goals into consideration to improve their performance (Rose & Levinson, 2004). The same idea is supported by Moffat et al. (2017), defending that SE effectiveness metrics should be goal sensitive. Tasks are divided into three primary goals: Navigational, informational & resources. The first goal refers to users trying to navigate to a website that they know in advance. Some example queries would be `Yale university` or `Facebook`. The second, informational, refers to users trying to learn something by reading or viewing web pages (`how to recycle`). Finally, resource refers to tasks that aim to obtain something, a resource available on web pages that is not explicitly information. These are also divided into categories such as download (`download Skype`), entertainment (`watch Gilmore Girls`), interact (`online percentage calculator`), or obtain (`standard Danish rent contract`) (Vuong et al. 2019; Rose & Levinson, 2004).

Individual intentions

Whereas task goals determine the desired outcome of a search task, individual intentions affect the search process. Depending on the search query and the user’s contextual background, an individual can be driven by different intentions when searching (Vuong et al., 2019). Also, users can have very different SE usage expectations, even though task goals are the same. For example, users can intend to browse a different number of documents for each query or issue a different number of queries to fulfill the same goal (Bailey et al., 2015). Therefore, each search process will depend on the searcher’s expectations towards a particular search task (Moffat et al., 2017). Vuong et al. (2019 p. 1254) separate individual intentions into four different categories: (1) Tasks for “Being Creative” such as “writing/composing documents”; (2) Tasks to “Enjoy Oneself” through social media, video streaming, or music consumption; (3) tasks to “Gain Knowledge,” which are learning or research-related activities; and (4) “Daily Activity,” referring to routinary activities like online shopping, looking up the news, or making travel plans. These intentions are also bound to the needs of a user. These needs can be informational, transactional or navigational. Informational refers to looking for information about a particular topic. Transactional searches consist in the user looking for a product or service. Navigational intentions are oriented towards finding particular websites that the searcher knows in advance of searching (Schultz, 2020).

Substance domain

Finally, a task's substance domain refers to the field the task belongs to. Some examples are free time, business, programming, social life, or studying. All tasks inside a domain, such as business, belong to the business domain, regardless of the goal or the intention (Vuong et al., 2019).

2.1.3 Searcher behavior

Users will conduct their search tasks in sessions and issue queries iteratively to reach a goal. Every search session includes subtasks: Query formulation, result scanning, document link clicking, document reading and judgment, and stopping. These subtasks are associated with certain behavioral patterns and determine how users interact with a SE (Baskaya, Keskustalo & Järvelin, 2013).

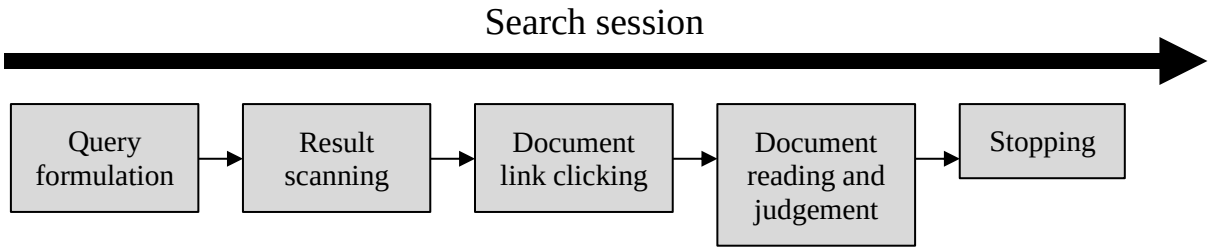


Figure 2. *Subtasks in a search session (Baskaya, Keskustalo & Järvelin, 2013).*

Traditionally, user behavior has been ignored to evaluate the effectiveness of IR systems (Moffat et al., 2017; Ellis, 1989) despite a variety of factors influencing user behavior, including whether the user encounters relevant documents, the user's stage in the search process, or the task difficulty (Moffat et al., 2013). Variability in users' expectations and individual intentions has also been crucial for interpreting searching behaviors. Therefore, these factors should be considered during system assessment (Moffat et al., 2017).

2.2 Rankings on SEs

SEs contribute to simplifying people's decision-making processes (e.g., Beer, 2016; Latzer et al., 2016; Steiner et al., 2020). The feature responsible for this simplification is the rankings that display the results estimated to be most relevant to a given search query from a plethora of web pages available online (Pan et al., 2007; Aggarwal, 2018). Ranking is a natural step of IR (Baeza-Yates & Ribiero-Neto, 1999) which gives users access to information by avoiding having humans making individualized choices regarding document relevance (Gao & Shah, 2020; Goldman, 2005). In other words, the role of a ranking algorithm

is to make choices, classify, sort, order, and rank, meaning that the algorithm decides "what matters" and "what should be most visible" (Beer, 2017, p. 6).

A principle behind rankings is that every IR system will return a ranked list of documents. It is a common approach in academia to evaluate the quality of IR systems' rankings through so-called test collections, consisting of search tasks made through synthetic users designed through user models, used to interpret how well a ranked document satisfies a particular query. Experts embrace this evaluation style in the field of IS, known as the TREC evaluation framework, which is more than half a century old (Harman, 2012). However, this widely accepted framework has been pointed at for its lack of external validity. These evaluations are hard to transfer to real-world circumstances, where (1) users are very different and range from stay-at-home parents to retired intelligence analysts; or (2) the interactions vary depending on the task the user is involved in (Moffat et al., 2017).

Rankings have also been evaluated by experts regarding the substance of the presented content. It is known that the presentation of the ranked results on commercial search engines such as Google is neither neutral nor fair (e.g., Gao & Shah, 2020; Bilić, 2016; Goldman, 2005). These findings suggest that the semantic ranking quality is a phenomenon of interest for academia, as rankings can induce cognitive biases (Gao & Shah, 2020).

Furthermore, the algorithm responsible for rankings on SEs such as Google's is updated regularly. Hence, rankings are not permanent but rather dynamic (Ziakis et al., 2019). Therefore, studies evaluating ranking configurations are likely to have validity concerns due to rankings' fluctuating nature and shifting algorithmic influence discussed in the following section.

2.2.1 Algorithmic influence

IR systems and, by extension, SEs, connect people in a variety of contexts such as e-governance, agriculture, education, information, etc. (Madankar, Chandak & Chavhan, 2016), and have set "expectations for billions of users" (Rosenberg, 2018, p. 29). The underlying SE algorithms have become an apparatus of social order and crucial tools of people's daily routine (Ziakis et al., 2019). They are seen as a life simplifier: They can recommend music, books, travel plans, or sources to get information by avoiding information overloads (Latzer et al., 2016).

The concept of an algorithm is ambiguous. They can be framed as "lines of code", as well as "social processes" with societal impacts (Beer, 2017, p. 4). For the aim of this study, the latter approach will be favored, even though the underlying technical logic is neither ignored.

Societal algorithm effects

The more people are exposed to systems run with an algorithmic logic, the more likely they are to be socially accepted (Shin, Zhong & Biocca, 2020; Rosenberg, 2018). It has become so natural to make choices relying on the criteria of algorithms that they have also been referred to as a tool of "social order" (Latzer et al., 2016 p. 1). Therefore "it is far more common for algorithmic processes to pass us by without being noticed" (Beer, 2017, p. 2).

Mager (2012) goes a step further and criticizes how SE algorithms, particularly Google's, have impregnated specific ideological beliefs in today's society. She defends that a "techno-fundamentalist" ideology has aligned with the capitalist ideology to sustain today's SEs business models. In that sense, SEs have constructed a dependency system for both businesses and individuals to generate profit through a "capitalist spirit." This ends with the situation where the algorithm is fiercely defended by the company's engineers and marketed to end-users by perpetuating an algorithmic ideology (Mager, 2012, p. 775- 782).

The relationship between humans and algorithms

Algorithms do not only affect the economic and social order, but they also influence people on an individual level. Users have a "romanticized view" towards SEs. Despite that, it is known that SEs control searcher behavior and users' perceptions thanks to this algorithmic logic. Although algorithms are powered by Machine Learning tools that are said to serve the public neutrally, SEs should rather be seen as media companies that are forced to make editorial choices when indexing and ranking pages for the user (Goldman, 2005, p. 189).

Hence, algorithms have the power to suggest which facts and information sources should be perceived as relevant by SE users (Beer, 2017). Simultaneously, algorithms are human inventions, meaning that human bias is inherited in the system's code that makes choices (Caliskan, Bryson & Narayanan, 2017; O'neil, 2016). Hence, a human actor will always be involved in deciding which factors to include in the ranking algorithm, even though they are still boldly promoted as neutral tools, guaranteeing the best possible output to the user's query (Mager 2012; Goldman, 2005).

A result of that is that users have been found to trust Google's selection criteria when performing searches, leaving their own evaluations aside. Google's algorithm reduces users' cognitive effort by sorting results through ranking mechanisms, depending on whether they can fulfill the user's information needs. In that sense, Google contributes to reduce mental effort and induce trust in the system, and users hesitate to question the selection of search results (Pan et al., 2007; Schultheiß & Lewandowski, 2021).

Furthermore, SE algorithms are designed to provide intercultural interconnectedness by allowing access to the Internet. Therefore, they also affect people from different cultures differently. The literature proves that when users get services in local languages, they have been majorly accepted and used, so that is why SEs such as Google have adapted its algorithm to specific publics and regions (Madankar, Chandak & Chavhan, 2016). Despite that, it is still known that specific groups of users are underserved. These differences mainly depend on demographic factors such as the user's age, gender, or location (Mehrotra et al., 2017).

2.2.1.1 SEO culture

Another consequence of the rise of algorithmic influence is search engine optimization (SEO). This widely extended marketing strategy leverages SEs' algorithmic logic to affect the economic order (Latzer et al., 2016). SEO refers to the study and implementation of techniques that allow websites to appear more valuable on SEs through higher rankings. Companies worldwide have learned to master the game of SEO to increase their exposure and revenue (Ziakis et al., 2019). Many different businesses are prone to play under SEs' rules and are trying to decipher Google's algorithm to get a more detailed understanding of how to improve their position on rankings (Bilić, 2016).

SEO is no longer seen as an option but instead as a fundamental marketing tool for ensuring a website's visibility (Schultheiß & Lewandowski, 2020). The activity of SEO is sustained thanks to a SEs system of punishments and rewards to boost a websites' position on Google's ranking (Röhle, 2009). Hence there is a tendency of trying to "please" Google (Bar-Ilan, 2007b, 163). Consequently, Google is currently shaping how online content is being produced globally, affecting users' search results and information habits. In turn, SEs will inevitably determine what should be considered to be a positive experience (Bilić, 2016). This type of information is commonly used among SEO and content marketing specialists to create content on their websites. However, studies have proved that quality standards are left behind when relying on Google's ranking mechanisms and that unreliable information can emerge as top results (Modave et al., 2014).

Experts argue that end-users do not necessarily understand SEO activity and might see it as a spamming technique, and that the information literacy of users affects how users interpret rankings and their trust towards the SE (Schultheiß & Lewandowski, 2021; Schultheiß & Lewandowski, 2020).

Optimizing websites

A SEO specialist's role is to discover what aspects of a website need to be optimized to guarantee a website's presence when a specific query is entered (Ziakis et al., 2019). Ziakis et al. (2019) investigated 24 SEO factors that would increase positioning on Google that were identified in previous literary works. Their study aimed to conclude which of those factors would be better contributors to ensure higher positions. They found the quantity and quality of backlinks to be a crucial factor to increase positions on rankings, followed by the presence of an SSL certificate and the bounce rate. The most mentioned factors in earlier literature included the quality and quantity of backlinks as well. However, social media support, the keyword's presence in the title tag, the URL length, and the website structure were also prevalent according to their updated research. This inconsistency among Ziakis et al.'s (2019) findings and previous theories proved that it is crucial to remain updated on the SE's algorithm altering, as it tends to change often (Ziakis et al., 2019, p. 10).

Google has also declared that they regularly add new ranking factors. Some of the newest ones have to do with mobile-friendliness or loading times (*Evaluating Page Experience for a Better Web*, 2020). Hence, studies on variables that influence SEO have shown to lack validity by nature, as the object being studied is modified throughout time.

2.2.1.2 The Page Rank algorithm

The Page Rank algorithm is the algorithm that drives Google search results which is able "to sort and prioritize the media we encounter" (Beer, 2017, p. 2). Hence, it is responsible for creating the rankings seen on the Google SE. The algorithm calculates the online popularity of a page by using a crawler that counts the number of links that points to it, and the resulting score of each page is used to generate SE rankings (Bilić, 2016; Modave et al., 2014; Mager, 2012; Najork, 2009).

Why Google won the algorithm game

Google proposed the Page Rank algorithm as a solution to "an untraceable clutter of websites in the 1990s", completely wiping out the competition due to its superiority in usefulness (Bilić, 2016, p. 2). Thanks to the PageRank algorithm and a smart business strategy, Google rapidly became a dominant

player on the market (Mager, 2012). Google's algorithm exploits the collective intelligence on the web, as the popularity of a site is measured through a link representing a sign of intelligence used to create value (Pasquinelli, 2009). Therefore, the literature suggests that Google won the algorithm game by creating a technologically superior product that reduces the cognitive effort of the user (Pan et al., 2007).

However, a lot has happened since 1990, and there is currently an active competition among webmasters to rank high on Google (Bilić, 2016). Also, rankings keep changing for a given query, as new content is created, and existing material is continuously updated by users and publishers (Kim & Caravalho, 2011). This design allows for manipulation, leading to the use of SEO (Ziakis et al., 2019), or illicit techniques such as Google bombing, a concept that refers to an "active manipulation of the results in Google" by intentionally creating links that point to a site that the attackers want to be placed on high-ranking positions (Bar-Ilan, 2007a, p. 912). Google has therefore put effort into rethinking its algorithm. In 2016, experts estimated that Google used more than 200 factors besides Page Rank to determine a page's relevance (e.g., geographic location, previous search history, or recommendations on social networks). This action helps Google sell its neutrality speech and promote its service of a completely objective SE based on state-of-the-art technologies (Bilić, 2016).

It is not rare for Google to publish information about new algorithm updates and the intentions behind them. For example, from 2020, Google has put a lot of focus on explaining how they want their algorithmic structure to favor the UX by updating their ranking factors (*Evaluating page experience for a better web*, 2020). The next algorithmic update is planned for June 2021 with an update "designed to highlight pages that offer great user experiences" (*More time, tools, and details on the page experience update*, 2021, n.p.).

Bilić (2016, p. 3) demonstrates that ranking generation does not stop there and points at the so-called "human quality raters." In essence, these raters perform rating tasks by evaluating the quality, utility, and relevance of websites in a subjective manner. This phenomenon is described as "hidden labour" as Google has not been open about this role, but the author states that the so-called Search Quality Rating Guidelines have been previously leaked in 2004, which are used by human raters to evaluate rankings and adapt them according to Google's interests. In that sense, the PageRank algorithm is not the only actor responsible for ranking the results on Google, and it raises questions towards Google's business model and the company's tendency to make the algorithm responsible for any outcome, controversial or not, on SE rankings (Bilić, 2016).

2.3 User experience (UX) on SEs

Ensuring that the people behind the screens get an excellent experience and are willing to stay loyal to the brand is key to sustaining Google's current business model, as users are labeled as Google's "goldmine" (Mager, 2012, p. 776). Google themselves have been transparent about their objective of offering "great user experiences" (*More time, tools, and details on the page experience update*, 2021; n.p.). Therefore, this chapter provides an overview on user experience (UX) as a construct and assesses its relevance in the context of SEs.

The definition of UX is still a recurrent discussion in the academic field, but some widely accepted definitions have emerged. One of them is the ISO definition, that sees UX as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" (International Standards Organization, 2010). However, literature has not agreed on a way to measure UX, as the concept is associated with terms that are hard to operationalize such as emotion, effect or aesthetics (Law et al., 2008; Law & Van Schaik, 2010). Some authors have overcome this challenge by bringing across reliable measurements of the construct. Lachner et al. (2016) proposed an instrument for measuring UX based on 9 dimensions they designed based on data from qualitative interviews with UX experts. When used as an instrument, values between .74 and .96 were reported as Cronbach's Alpha. Despite not presenting any validity measures, UX experts have therefore managed to provide a reliable measurement instrument (Lachner et al., 2016). Despite that, it is crucial for UX measurement instruments to be adapted to the application domain being assessed instead of using a standardized scale for a variety of products (Lallemand & Koenig, 2017; Bernhaupt & Pirker, 2013)

Google as a reference experience

The UX on SEs has been defined as a "reference experience." Users are accustomed to specific UX patterns that have become socially accepted thanks to popular products and services. Google serves as an example, as users have learned and accustomed themselves to using search features the same way they use Google. Hence, the internet society has embedded searching behaviors into their everyday lives (Rosenberg, 2018, p. 29). As Gillespie (2014, p. 187) puts it, SEs such as Google are tools "that billions of people use every day, most of whom experience it as something that simply, and unproblematically, works." SEs have been labeled as habit-forming products that can trigger their usage when people feel in a particular manner (in the case of SEs, uncertain about something) (Liu & Li, 2016).

How rankings enhance the experience

SEs enhance the UX by connecting user intentions with search results that are organized in rankings. In that sense, they appear as "displayed estimates of utility" related to the users' intentions (Bilić, 2016 p. 4). Given that IR systems can retrieve many documents for a search query, the user cannot examine them all. Hence, UX on SEs should be evaluated on the system's ability to rank the most relevant documents for the user at the top of the results list according to the user preference (Zhou & Yao, 2010).

Feedback loops on SEs

Modern SEs also incorporate a rewarding system for users to gain a "sense of satisfaction" after searching (Liu & Li, 2016). These rewarding systems can be strengthened through feedback loops, facilitating the communication between the user and the system. These feedback mechanisms become crucial patterns for developing satisfactory experiences on SEs. The feedback loop is defined as an HCI that consists of inputting a query, the process of obtaining a text response to this query, the text of the response, and an interpretation of the appropriateness of the delivered text. This loop is user-initiated as soon as the user enters a query and consists of the user interpreting how adequate the system's response is to their information need (Spink & Saracevic, 1998).

Device-dependent UX on SEs

State-of-the-art literature has also placed mobile devices in the spotlight to evaluate search experiences, given the increasing use of smartphones and tablets, that account for more than half of web traffic worldwide (StatCounter, 2020). Responsive sites (i.e., sites that can adapt their layout to the viewing environment, in this case, smaller screens) contribute to higher SE rankings and a better search experience (Ziakis et al., 2019). Google is taking a "mobile first" approach, meaning that the algorithm favors these pages offering the best mobile experience (Strzelecki & Rutecka, 2020, p. 9). In that sense, Google has declared they have added UX criteria as "mobile-friendliness" as a factor for ranking results (*Evaluating Page Experience for a Better Web*, 2020).

What is particularly remarkable about mobile devices is that they are used to access the Internet in various scenarios, especially during "on-the-go search tasks." People use mobile devices in situations where all the attention is not put solely on the search task (e.g., walking), which increases the cognitive load on the user (Harvey & Pointon, 2017, p. 293). Therefore, access to information must be simplified, as search behavior changes depending on the screen being used (Kim et al., 2017). These different behaviors on desktop and mobile devices should be considered when measuring the UX on SEs, as search behaviors

are device dependent. For example, desktop users view and click on more results and issue more queries for satisfying their information needs. In contrast, mobile users achieve higher search accuracy for tasks with increasing numbers of relevant results (Ong et al., 2017). Mobile users are also more susceptible to ranking effects and have a higher tendency to click on top results; and are also more prone to click on results that appear geographically close to their location (Ghose, Goldfarb & Han, 2013).

In parallel, Google favors voice search, which consists of search results read aloud by voice assistant devices such as Google Home. In that sense, it promotes the presence of featured snippets for a wide variety of queries, as it becomes more legible both for mobile users and these systems (Strzelecki & Rutecka, 2020).

2.3.1 Searcher satisfaction

The literature has also focused on the concept of searcher satisfaction to evaluate the UX on SEs. Many authors have focused on exploring the construct after a task has finished (Al-Maskari, Sanderson & Clough, 2007; Fox et al., 2005; Hassan, Jones & Klinkner, 2010). Hence, satisfaction is understood as the fulfillment of the desired information need (Kelly, 2009).

The difficulty of assessing searcher satisfaction is that it is difficult to measure (Mehrotra et al., 2017). Authors have used signals such as clicks and an extended dwell time, considering they were correlated with satisfaction (Fox et al., 2005). Edwards & Kelly's (2017) findings challenge this view, as they have positively correlated clicks and increased behaviors on SEs with searcher frustration.

Some authors have focused on evaluating SE performance through individualized preference metrics and, specifically, considering each document's relevance for each user. Hence, a well-performing SE can retrieve the most relevant documents for a particular person (Zhou & Yao, 2010). According to these findings, satisfaction is an individualized factor.

The display device can also affect the experience, and Lagun et al. (2014) focus on scrolling and viewpoint features to understand user satisfaction on mobile devices. However, these signals often rely on demographics. For instance, older users might read slower than younger users (Mehrotra et al., 2017; Wolfram & Xie, 2002).

An appropriate way to assess searcher satisfaction is to analyze the user's explicit judgement that determines if the desired information need is fulfilled when a search task is finished (Vuong et al., 2019).

Difference between searcher satisfaction and UX

Searcher satisfaction and UX might appear to be similar constructs, and there is a risk of using them interchangeably. Based on Hassenzahl & Tractinsky (2006), UX focuses on the context of use, and the experiences of the user during the use of a product. Differently, user satisfaction focuses on the state of the user after the experience and wants to understand users' state when a search task has finished (e.g., Al-Maskari, Sanderson & Clough, 2007; Fox et al., 2005; Hassan, Jones & Klinkner, 2010). However, according to Lachner et al.'s (2016) dimensions of UX, assessing the outcome satisfaction of a product is relevant for assessing the overall UX. Hence, it can be interpreted that in the context of SEs, searcher satisfaction contributes to measuring the UX.

2.4 Search engine bias (SEB)

One of the major research interests in IS and sociology is exploring to what extent SE rankings are biased, and how this phenomenon is overcome. Particularly, experts are concerned with the concept of search engine bias (SEB). SEB started gaining popularity in research at the beginning of the 21st century. It refers to ranking algorithms favoring certain types of content in front of others due to the SE's editorial choices. Hence, commercial SEs offer skewed search results by nature (Goldman, 2005).

These biases on the SE's side affect the people using SEs and the sources of information they are exposed to (e.g., Mager, 2012; Steiner et al., 2020). SEs rank web pages optimized with SEO techniques higher, allowing low-quality pages to be more visible to Internet users. On the other hand, high-quality information from governments, medical organizations or universities are left in lower-ranking positions (Modave et al., 2014).

When exploring bias on SEs, studies have detailed the differences on "indexal bias" and "content bias." The former refers to the snippets on the rankings containing information such as titles or descriptions that can be biased. The latter refers to the substance of the content itself from a selected document (Mowshowitz & Kawaguchi, 2002). "Indexal bias" is the bias analyzed in this current study.

Consequences of SEB

Experts discuss whether SEB is a positive consequence of the SEs algorithmic structure or whether it needs to be corrected. In Goldman's eyes (2005, p. 188), SEB is a "beneficial consequence" of SEs, as it allows to optimize content to users. At the other end of the spectrum, more recent studies label SEB as a "threat" (Steiner et al., 2020, p. 3); "troubling" (Haim et al., 2018, p. 339), or something that users should be alerted of (Epstein et al., 2017). Goldman (2005) believed that the problems associated with SEB would disappear, but state-of-the-art literature indicates the opposite (e.g., Steiner et al., 2020; Gezici et al., 2021; Baeza-Yates, 2018).

Studies have focused on understanding to which extent bias is present and how much it simplifies decision-making processes (e.g., Beer, 2016; Latzer et al., 2016; Steiner et al., 2020). Bias that remains uncorrected leads to cognitive biases, and a skewed layout of search results on rankings affects "users' credibility judgment, selection making, and belief and attitude shaping of information" (Gao & Shah, 2020, p. 2). For instance, search results polarized towards a particular political option can change the voting preferences of undecided voters by 20% (Epstein & Robertson, 2015). Also, Treen, Williams & O'Neill (2020) list Google as one of the responsible actors for misinformation regarding climate change due to its ranking mechanisms by directing users to low-quality content. On April 1st, 2020, Google responded to this issue and declared that they would stop funding organizations that are climate change deniers (Pichai, 2020), but academia has highlighted other topics such as politics or health advice where bias is still present and nor Google or other SEs have acted on the issue (e.g., Modave et al. 2014; Epstein & Robertson, 2015; Gao & Shah, 2020).

Trust makes users blind to SEB

Users do not only trust the SEs ability to rank but also the content provided on top search results. This phenomenon has been described as "trust bias" (Joachims et al., 2017, p. 4; Pan et al., 2007).

Experts see trust bias as a regrettable SE outcome, as top results usually consist of big sites that get even more significant by Google's rewarding mechanism, whereas "smaller, less affluent, alternative sites are doubly punished by ranking algorithms and lethargic searchers" (Joachims et al., 2017, p. 4). This pattern favors a winner-takes-it-all market. Also, users also face the "vicious cycle of bias on the web," meaning that they are more likely to interact with high rankings. Clicking on high rankings leads to more clicks on top-positioned content providers and to the situation where "the richer get richer." However, the

consequence is that this only makes it more challenging to distinguish quality pages from flawed pages (Baeza-Yates, 2018, p. 56).

2.4.1 Stance bias (SB)

The term "stance bias" (SB) has been previously used in literature in the context of web browsing and online information consumption (e.g., Gezici et al., 2021; Johnson & Goldwasser, 2016; Ma et al., 2019). Despite that, the construct has not been defined formally. Instead, authors relied on the idea of "stance." The concept of "stance" is defined as being "in favor or against" a topic. "Stance" is not a synonym of "ideology," as the latter refers to "the specific ideological group as conservatives or liberals that supports the corresponding topic." (Gezici et al., 2021, p. 3).

Experts have presented examples of how SB can affect SE users. For instance, stances can emerge on SEs on searches related to controversial issues such as abortion, medical marijuana, or gay marriage (Gezici et al., 2021). However, topics that are not necessarily labeled as controversial can also induce SB. For example, in a scenario where a searcher would query coffee health, users should face both the benefits and harms of coffee (i.e., reasons to be in favor of drinking coffee and against drinking coffee). However, this is not the case with current SEs, where users could be presented with a featured snippet highlighting the benefits of coffee (Gao & Shah, 2020).

SEs do not systematically lean towards a particular stance (Gezici et al., 2021) despite SEs being media companies that "reinforce and perpetuate existing power structures" (Goldman, 2005, p. 193). It remains unanswered whether SB appears on SEs such as Google with the objective of enhancing the UX to reduce cognitive efforts during information seeking (Gillespie, 2014; Pan et al., 2007). In this case, Google can go a step further and leverage existing individualized data to display information that aligns with the user's preferences to introduce confirmation bias (Bilić, 2016).

2.4.2 Confirmation bias (CB)

Even though biases emerge on the SEs side, this can also induce biases affecting the user. The concept of confirmation bias (CB) refers to prioritizing the information that supports one's opinion or hypothesis in an unconscious manner (Nickerson, 1998). In an information-seeking context, it results in users browsing for information that supports their prior beliefs. This affects the user's assumptions on a wide range of topics, from food and clothing to health and politics (Suzuki & Yamamoto, 2020; White, 2013). The main threat of CBs is that it leads to the diffusion of misinformation online (Treen, Williams & O'Neill 2020).

Meppelink et al. (2019, p. 137) prove how users on the Internet choose to "expose themselves to belief-consistent information rather than disconfirming information when they seek online information." Traditionally, the effect of CBs has been attributed to the user itself. However, experts have pointed to SEs as responsible for this effect. SEs can intensify CBs "by generating results that consist only of confirming evidence for search contexts where disconfirming evidence is identified using different terms or phrases." Therefore, depending on the search queries issued, a system can predict which opinions or hypotheses should be favored for a particular user (Kayhan, 2015, p. 1).

2.4.3 Difference among fairness and SEB on rankings

"Fairness and bias are often considered to be two sides of the same coin," and the concepts can even be viewed as synonyms, depending on the area they are being studied (Gao & Shah, 2020; p. 3). Both phenomena have in common that their effects arise from the SE algorithm and impact the end-user and society (Mager 2012; Pan et al., 2007; Gao & Shah, 2020, Lewandowski, 2017).

Fairness in rankings is understood as to how different individuals and groups are treated by SEs ranking mechanisms and their exposure. In contrast, SEB focuses on the content itself by favoring certain kinds of content "through the assumptions inherent in its algorithms" (Lewandowski, 2017, p. 7).

2.5 The UI on SEs

SEs wrap the algorithmic experience into a user interface (UI). A UI should be understood as the means a person uses to interact with content to accomplish a goal, in this case, a search task (Blair-Early & Zender, 2008). User interactions with the UI have, in some cases, been ignored in previous literature despite having an impact on IR effectiveness (Moffat et al., 2017).

Users behave differently depending on the UI they are presented with (Azzopardi, Kelly & Brennan, 2013). They prefer UIs where rules are simple, self-explanatory and intuitive on commercial SEs (Xie, 2004). SE interfaces are very diverse, such as the map interface from Google Maps, which has various input options such as "typed queries, drags, clicks" (Vuong et al., 2019, p. 1253), or many times simply consisting of a search bar for inputting queries (Khoo & Hall, 2012). The UI on SEs is also becoming more user-centric thanks to features designed to help the user through a search task. For example, SEs can now correct misspellings or suggest queries (Bailey et al., 2015).

SE interfaces tend to be very similar and based on a textbox where users can input a query that they believe would satisfy their information needs. The SE returns a rank with the relevant documents to the query in the form of snippets, as described in Spink & Saracevic's (1998) feedback loop. This widely accepted UI logic affects the final user, as small changes to the layout can impact the behavior. For example, as the snippet length increases, users pay more attention to the snippet and less attention to the URL (Cutrell & Guan, 2007).

The UI on Google as a facade of user control

Google's UI has been widely accepted among SE users, who see it as something that "unproblematically works" (Gillespie, 2014, p. 187). When users were asked about their feelings towards altered rankings, where search results would differ from a typical search scenario, they would blame themselves instead of the system for the non-optimal results, saying they could not form the right queries (Pan et al., 2007). It remains unanswered whether confounding factors are intervening, such as the user's belief that the experimental procedure evaluates their ability to evaluate a system rather than the system itself, a common misconception during usability testing (Lauesen, 2005). Another plausible explanation to why users tend to blame themselves instead of Google when faced with unsatisfactory results is a consequence of Google's UI as a tool to dress up the underlying algorithmic logic. Google's UI forms "a facade of user control," and it contributes to "promote itself as an objective courier of online information empowering global internet users" (Bilić, 2016, p. 7).

How the UI adapts to new search trends

Interfaces on SEs need to adapt to new environments of use and to-go searches performed in circumstances where all users' cognitive effort is not put on the search task. UIs on SEs must adapt well to smaller screen sizes, such as mobile phones, which are more likely to be the chosen devices under spontaneous search tasks. The UI on smaller screens affects how people browse the web, as mobile environments are less "Internet-like." In that sense, ranking effects become higher on smaller screens, and location-based results become more relevant (Ghose, Goldfarb & Han, 2013, p. 614; Harvey & Pointon, 2017).

2.5.1 SERPs

The SE results page (SERP) is a core feature of a SE's UI. SERPs are crucial for web SEs, as they act as the gateway for users to interact with the search results to their given query (Bailey et al., 2010). The SERP is a relevant element of HCI, given its ability to influence the users' decisions (Lewandowski, 2017).

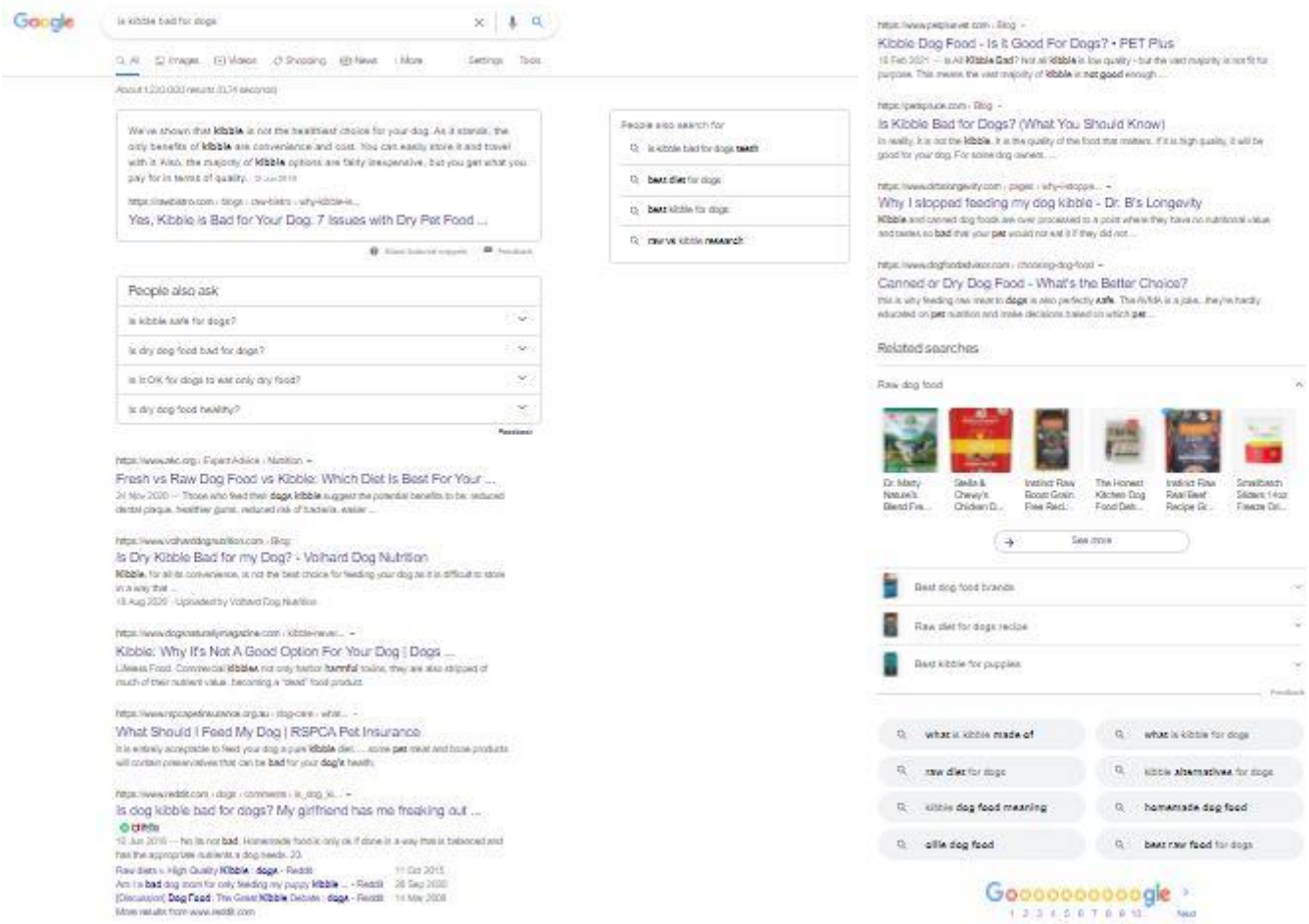


Figure 3*. Example of different elements appearing on SERPs. In this search, the user is faced with a featured snippet on the top, and diverse elements such as suggestions of follow-on queries, advertisements, or mini rankings.*

The typical SERP layout consists of each search result including a description, in some cases introducing multimedia, spelling corrections to the queries, suggestions of follow-on queries, advertisements, or mini rankings from other search features such as images, news, or video search (Bailey et al., 2010). SERPs commonly feature a textual listing of 10 results on each page and, despite many experts altering this layout for experimental purposes, the public accepts SERPs in a variety of different layouts (Kelly & Azzopardi, 2015).

The distribution of documents on SERPs also determines search behavior and is crucial for ensuring a positive UX. As noted by Lewandowski (2017, p. 14), commercial SEs "are able to lead users to certain

types of results merely through results presentation." They can do so through eye-catching layouts on the SERP, leading the user to click on them (Lewandowski & Sünkler, 2013).

SERPs have become richer over time. Commercial SEs such as Google have focused on enhancing the SERPs with new features. Richer search results contain semantic information, multimedia elements, recommendations, or location-based snippets. Although these can capture users' attention, rankings are still the most determinant factor for users to decide which document to click on (Marcos et al., 2015).

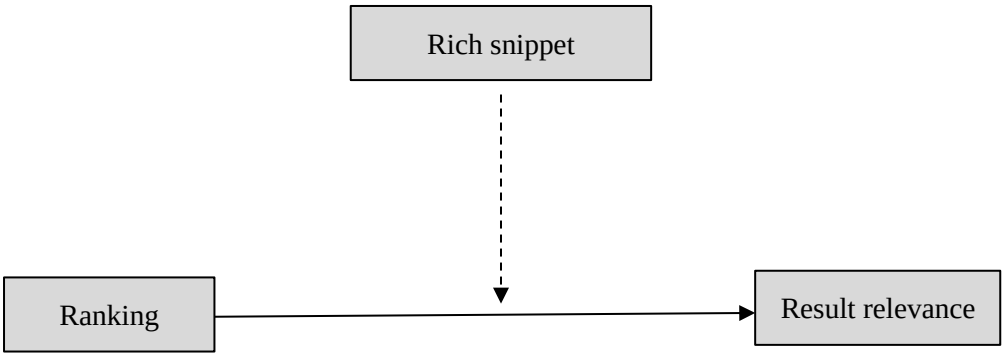


Figure 4. *Model introducing rich snippets as determinants of result relevance based on Marcos et al. (2015).*

A popular type of rich snippets is featured snippets (FSs). Zhang, Zhang & Wang (2018) describe a FS as an answer to a question extracted from an URL. The quality of snippets has traditionally been evaluated based on their ability to summarize a document (Ageev et al., 2013). FSs are usually taken from trustworthy websites such as Wikipedia and presented in a paragraph that can be read aloud by systems enabling voice search (Strzelecki & Rutecka, 2020). However, Gao & Shah (2020) show how stances can emerge on FSs as well by favoring one point of view over another.

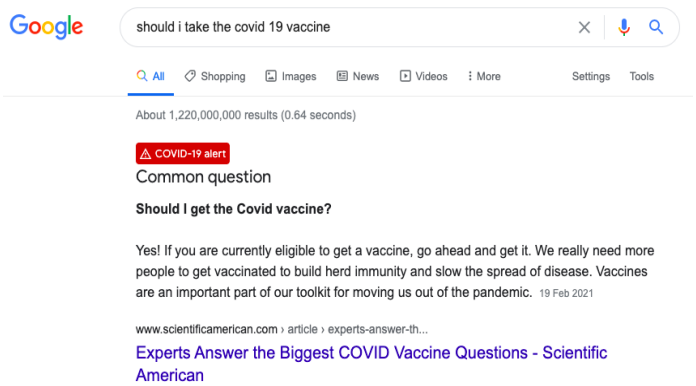


Figure 5. *Example of a featured snippet appearing on Google search results when querying "should I take the covid 19 vaccine". Retrieved on 11/02/2021.*

The literature has not focused on understanding how FSs impact the user. Previous studies have been rather observational (i.e., trying to understand how FSs appear), such as Strzelecki & Rutecka (2020). Hence, a significant limitation is that no theories illustrate how featured snippets affect the UX.

One crucial aspect to include when evaluating the SERP appearance is paid placements or search engine advertising (SEA), a Search Engine Marketing (SEM) technique used hand in hand with SEO. Paid placements on SERPs are purchased by sellers who "advertise and promote themselves on search engines." Paid placements on SERPs target buyers looking for information related to a specific product. Marketers prefer to invest in SEA instead of SEO, as the latter is proven not to be an "optimal SEM strategy" to ensure successful business results (Sen, 2005, p. 21-22; Jafarzadeh et al., 2019). Hence, advertisements are also a natural element on a SE's UI.

2.5.2 Whole-page relevance

In order to assess the UI on SEs', Bailey et al. (2010) came up with the concept of "whole-page relevance" (WPR), an evaluation defining how well a SERP and its holistic appearance satisfies the users' information needs. In other words, WPR considers how different elements such as rich snippets, conventional snippets, and paid results are distributed on the page to provide the best UX.

The technique used to determine WPR is using the School Assignment Satisfaction Index (or SASI method), which "focuses on assessing the user's experience". It is designed to capture the quality of a SERP holistically by assigning scores to the different elements on the page. The SASI method focuses on evaluating: (1) "the surface-level presentation on the SERP rather than the full content"; (2) "all components of the SERP rather than top-ranked algorithmic search results"; and (3) "the SERP components in context" instead of "each document in isolation."

3. Outline of the research design

The previous theoretical underpinnings chapter fulfills two objectives: (1) To gain a deep understanding of the concepts included in the RQ (rankings, stance bias, featured snippets, and UX); and (2) to provide a state-of-the-art holistic understanding of the study and use of SEs to detect factors to be controlled during the experimental setup. Including those topics in the theoretical underpinnings makes it possible to frame a valid study, as factors that could alter the experimental outcomes can be ruled out during the design of the research strategy.

Three different hypotheses are established in this chapter to answer the RQ: (*How does the presence of stance bias and featured snippets on rankings affect the user experience on search engines?*) based on the explored theory. These hypotheses build on the reasoning presented in this study's introduction. The fact that elements such as bias and rich snippets are encountered on rankings (e.g., Gao & Shah, 2020; Gezici et al., 2021) and the fact that Google invests time in updating its ranking factors with "user experience criteria" (*Evaluating page experience for a better web*, 2020, n.p.; Mager, 2012) are used to reason that there is a one-way relationship between Google's algorithmic choices on rankings and the UX.

It should be noted that Google focuses on measuring the UX on websites, not directly on rankings. Hence, when the SE refers to improving the UX, it only applies to the document clicking and document reading and judgement phase of a search session (Baskaya, Keskustalo & Järvelin, 2013). Hence, whether there is a direct link between bias and rich snippets and the measured UX on rankings is a novel link to be explored. Still, testable hypotheses can be formed by using abductive reasoning.

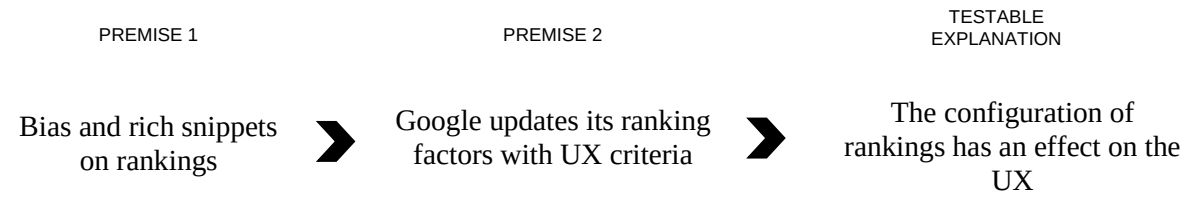


Figure 6. *Reasoning behind the established hypotheses.*

The following hypotheses are framed to guide the methodology:

H1. The presence of SB on rankings affects the UX.
H2. The presence of FSs on rankings affects the UX.
H3. The presence of both SB and FSs on rankings affects the UX.

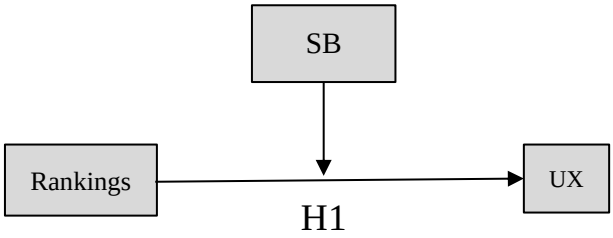


Figure 7. *Hypothesized relationship in H1.*

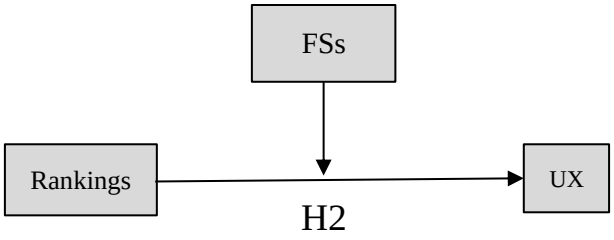


Figure 8*. Hypothesized relationship in H2.*

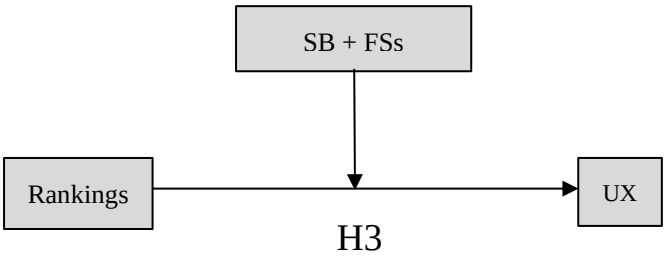


Figure 9. *Hypothesized relationship in H3.*

4. Methodology

This chapter describes the research design used to answer the established RQ (*How does the presence of stance bias and featured snippets on rankings affect the UX on search engines?*). The RQ will be answered through the assessment of the hypotheses defined in the previous chapter. These hypotheses have been developed by building on the existing theory and will be answered with hypothesis testing techniques (Anderson, 2001). To do so, empirical data is going to be collected throughout planned observations (Patten & Galvan, 2019). This chapter goes through all the steps that are necessary to collect and analyze the data and is distributed as such:

Figure 10. *Outline of the sections in the methodology chapter.*

4.1 Research purpose

The research purpose is explanatory and tries to complement existing research trends that have been mainly descriptive, exploratory and evaluative.

A wide range of studies on SEs and, by extension, IR systems were included in the theoretical underpinnings chapter. These works were created with different purposes, often depending on the field of study. Studies framed by experts from a sociological perspective resulted in very descriptive and exploratory natures, two research purposes that are often combined, as descriptive research is often a part of exploratory research (Saunders, Thornhill & Lewis, 2019). For example, research from the field of sociology provides an understanding of how SEs, ergo, the underlying algorithms that comprise them, have various societal impacts. Some examples would be SEB (Goldman, 2005), the perpetuation of a capitalist ideology (Mager, 2012), market dominance and user control (Bilić, 2016), or the promotion of certain points of view (Beer, 2017). Hence, previous studies have focused on developing a description of the role of SEs in today's society by providing details on why these have become tools of social order. This research has also been mainly qualitative. Some techniques used are in-depth interviews (Mager, 2012) or the analysis of secondary qualitative data such as Google's Search Quality Rating Guidelines or media reports (Bilić, 2016). These studies have also had a hard tone on the industry of SEs, particularly Google, which has received critique from authors in both the field of sociology and IS such as Mager (2012), Bilić (2016), Gao & Shah (2020) or Lewandowski (2017).

Studies from the field of IS tend to have two different objectives. The first one is to engage in descriptive research. Examples of descriptive studies are seen with Ong et al. (2017) or Ghose, Goldfarb & Han (2013), who provide patterns on how users interact with SEs from different devices. The same example is seen with Strzelecki & Rutecka (2020), who illustrate when featured snippets appear on SEs by analyzing secondary data. These studies are used to provide information about the phenomena being studied but do not explain why the results are reported as such.

Experts coming from the world of IS have also traditionally framed their studies as problem-solving tasks. They present evaluative research where new models are developed to solve SE issues that have received criticism, such as fairness on rankings (Gao & Shah, 2020) or lack of query precision (Baeza-Yates, Hurtado & Mendoza, 2004). Hence, there has been a trend for more evaluative studies that test the performance of different solutions the authors have designed (Saunders, Thornhill & Lewis, 2019). Despite that, the authors of these studies do not question the underlying reasons behind the status quo and

whether the architecture and outcomes of current algorithms are intentional or can be explained throughout extraneous variables.

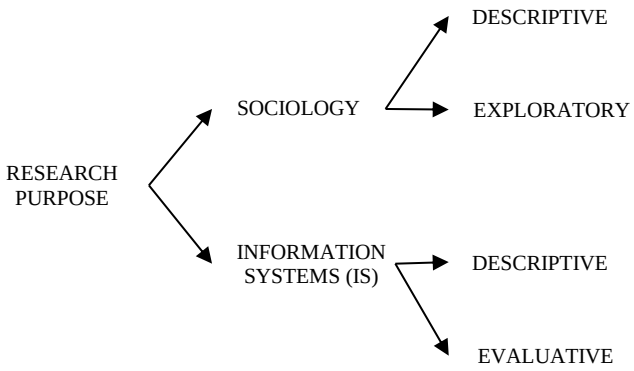


Figure 11. *Outline of the predominant research purposes in the fields of sociology and IS.*

This current research leverages the knowledge gained through descriptive studies both in IS and sociology. It also collects the critique oriented towards SEs from social scientists to frame an explanatory research. Hence, this work takes a step back from the classical evaluative approach in IS and builds on the state-of-the-art exploratory research describing SEs and, mainly, Google's rankings as a tool with implications on users. Therefore, this current research purpose is explanatory and attempts to define a relationship between the variables included in the hypotheses. Previous studies contribute to making this a theory-informed study (Waters, 2007; Saunders, Thornhill & Lewis, 2019).

4.2 Research philosophy

The term "research philosophy" refers to "a system of beliefs and assumptions about the development of knowledge" (Saunders, Thornhill & Lewis, 2019, p. 130). These beliefs and assumptions refer both to the nature of society and the nature of science (Burrell & Morgan, 1979). The underlying research philosophy will shape a study's research design and outcome (Crotty, 1988).

The nature of society refers to the idea that a researcher needs to define how societies evolve. When researchers attempt to explain why societies remain united and can be "maintained as an entity," they describe sociology as "regulatory." On the contrary, researchers that analyze why individuals are trying to emancipate themselves from societal structures explore the "sociology of radical change." These authors are "concerned with what is possible rather than with what is." Therefore, the sociologies of "regulation" and "radical change" are viewed as opposed views (Burrell & Morgan, 1979, p. 16-17).

When it comes to the nature of science, the researcher must pick an objective or subjective approach to refer to the ontology and epistemology. Objectivism and subjectivism are seen as two opposite schools of thought. The former refers to the idea of social actors being external to the researcher, leading to studies that use methods derived from natural sciences. In contrast, the latter refers to reality as a result of the perceptions of social actors, adopting a logic from the world of arts and humanities (Burrell & Morgan, 1979, p.16-17; Saunders, Thornhill & Lewis, 2019); Holden & Lynch, 2004).

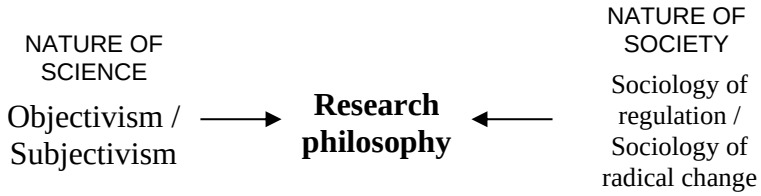


Figure 12. *Illustration on how the nature of science and nature of society define the research philosophy.*

The nature of science and the nature of society determine four paradigms of social science: Radical Humanism, Radical Structuralism, Interpretivism, and Functionalism.

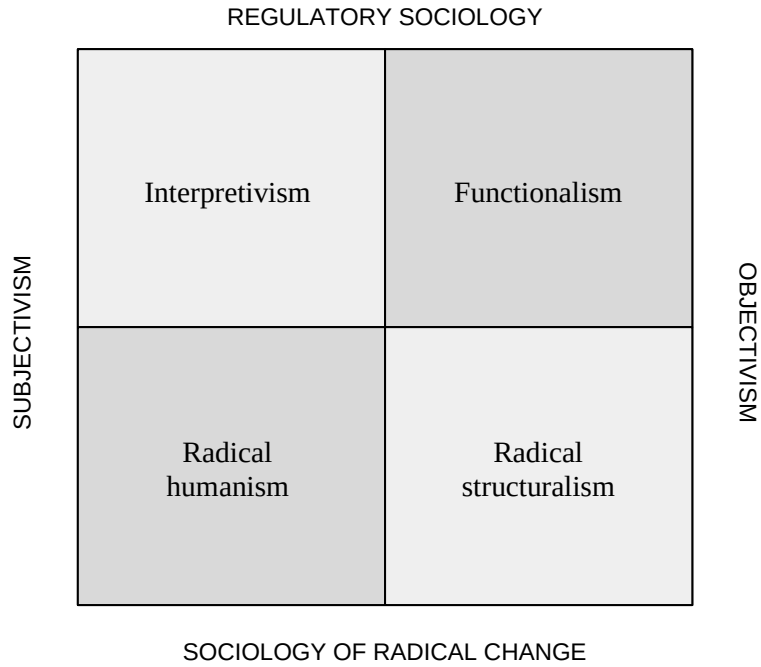


Figure 13. *The four paradigms of social sciences based on Burrell & Morgan (1979)*

Radical Humanism mixes subjectivism with the sociology of radical change, whereas Radical Structuralism relies on subjectivism and the sociology of regulation. On the other hand, Interpretivism

and Functionalism approach research from the sociology of regulation, but the first does it with a subjectivist approach, whereas the latter does it with objectivism (Burrell & Morgan, 1979).

A Functionalist approach will be the perspective used to frame this research. The reason for this is that previous studies from the field of sociology focusing on SEs have relied on the sociology of radical change, trying to question the societal implications of SEs and highlighting how the status quo can be challenged (e.g., Mager, 2012; Bilić, 2016; Beer, 2017). It is also typical for IS research to focus on the context within which technologies are used. In other words, IT artifacts such as SEs are seen as something static and independent used in a society that changes around these artifacts (Orlikowski & Iacono, 2001). Despite the constant evolution of SEs, this study assumes that SEs are an external entity that is defined objectively, and hence it must be highlighted that this is a cross-sectional study that wants to understand the impacts of SEs for the time being, assuming these are embedded and regulated into today's societal structures. In other words, this research intends to use a regulation perspective to understand SEs as tools of social order (Saunders, Thornhill & Lewis, 2019; Burrell & Morgan, 1979).

Objectivism is also used as an approach as it contributes to providing "rational explanations" and developing "sets of recommendations within the current structures." The objective is to explore the different constructs in the research from an external point of view, assuming these do not depend on individualized factors (Saunders, Thornhill & Lewis, 2019, p. 140; Burrell & Morgan, 1979). This work, therefore, relies on a Functionalist philosophy.

4.2.1. Research paradigm

As this study is framed from a Functionalist approach, it is common practice that the positivist paradigm will also be adopted. Hence, this research is referred to as "positivist-functionalist" (Saunders, Thornhill & Lewis, 2019; p. 141). Positivism is a widely and traditionally adopted research philosophy in the field of IS (Benbasat & Zmud, 1999). It is relevant to define what is meant by positivism in this study, as Crotty (1988) defined that this research philosophy can be categorized into many different varieties within itself.

In this case, positivism refers to the belief that unambiguous and objective knowledge can be achieved by using a method that allows gathering empirical data. Hence, to produce consistent findings, it is crucial for the researcher to keep neutral during the data analysis (i.e., using statistical tests instead of subjective interpretations) and collect it through methods that allow for neutral measurements and quantifying results and effects, as is typical in scientific studies. Another characteristic of a positivist approach is that it is

common to use theory to create hypotheses as it is done in this study (Saunders, Thornhill & Lewis, 2019; York & Clark, 2006).

When it comes to theory development, a pragmatist philosophy is also leveraged in combination with positivism. Whereas the positivist philosophy allows for reporting relationships through collected data, pragmatism allows putting the gathered knowledge into specific contexts to enable organizations to take action. A pragmatist philosophy empowers the researcher to focus on problems and phenomena detected in the literature (e.g., trust bias or the effects of SEO) and come with informed recommendations for future practice as a contribution (Saunders, Thornhill & Lewis, 2019).

4.2.2 Ontology

Ontology refers to the study of what is assumed to exist and the nature of reality or being. A positivist ontology is objectivist and assumes that reality is external, real and independent, and does not depend on any subjective perceptions or interpretations. It is also objective and universal, and it can be ordered and granular (i.e., separated into entities or things) (Saunders, Thornhill & Lewis, 2019). For example, in this study, the object of a SE ranking is not seen as something that is defined depending on the user. Even though the contexts and individuals using it can change, the SE ranking is external to those and is exposed equally to every individual. The same is true for UX. UX is a construct that is granular and is operationalized through different dimensions. Furthermore, even though the construct focuses on collecting user's perceptions and subjective responses towards a product, which a priori would seem inconsistent with a positivist philosophy, it should be noted that the objective is to externalize user reactions to be quantified and compared. Therefore, UX is not an individual-dependent construct, but it is product-dependent and used to quantify the "quality-in-use" of interactive solutions and, hence, it derives to "a common organizational understanding of a product's UX" (Lachner et al., 2016, p. 1; Hassenzahl, 2008; Law et al., 2009).

4.2.3 Epistemology

Whereas ontology refers to what is assumed to exist, epistemology is the study of knowledge and the process of understanding and theory development. In other words, epistemology focuses on understanding what means are used to constitute "acceptable knowledge." From a positivist-functionalist perspective, it becomes crucial to use the scientific method to generate knowledge, and facts need to be observable and measurable. In other words, it is essential to rely on numbers and quantifiable realities, which allows the researcher to establish explanations about a phenomenon (Saunders, Thornhill & Lewis, 2019, p. 133).

Therefore, this study focuses on measuring the constructs included in the RQ to report simple effects and interaction effects among them based on their impact on a dependent variable (UX). For instance, SB and FSs have been previously quantified in the literature and are easily measurable on rankings (e.g., Gezici et al., 2021; Gao & Shah, 2020; Strzelecki & Rutecka, 2020). UX is also a quantifiable construct, and authors such as Lachner et al. (2016) have reported reliable measures for the concept that can be used in various domains. In the case of this study, different measurements of SB and FSs are going to produce dependent measurements of UX that can later be used to report a quantifiable effect.

4.2.4 Axiology

Axiology is related to the values and ethics present in a research process, impacting the results. Therefore, this current research adopts a positivist approach and is designed to be value-free. The researcher does not make any value judgements on the researched topic and keeps an objective stance (Saunders, Thornhill & Lewis, 2019). To illustrate so, it is relevant to keep in mind the examples in the research purpose section, where the studies described were characterized for their critical tone, trying to manifest preconceived ideas and values on which actions should be undertaken to regulate SEs and the algorithmic influences of the technology. For example, Mager (2012, p. 769) advocates for "a shift of perspective" to "renegotiate search engines and their algorithmic ideology in the future," whereas Steiner et al. (2020) labeled the lack of diversity on SEs as a "concern" and they "call for media policy measures". Contrary to this, the current study has not relied on any values to set the research agenda but wants to critically assess the existing literature and contribute by establishing a relationship explaining how ranking configurations impact the UX. Even though the study design has not been based on any values or stances towards the constructs explored, a pragmatist and reflective approach is adopted in the discussion to assess the findings in a larger context (Saunders, Thornhill & Lewis, 2019).

4.3 Research approach to theory development

Positivist research typically has a deductive approach to theory development and relies on scientific methods to generate knowledge (Saunders, Thornhill & Lewis, 2019). Despite that, this research will use a combination of abductive and deductive reasoning to contribute to knowledge generation.

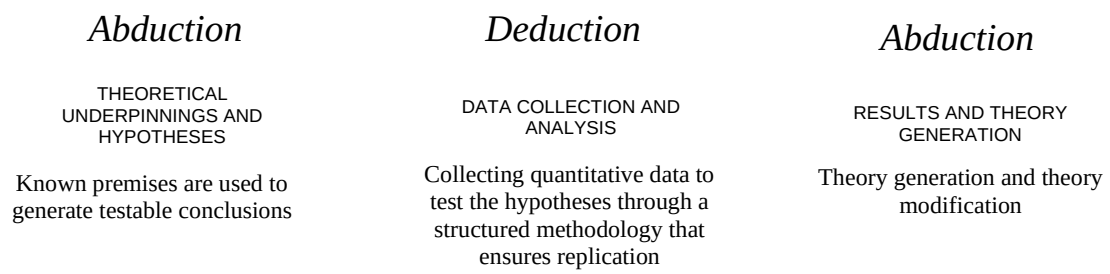


Figure 14. *Use of the research approaches in different steps of this work (adapted from Saunders, Thornhill & Lewis (2019) & Blaikie & Priest (2010)).*

Abduction is the chosen approach for framing this study. The principle behind abductive reasoning is observing empirical patterns that can be used to generate testable conclusions. In this case, the observed patterns are the presence of elements such as bias and rich snippets on rankings, knowing that Google simultaneously concentrates on updating its ranking algorithm to offer a better UX. Hence, it can be tested whether the configuration of rankings has consequences on the measured UX. It is known that the two first premises are true, but not whether the testable conclusion is true (Saunders, Thornhill & Lewis, 2019, p. 155).

Building on the positivist-functionalist logic, this study assumes that realities are granular and quantifiable (i.e., can be split into different entities that are countable). These entities were explored in the theoretical underpinnings to develop theory-driven hypotheses, which can be tested through a systematic data collection that allows performing an analysis that can assess the premises in a deductive manner with empirical data. Deduction is chosen at this stage as it complements abduction for testing plausible theories. An abductive approach is adopted again to build new theories, but also for modifying existing theories seen in previous literature (Saunders, Thornhill & Lewis, 2019; Blaikie & Priest, 2019; Van Maanen, 2011).

Abduction is not the chosen approach regarding the data collection, as it involves exploring a wide range of themes and patterns in detail (Saunders, Thornhill & Lewis, 2019) and the scope of this study is reduced to understanding how SB and FSs on rankings affect the UX. Hence, a deductive approach that tests the established hypotheses becomes more appropriate.

A deductive approach is also used for the scale development of the measurement instrument employed in the method, as the theoretical foundation provides tools to develop the items of a questionnaire. The primary advantage of basing the scale development and construct definition on a deductive approach is

that items are better suited to capture the domain of interest and are recommended for situations where theory exists about the domain (Hinkin, 1998).

4.4 Research strategy

Positivist research strategies tend to be deductive and focus on building a structured method. It is also common to have large samples that provide data that can be measured quantitatively (Saunders, Thornhill & Lewis, 2019). That is why a mono-method quantitative design was chosen to answer RQ. Quantitative research explains phenomena through the collection of numerical data. This data is later analyzed using mathematical methods (Creswell, 1994). Hence, data needs to be in a numerical form to be analyzed statistically (Sukamolson, 2007). Also, the variables being measured should be quantifiable (Goertzen, 2017).

The research strategy that was used to do so was a 2x2 full factorial between experimental design, where data was collected throughout a 7-minute-long survey. The objective of experimentation is "to study the probability of a change in an independent variable causing a change in another, dependent variable" (Saunders, Thornhill & Lewis, 2019, p. 190). A factor refers to "a controllable experimental variable" thought to influence a response variable (Kohavi, Henne & Sommerfield, 2007). In factorial experimental design, all combinations of the factors are investigated (Montgomery, 2017).

Experimental groups		
	No stance bias	**Stance bias**
No featured snippet	Control (C)	Treatment 1 (T1)
Featured snippet	Treatment 2 (T2)	Treatment 3 (T3)

Table 1. *Treatments in the four experimental groups*

An efficient experiment requires a scientific approach to planning, emphasizing its statistical design to ensure that valid conclusions are drawn from the collected data. Hence, "statistical methods are the only objective approach to analysis" (Montgomery 2017, p. 11). This approach is necessary to assess the established hypotheses.

The established hypotheses also determined the control and treatment groups of the experiment. The control group received no treatment, meaning in this case that neither SB on rankings nor FSs appeared during experimentation. The second group (T1) was treated with SB on rankings but no FSs. The third

(T2) treatment included FSs on rankings, but SB was omitted. Lastly, the fourth group (T3) received both SB and FSs as a treatment. As this was a between-subject experimental design, participants would only be allocated to one combination of treatments, meaning that they would only be in one experimental group. A between design was chosen to avoid range effects, which refer to the situation where participants exposed to different treatments will systematically compare the different scenarios, they are exposed to except the first. Hence, exposing the user to various treatments has psychological consequences that are difficult to detect and fix, which will affect the results (Charness, Gneezy & Kuhn, 2012; Saunders, Thornhill & Lewis, 2019; Poulton, 1973).

Vignette design

The current experiment was designed as a vignette study, a research design usually leveraged in social sciences that requires participants to judge hypothetical situations through surveys. Ageev et al. (2013) suggested that snippet evaluations, featured or not, should be task-based, considering how the snippets help users fulfill their information needs for a given search task. Therefore, the vignette experiment was designed to face users with hypothetical search tasks.

The experimental design also received inspiration from Bailey et al. (2010), who introduced whole-page relevance (WPR) and the SASI method. WPR evaluates the UX by presenting the user with a SERP to assign a score to different aspects of the site. One of its objectives is to assess the elements holistically instead of ranking each search result in isolation as it is typical in IR evaluation. Therefore, the questionnaire consisted of pictures of rankings that the user was asked to assess holistically through questionnaire items. The reason for referring to the shown pictures as rankings instead of SERPs is that many SERP elements are omitted, such as paid results or other forms of rich snippets with multimedia that would generally appear on SERPs (Bailey et al., 2010; Sen, 2005). Moreover, the SERP was also not shown in its entirety, so only the top results were displayed. Also, the UX on SEs should be evaluated based on its ability to rank documents and provide relevance to the user (Zhou & Yao, 2010), and, hence, the object of rankings gains more interest in this study's settings than SERPs.

The images of SE rankings the participants assessed were created to suit smartphone screens, given that Google is having a mobile-first approach that enhances FSs and simplified access to information. Ranking effects are also higher on mobile devices (Strzelecki & Rutecka, 2020; Kim et al., 2017; Ghose, Goldfarb & Han, 2013). The text of two real responsive SERPs was modified, one with a featured snippet and one without a featured snippet, to fit the content that was planned for experimentation. Paid results and other elements that are not relevant to the study were deleted from the modified SERPs. Three hypothetical

information needs were formulated for the experiment. Search goals were also considered, and all three scenarios made users engage with informational searches to achieve comparable results among cases (Vuong et al., 2019; Moffat et al., 2017).

The three search situations the participants faced were: (1) A scenario where the participant wanted to gain an understanding of a non-existent creature called “robot parrot;” (2) another where the participant had to understand more about a free time activity called “pupet;” and lastly, (3) the participant had to consider whether to drink or not a beverage they are offered called “Pure drink”. The entirety of the experiment can be consulted in Appendix 1.

4.4.1 Factor manipulation

In experimentation, the researcher must be able to manipulate the independent variables (Kohavi, Henne & Sommerfield, 2007). The factors manipulated in this experiment were the presence of SB and FSs on rankings.

Presence of stance-biased results

Following a positivist approach, the amount of SB was quantified in the experimental procedure (Saunders, Thornhill & Lewis, 2019). It should be noted that the bias of interest during experimentation was indexal bias, and not content bias (Mowshowitz & Kawaguchi, 2002).

Each case received a holistic SB score for each group, depending on the predominant point of view on the ranking. However, as ranking mechanisms affect decision-making (e.g., Beer, 2017; Latzer et al., 2016; Steiner et al., 2020), it became more appropriate to develop a proportional score that quantifies SB that considered ranking positions as well. As the experimental rankings maximum showed four search results, the bias on the first result was multiplied with 4, the second with 3, the third with 2, and the fourth received a value of 1. These calculations resulted in a proportional score illustrating the stance in each experimental case. If a particular stance was highlighted in an FS, a total of 1 point was added based on Marcos et al. (2015), who describe rich snippets as an influential factor for determining the relevance of a search result.

	Presence of stance				
	In favor	Neutral	Against	**Holistic score**	**Proportional score**
C / Case 1	0	4	0	100% neutral	**100% neutral**
C / Case 2	0	4	0	100% neutral	**100% neutral**
C / Case 3	0	4	0	100% neutral	**100% neutral**
T1 / Case 1	1	0	3	75% against	2 p in favor / 8 p against **80% against**
T1 / Case 2	3	0	1	75% in favor	7 p in favor / 3 p against **70% in favor**
T1 / Case 3	3	0	1	75% in favor	9 p in favor / 1 p against **90% in favor**
T2/ Case 1	0	4	0	100% neutral	**100% neutral**
T2 / Case 2	0	4	0	100% neutral	**100% neutral**
T2 / Case 3	0	4	0	100% neutral	**100% neutral**
T3 / Case 1	1	0	3	75% against	2 p against / 9 p in favor **81% against**
T3 / Case 2	3	0	1	75% in favor	8 p in favor / 3 p against **72% in favor**
T3 / Case 3	3	0	1	75% in favor	10 p in favor / 1 p against **90% in favor**

Table 2. *Measurements of stance in the different experimental groups and cases*

Presence of FSs

The FSs were included in the corresponding treatment groups (3 and 4) by modifying the UI as described in the previous section. To avoid confounding variables and ensure that only the effect of the featured snippet was being measured in these two groups, the content included on the rankings in comparison to the control group and T1 remained the same, but the first search result was simply modified into an FS.

4.4.2 Response variable and measurement instrument

A questionnaire was created based on the dimensions and items proposed by Lachner et al. (2016) to measure the response variable, UX. Not all the original dimensions from Lachner et al.'s (2016) study were included as they were not considered relevant for the experimental setup. Experts have previously challenged the idea of using standardized scales for measuring the UX. Instead, these should be adapted to the application domain (Lallemand & Koenig, 2017; Bernhaupt & Pirker, 2013). Therefore, the text in

the different items was also changed to fit the domain. For example, the word "product" was replaced with SE or Google. The entirety of the measurement instrument is found in Appendix 2.

Five dimensions were included in total and measured with a 5-point Likert scale, as 5-point items have been used before during UX measurements (e.g., Hussain, 2017; Hassenzahl, 2008), which makes it more feasible to compare the results afterwards. Dimensions are included based on the findings in existing SE literature.

Included dimensions		
Dimension	**Reasons for inclusion**	**Support in literature**
Communicated information structure (CIS)	This dimension refers to the SEs ability to provide clear navigation and that the provided information is understandable. An essential feature on today's SEs is their ability to offer a system that provides relevant information to the user in the form of rankings. Therefore, this dimension refers to whether the ranking logic makes sense to the participant.	Zhou & Yao (2010), Bilić (2016)
Visual branding (VB)	Visual branding refers to the user's reactions towards the brand, namely the trust, thoughts and feelings triggered whenever they are exposed to it. Experts have labeled Google as a "reference experience", and the brand has acted as a contributing factor to assessing the UX on SEs.	Rosenberg (2018, p. 29), Gillespie (2014)
Outcome satisfaction (OS)	Assessing the outcome and, in consequence, the searcher satisfaction is an extended practice in IR research. This dimension is an outcome assessment based on what the participant sees on rankings after finishing the document scanning.	Felid, Allan & Jones (2010), Kelly (2009), Baskaya, Keskustalo & Järvelin (2013)
Task effectiveness (TES)	The effectiveness of SEs has previously been categorized as a positive aspect of SE by users participating in experimental studies. This dimension also contributes to understanding whether an information need is met, which is a common evaluation among SE users.	Xie (2004), Spink & Saracevic (1998)
Task efficiency (TEY)	Experts focusing on mobile search argue that access to information should be "simplified", as search tasks are performed "on-the-go". As this study focuses on mobile search, the efficiency during result scanning should be considered. This dimension explains how quickly the participants fulfil a task.	Strzelecki & Rutecka (2020), Harvey & Pointon (2017), Kim et al. (2017)

Table 3. *Selection criteria of the five included measurement dimensions.*

Excluded dimensions	
Dimension	**Description**
Appealing visual design	Given that the shown images are manipulated rankings, and elements such as paid results or rich snippets are eliminated, the assessed visual design lacks external validity.
Mastery	This dimension refers to the use and control of a product. As participants are not asked to interact with the SE by querying or clicking on results, the items in the dimension cannot be assessed.
Emotional attachment	The items in the emotional attachment dimension also refers to the use of the product, and, therefore, the dimension is considered not to be adequate for the current case.
Stability and performance	It cannot be assessed as it considers errors and speed, which cannot be evaluated as participants are not interacting with the product.

Table 4. *Criteria for the elimination of the four excluded measurement dimensions.*

4.4.3 Control of confounding variables

To confound means that the effects of two factors are difficult to separate. Confounding variables are "extraneous but difficult to observe" which can alter the results (Saunders, Thornhill & Lewis, 2019, p. 191; Anderson & McLean, 2018). Whenever an experiment is conducted, the researcher needs to account for uncontrollable factors (z) and control them to avoid an undesired experimental outcome. The objective is to design an experiment that allows for manipulating only the factors of interest, so they are not affected by other factors (Montgomery, 2017).

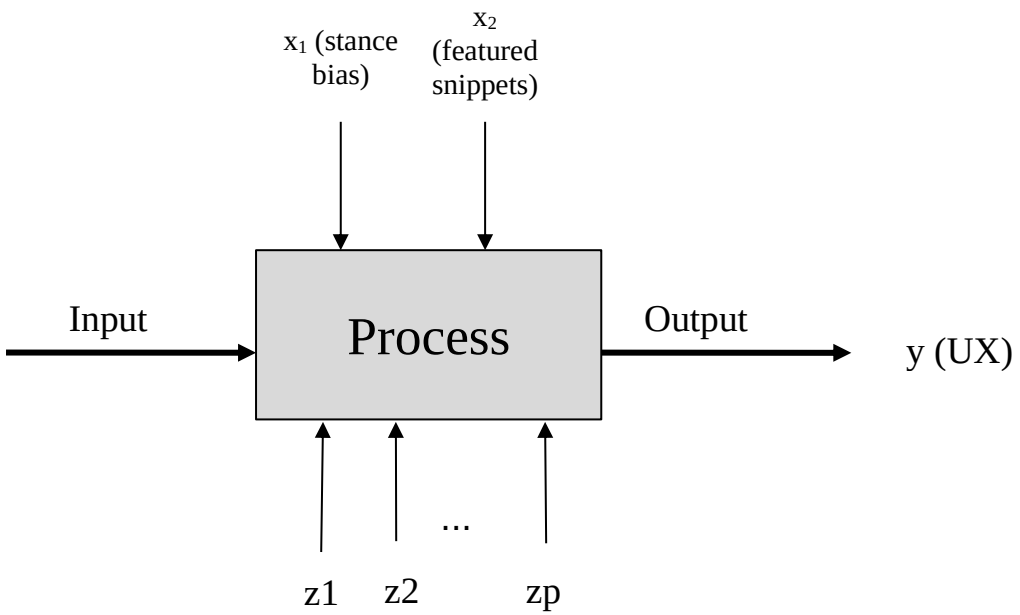


Figure 15. *Illustration on the factors influencing the experimental procedure. Every experiment receives an input (treatments), which are influenced by controllable factors and uncontrollable factors. This results in an output variable, in this case, the UX (adapted from Montgomery, 2017).*

The presence of confounding variables is reduced by exploring existing literature, making it possible to detect possible confounders in advance of the data collection (Rhodes et al., 1999). Hence, the technique employed to identify the maximum possible number of confounding variables has been through an exhaustive state-of-the-art literature review reported in the theoretical underpinnings chapter, that makes it possible to define extraneous variables that have the potential to influence the experimental outcome (Saunders, Thornhill & Lewis, 2019).

This is a complete list of the factors controlled during the experimental procedure to avoid confounding variables that might have altered the measurements of UX:

Controlled factors during experimentation			
Confounder	**Effects on the UX**	**Supported by**	**Impact reduction**
Variety in result types	A search that may cover any form of information: structured data, text, video, image, sound, music scores...	Grossman & Frieder (2012)	Eliminating types by including only text results.
Rich snippets	Rich snippets different from the FS can emerge on rankings and have different effects on the participants.	Marcos et al. (2015)	Including only FS as rich snippets.
Different types of IR systems	IR systems have different natures and are categorized differently depending on their characteristics.	E.g., Wolfram & Xie (2002). Gordon & Pathak (1999)	Using the same IR system throughout the whole procedure, in this case, Google.
Query cost, query variability, individual intentions & query-based IR effectiveness assessment	Querying comes at a cost for the user, which can be mental, physical, temporal, fiscal, or a combination of those. Furthermore, multiple query forms can represent a single information need, and a user may issue multiple queries in a search session to fulfill the same search task.	E.g. Azzopardi et al. (2013), Bailey et al. (2015), Pan et al. (2007), Vuong et al. (2019)	Suppressing the querying step to eliminate all the factors that are query dependent.
Task complexity	Success in information seeking depends on the search task, and its complexity determines the time spent on it.	E.g. Byström & Järvelin (1995), Wu et al. (2013), Wu et al. (2014), Pan et al. (2007), Baskaya, Keskustalo & Järvelin (2013)	Facing all participants with the same task to reduce differences in complexity and achieve comparable results.
Subtasks	Each search session includes subtasks associated with different behavioral patterns.	E.g., Baskaya, Keskustalo & Järvelin (2013), Moffat et al. (2017)	Reducing the search session to only result scanning. According to Moffat et al. (2017), assessing result

			scanning already gives information about the performance of a SE.
User variability	Not all users interact with a system equally, and individual factors affect the behavior on SEs.	E.g., Carterette et al. (2012), Moffat et al. (2017), Bailey et al. (2015)	Suppressing any interaction phase (i.e., querying and document clicking and exploration) so all participants would see the same results regardless of their searching habit and demographic conditions.
Contextual variability	The same user can interact with a system in different ways depending on the context of use.	E.g., Carterette et al. (2012), Vuong et al. (2019), Harvey & Pointon (2017)	Suppressing any interaction phase (i.e., querying and document clicking and exploration) so all participants would see the same results regardless of the environment in which they perform the experiment.
Task goal	The task goal refers to why the user has decided to search for something and determines whether the SE is successful at IR.	E.g. Vuong et al. (2019), Rose & Levinson (2004), Moffat et al. (2017)	Presenting the user with cases where the task goal is explicitly established so all participants evaluate the outcome under the same premises. All the information needs are posed as informational queries where the user looks for advice based on Vuong et al. (2019) so all three cases are comparable.
Substance domain	A task's substance domain refers to the field the task belongs to and affects the use of the SE.	Vuong et al. (2019)	Reducing the cognitive effort by facing the user with hypothetical tasks related to topics that do not require specialized knowledge: Pets, free-time activities, and nutrition.
Feedback loop	Spink & Saracevic (1998) defined the feedback loop as an HCI that consists of inputting a query, the process of obtaining a text response to this query, the text of the response, and an interpretation of the appropriateness of the delivered text. The objective of modern SEs is to give a user a sense of satisfaction after inputting a query, which can alter the rating of UX.	Spink & Saracevic (1998), Liu & Li (2016)	By removing the querying phase, the feedback loop was broken and, hence, non-existent in the experimental procedure.
Device-dependency	The device used for search also determines the experience quality. Google is taking a "mobile first" approach, meaning that the algorithm favors these pages offering the	E.g., Strzelecki & Rutecka (2020), Harvey & Pointon (2017), Ghose et al. (2013)	Even though the experiment could be conducted from any device, the rankings were all shown in a mobile format to achieve comparable results.

	best mobile experience. Behaviors are device dependent.		
Voice search	Google favors voice search, which consists of search results read aloud by voice assistant devices. Voice search is typically used for informational searches.	Strzelecki & Rutecka (2020), Guy (2018)	Suppressing the possibility of voice search in the experimental procedure made sure that every user experienced the system in the same manner.
Quality standards of search results	Studies have proved that quality standards are left behind when relying on Google's ranking mechanisms and that unreliable information can emerge as top results.	Modave et al. (2014)	Using imaginary topics so quality does not become a factor.
Google's UI	As Google is described as a habit-forming product that has a simple UI that users are familiar with, the use of the technology is simplified.	E.g., Rosenberg (2018), Liu & Li (2016)	To achieve results with external validity, Google's UI received the minimum manipulation possible once unnecessary SERP elements were removed.
URL importance	In some search circumstances, users pay attention to the URL to evaluate the quality of a search result.	Cutrell & Guan (2007), Modave et al. (2014)	The URL was suppressed so the user only focused on the experimental factors when assessing rankings.
Threats in usability testing	Whenever a user is asked to assess a system there is a risk, they misunderstand the task and blame themselves for undesired outcomes.	Lauesen (2005), Pan et al. (2007)	The experiment did not make the user interact with the technology to avoid the user assessing the rankings based on the outcomes of their actions.
Paid results	SEA has consequences for searchers and turns information needs into consumer desires through commercials related to the search query.	E.g., Sen (2005), Mager (2012), Taylor & Bicak (2020)	Paid results were omitted so only organic search results were assessed.
Fairness on rankings	Fairness on rankings refers to providing diversity in search results in relation to protected groups and the overall population.	E.g., Gao & Shah (2020), Lewandowski (2017)	Unfairness was avoided by designing hypothetical situations. As the topics were imaginary, they would not affect any population groups, ruling out the effects of unfairness.
SEB	Ranking algorithms favor certain types of content in favor of others due to their editorial choices, and, hence, commercial SEs offer skewed search results. SEB influences people's decision-making.	E.g., Goldman (2008), Mager, (2012), Steiner et al. (2020), Epstein et al. (2017), Gezici et al. (2021), Gao & Shah (2020)	Although bias is one of the controllable factors of this experiment, emphasis is put on manipulating solely the SB and avoiding other categories of bias such as ideologies.
CB	The concept of confirmation bias (CB) refers to prioritizing	E.g., Nickerson, 1998; Meppelink et al. (2019)	Making up hypothetical situations that participants

	the information that supports one's opinion or hypothesis in an unconscious manner. Users prefer exposing themselves to information that supports their prior beliefs, which could influence the evaluation of UX.	Kayhan (2015)	will not have formed an opinion on rules out the effects of CB.

Table 5. *Summarizing table on potential confounders and how they are transformed into controllable factors.*

4.4.4 Sampling method

The studied population in this research are individuals using Google as a preferred SE. Sampling was done in a convenient manner, where users were recruited through different Internet platforms such as social media, websites, forums, and chat services. In order to guarantee that the recruited participants were familiar with the assessed technology and part of the studied population, a filter question that ensured that the participants were active Google users was asked before being granted access to the experiment.

To achieve external validity, experiments should be designed with special emphasis on carefully choosing the survey samples to match the target population and create experimental designs that "motivate respondents to seriously engage with hypothetical choice tasks to mimic the incentives that they face when making the same choices in the real world" (Hainmueller, Hangartner & Yamamoto, 2015, p. 2400; Atzmüller & Steiner, 2010). This was done by ensuring that all participants were Google users to make the experimental procedure relevant and relatable.

In order to gain an understanding of the characteristics of sampled respondents, basic demographic information was collected at the last stage of the experiment to be able to outline potential participant attributes that could affect the results of the study and to give information about the respondents' characteristics (Toepoel, 2015).

4.4.5 Group allocation

Randomization was used to allocate participants in the experimental groups. This contributes to overcoming the effects of confounders (McNamee, 2003). Hence, randomization allows observations to be equally distributed for the statistical analysis and allows to ensure causality in the results by guaranteeing that the experimental outcome is not explained by differences in sample characteristics between groups (Montgomery, 2017; Kohavi, Henne & Sommerfield, 2007). Participants were allocated in the experimental groups with the randomization function in Qualtrics, the technology used for setting up the experimental procedure.

4.5 Collected data and variables of interest

A total of 11 variables attributable to each participant were collected in the survey-based experimental procedure:

Data collected from each participant		
Variable	**Description of the collected data**	**Type**
Control question	*Is Google your preferred search engine to perform searches online?*, 1 response per participant (0 or 1)	Discrete (binary)
Assignment group	Control, T1, T2, T3	Discrete
SB	Presence of SB during the experimental procedure (0 or 1)	Discrete (binary)
FSs	Presence of FSs during the experimental procedure (0 or 1)	Discrete (binary)
Mean overall UX score ***(Response variable)***	From 1 to 5. 1 response per participant.	Continuous
Score for each item	From 1 to 5. 45 responses per participant.	Continuous
Filter question	"Neither agree nor disagree" is the correct answer, which equals a value of 3. 1 response per participant.	Continuous
Gender of the participant	Male or female. 1 response per participant.	Discrete
Age group	8 different intervals. 1 response per participant.	Discrete
Educational level of the respondents	6 different groups. 1 response per participant.	Discrete
Other comments	Blank field for the participant to optionally share additional information that could become relevant for the data analysis.	Textual qualitative data

Table 6. *Summary of the collected variables.*

A common debate on Likert scales is whether values should be treated as continuous or discrete, as researchers have not agreed on a consistent way to analyze the scales' results (Brown, 2000). In this research, the 5-point Likert scale is seen as continuous and treated in the same way as the original questionnaire developers, who quantify UX by calculating the average score of a product's UX based on the average score of each observation's item assessments (Lachner et al., 2016).

The variable of interest during experimentation is the response variable (mean overall UX score), which was assessed during the data analysis to confirm whether SB and FSs influence the UX.

4.6 Data preparation and analysis

In this section, the chosen statistical test to conduct the data analysis, a two-way ANOVA, is explained; and the data preparation and analysis process is described and justified. Additional insights from the data treatment performed in the Jupyter Notebook software is found in Appendix 3.

4.6.1 Two-way ANOVA

An ANOVA (analysis of variance), proposed by Fischer in 1925, "compares the variance of scores between groups with the variance within the groups." If the former is larger, it can be concluded that the groups differ and, hence, are statistically significant. In other words, the null hypothesis (H_0) in an ANOVA test is that there are no differences between the means. Hence, a $p < 0.05$ accounts for a significant result, and this value is used as a significance level to assess the outcomes of the test (i.e., $\alpha = 0.05$) (Kent, 2015; Cardinal & Aitken, 2013).

A two-way ANOVA is the chosen statistical test for assessing the collected data in this research. The test is described as a multivariate and parametric technique used to evaluate whether two categorical variables (the presence or absence of SB and FSs) affect a continuous dependent variable, which in this case is the mean overall UX score for each participant (Kent, 2015).

The two-way ANOVA has been used in similar domains for evaluating experimental results, such as measuring SEB on different SEs (Mowshowitz & Kawaguchi, 2005), keyword frequencies on different SEs (Zhang & Dimitroff, 2005), or interface design and trustworthiness in IR (Kammerer & Gerjets, 2010). Common for those experimental studies concerning SEs is that they test how an independent continuous variable is affected by two categorical variables. In this experimental case, the first categorical variable is SB, which is classified as present (1) or absent (0). The second, being FSs, is also categorized in present (1) or absent (0).

4.6.2 Data exploration and preparation

The data from the experimental survey was downloaded as comma-separated values (.csv) files as four different documents corresponding to each group. The files were cleansed and analyzed with the Python programming language in a Jupyter Notebook.

Before looking at the variables of interest, the collected demographic data was explored to understand the configuration of the four experimental groups and the allocation of participants according to their age,

gender, and education. Insights on the participants were created to include in the discussion of this paper, which gives richer insights on the results.

The next step consisted of making the data suitable for analysis. Four different data frames were created in Python containing the treatment conditions as dummy variables, taking the values of 1 when a specific treatment was present. In contrast, the number 0 symbolized the absence of treatment. Several libraries were leveraged to treat the data and conduct the desired statistical tests, such as pandas, statsmodels, or pingouin (McKinney, 2011; Vallat, 2018; Seabold & Perktold, 2010).

To conduct a two-way ANOVA, the data had to guarantee the different assumptions required for the statistical test. These are the normality of the data, homogeneity of the variance, the independence of observations, and the same sample size between groups (Cardinal & Aikten, 2013).

Data distribution

To ensure normality in the collected data, the distributions were plotted with the help of a distplot method from the seaborn library, which illustrates the distribution of a variable in the data. Next, a Shapiro-Wilk test was performed with the scipy stats shapiro-method to test for normality in the data. The H_0 of the Shapiro-Wilk test states that the data is normally distributed. It can be assumed that the data is normally distributed when $p > 0.05$. The test complements the graphical representation of normality (Ghasemi & Zahediasl, 2012). The Shapiro-Wilk test has been recommended as the best choice for testing data normality (Thode, 2002).

Homogeneity of variance

The homogeneity of variance assumes that all groups have the same variance (Cardinal & Aikten, 2013). Bartlett's test provides information about the equality of variance in different groups. The test was performed with the scipy stats bartlett-method on the four experimental groups. A prior assumption for this test is that the data must be normally distributed. The test's H_0 is that all variances are equal, as seen in the following formula. Hence, when $p > 0.05$, the variances among populations are assumed to be equal (Wu & Wong, 2003).

Independence of observations

The independence of observations refers to the assumption that one observation should not be able to explain anything about another observation as they all are separate entities. Independence is achieved by

randomization in the experimental procedure, meaning that participants were assigned to different groups randomly (Cardinal & Aikten, 2013).

Equal samples in each group

To ensure the same number of samples in each group, samples are dropped from the most populated data frames to 34 observations, which corresponds with the number of subjects available in the least populated data frame.

4.6.3 Data analysis

Lastly, the data analysis was concluded by performing the statistical test of interest: A two-way ANOVA. The data was fit into an ordinary least squares (OLS) regression model, which is a statsmodels prerequisite to conduct a two-way ANOVA (Seabold & Perktold, 2010). The OLS method is used for predictive data analysis and minimizes the sum of squared residuals (i.e., the distances between the recorded observation and the observation line) which establishes a regression line that can forecast the values of a dependent variable given the values of an independent variable (Sharda, Delen & Turban, 2016). Using the formula.API module from statsmodels, an intercept is defined and OLS results are also displayed when performing the statistical test (Haslwanter, 2016).

The adjusted R^2 is reported to assess the fit of the model. This score reports whether the results in the predicted regression line are close to the observations, indicating whether the model correctly predicts the dependent variable. The adjusted R^2 takes into account the number of independent variables in the model. R^2 values range from 0 to 1, the first symbolizing a poor fit, and the latter indicating that the model produces exact predictions. A R^2 value of 0,3 can be considered a good enough fit in social sciences settings. In order to improve a regression model, explanatory variables can be added or removed from the model (Sharda, Delen & Turban, 2016; Argyrous, 2011).

When reporting the results of an ANOVA, showing only statistical significance is not sufficient. A score that quantifies the reported effect puts the results into a larger perspective (Yigit & Mendes, 2018). This is done with the partial η2 (eta-squared), which provides information about the effect sizes and to which extent UX can be explained through a significant independent variable (i.e., variance explained) (Vacha-Haase & Thompson, 2004). The partial η2 was reported through the pingouin library ANOVA method.

As “one-way-fits-it-all rules of thumb are not always very helpful in interpreting effect sizes”, it is important to report effect sizes considering the outcome that is being studied, in this case, how the reported treatment affects the UX on SEs. Also, in studies where it becomes impossible to compare the effect size with existing literature trying to report the same effect, the benchmarks described by Cohen (1968) should be leveraged, who described that a partial η2 close to 10% symbolizes a medium effect, whereas a partial η2 close to 25% symbolizes a large effect (Vacha-Haase & Thompson, 2004, p. 478; Cohen, 1968). This approach is leveraged in this data analysis, given that no studies allow for a direct comparison of the explored effect size, to the knowledge of the researcher.

4.7 Quality assessment

The validity and reliability measurements are used to assess the quality of this current research. The former refers to whether the used measures of a construct provide an accurate analysis and generalizability of the findings. The latter refers to the consistency of a research strategy, meaning that the method can be replicated, and the same findings would be achieved. Lastly, ethical principles are also discussed by reflecting on the ethical concerns appearing during the overall research process (Saunders, Thornhill & Lewis, 2019).

A construct is understood as a representation of something that cannot be directly observed. Hence, very abstract constructs are difficult to measure (Nunnally, 1978). In order to provide a valid and reliable analysis, the survey instrument must adequately represent the construct being examined (Hinkin, 1998).

4.7.1 Validity

Validity refers to measuring what is intended to be measured (Field, 2005). Data from a valid questionnaire will measure the construct of interest (in this case, UX) accurately (Saunders, Thornhill & Lewis, 2019). Several techniques have been employed to assess the validity of a questionnaire, such as face validity, content validity, and construct validity (Taherdoost, 2016).

External and internal validity

Experiments in social sciences are often performed in the context of a laboratory rather than in the field, making it possible to have control over the manipulated variables. Having the opportunity of directly manipulating the variables makes it easier to achieve internal validity, which refers to the extent which experimental findings can be attributed to the treatment instead of design errors in the research strategy. In order to account for internal validity, a control group and randomization should be ensured during the

experimental design. By having a control group and allocating participants in a randomized manner, alternative explanations to the treatment can be avoided. In that sense, the intervention becomes the only explanation for an effect on the outcome variable (Saunders, Thornhill & Lewis, 2019; Kohavi, Henne & Sommerfield, 2007).

Potential confounders are also detected through a literature review and considered in the experimental design to ensure that extraneous variables do not alter the experimental results to ensure internal validity (Saunders, Thornhill & Lewis, 2019; Rhodes et al., 1999).

In contrast, laboratory experiments also face challenges with external validity, as laboratory settings do not necessarily correspond with the external reality. Therefore, laboratory experiments such as the one performed in this current research tend to face problems generalizing the findings (Saunders, Thornhill & Lewis, 2019). Despite that, studies have shown that lab experiments have still been used as appropriate methods to generalize the findings into a larger context (Schulte-Mecklenbeck & Huber, 2003), and therefore, performing laboratory experiments does not necessarily mean that external validity cannot be guaranteed.

Particularly in the field of IR, authors have criticized laboratory experimental procedures due to their lack of external validity. A reason for that is that these studies do not consider the variability of users, as they are created synthetically, making it challenging to shift the findings to real-world circumstances. Another issue involves the nature of the search task, which can shift from experimental settings to real scenarios (Moffat et al., 2017). To increase external validity in the experiment in this research, real users have been used to collect data and the search tasks were clearly defined as real-life scenarios during the experiment that the participant would compare to real-world circumstances.

4.7.1.1 Face validity

Face validity refers to a "subjective judgment on the operationalization of a construct" and "evaluates the appearance of the questionnaire in terms of feasibility, readability, consistency of style and formatting, and the clarity of the language used" (Taherdoost, 2016, p. 29). The face validity was judged through a comment section at the end of the questionnaire, where participants had the option to leave remarks about the survey they just completed. No comments were included regarding the appearance of the construct in the questionnaire. Still, a single participant suggested randomizing responses followed by a statement: "I felt like I was almost getting into a flow of what I thought I'd respond." As no other comments were added

about the measuring instrument, it can be interpreted that the questionnaire has face validity, and that it became clear to the participants how to assess the construct.

4.7.1.2 Content validity

Content validity is applied while a new measuring instrument is developed and ensures that all the essential items are included. Its objective is to avoid undesirable items for measuring a construct (Taherdoost, 2016; Lewis, Snyder & Rainer Jr, 1995; Boudreau, Gefen & Straub, 2001). The content should be adapted to the studied domain, which in this case is the use of commercial SE rankings (Lallemand & Koenig, 2017; Bernhaupt & Pirker, 2013). Content validity is ensured by designing the research instrument through existing literature, based on Lachner et al. (2016), and suppressing the dimensions that were not relevant for this context as described in section 4.4.2. The items proposed by Lachner et al. (2016) were adapted to the context of this experiment, replacing the terms "brand" and "product" with "Google" and "search engine" to strengthen both the content and face validity.

4.7.1.3 Construct validity

Construct validity refers to which extent a construct has been transformed into a "functioning and operating reality." To assess the validity of the construct that concerns this study, UX, factor analysis is conducted. A factor analysis makes it possible to simplify a measurement and uncover patterns in a set of latent variables through mathematical procedures. Items with a loading > 0.40 can be considered to pass the threshold to be considered for further analysis (Taherdoost, 2016, p. 31; Child, 2006; Dawson, 2016). As a hypothesized model is established before the factor analysis, it becomes a confirmatory factor analysis (CFA). CFA "compares the actual relationships between items with the relationships that are suggested by the hypothesized structure," and, in that sense, CFA quantifies how well a hypothesized model fits with the observed data (Dawson, 2016, n.p.; Thompson, 2004). CFA is preferable when there is a belief of what items belong to a specific scale (i.e., typically in deductive scale development), which is the case of this experiment, where scales and dimensions were developed based on previous literature (Dawson, 2016; Hinkin, 1998).

The factor analysis was performed with the Stata software by building a model that links the different items with a single latent construct (UX) with 5 different factors. The analysis was conducted for each experimental case in each group, resulting in 12 analyses. Items with factor loadings under the cutoff of 0.4 were excluded for the data analysis based on Taherdoost (2016).

Confirmatory Factor Analysis (CFA) results												
	C-1	**C-2**	**C-3**	**T1-1**	**T1-2**	**T1-3**	**T2-1**	**T2-2**	**T2-3**	**T3-1**	**T3-2**	**T3-3**
TES1	.565	.740	.797	.531	.610	.681	.729	.681	.677	.593	.676	.743
TES2	.593	.845	.744	.861	.746	.790	.832	.917	.806	.798	.913	.948
TES3	.616	.804	.794	.905	.808	.842	.882	.905	.649	.775	.925	.886
TEY1	.497	.605	.685	.503	.644	.727	.649	.590	.682	.600	.622	.668
TEY2	.769	.602	.733	.626	**.380***	.867	.451	**.381***	.637	.487	.650	.554
TEY3	.571	.488	.597	.590	.426	.771	.509	.488	.706	.592	.596	.435
CIS1	.456	.706	.688	.481	.642	.601	.519	.547	.725	.605	.684	.694
CIS2	.522	.656	.688	.440	.820	.672	**.352***	.452	.663	.591	.601	.621
CIS3	.505	.680	.597	.495	.610	.581	.643	.515	.648	.569	.565	.713
OS1	.857	.756	.910	.681	.760	.774	.937	.933	.879	.846	.912	.927
OS2	.756	.834	.852	.635	.717	.706	.895	.925	.767	.783	.866	.902
OS3	.886	.868	.866	.703	.798	.853	.963	.877	.822	.858	.820	.936
VB1	**.366***	.475	.467	.412	.635	.550	.408	**.378***	**.384***	.595	.646	.684
VB2	.557	**.327***	**.208***	**.130***	.685	**.292***	**.137***	**.155***	**.194***	**.387***	.560	.727
VB3	.596	**.382***	**.119***	**.227***	.724	.454	**.159***	**.068***	**.157***	**.368***	.420	.688

Table 7*. Factor loadings for the different experimental cases in each experimental group.*

21 items were discarded for further analysis given their low factor loadings:

Discarded items for each experimental group	
Control	VB1-C1, VB2-C2, VB3-C2, VB2-C3, VB3-C3
T1	VB2-C1, VB3-C1, TEY2-C2, VB2-C3
T2	CIS2-C1, VB2-C1, VB3-C1, TEY2-C2, VB1-C2, VB2-C2, VB3-C2, VB1-C3, VB2-C3, VB3-C3
T3	VB2-C1, VB3-C1

Table 8. *Items excluded for further data analysis.*

4.7.2 Reliability

Reliability was addressed by assessing the internal consistency of the items in the experimental survey. The measure to ensure that items in a scale are reliable is Cronbach's Alpha, meaning that the scale measures what it is intended to measure (Cronbach, 1951). Cronbach's Alpha is a statistic used to assess the different dimensions of the construct, and each of these dimensions receives a value between 0 and 1. It has been widely accepted among the academic community that alphas of 0.7 or higher were found to indicate that the items in a scale were internally consistent in measuring a construct (Saunders, Thornhill

& Lewis, 2019). However, some authors have challenged this approach and argue that there are "limited grounds for adopting such a heuristic" and that a high alpha value is not always a good indicator. Instead, a construct measured with dimensions that receive values between 0.45 and 0.8 is considered to have acceptable alphas (Taber, 2018, p. 1288).

The alpha was reported through the pingouin statstical package method cronbach_alpha for the Python programming language. Note that alpha cannot be reported for the VB in T2, given that all items were deleted to ensure measurement validity except one.

Cronbach's Alpha results				
Dimension	**Control**	**T1**	**T2**	**T3**
TES	.676	.646	.683	.690
TEY	.587	.634	.605	.598
CIS	.615	.531	.476	.586
OS	.677	.667	.706	.705
VB	.860	.638	-	.672

Table 9. *Alphas for the different dimensions for each experimental group.*

Reliability is not only measured by assessing the internal consistency, but it can also be done with test-retest and an alternative form. The first refers to correlating the experimental survey data with those from an identical questionnaire that has been collected in similar conditions (Saunders, Thornhill & Lewis, 2019). The latter refers to comparing responses to an alternative question form (i.e., asking the same question in two very similar ways) (Mitchell, 1996). Test-retest was not done for this study due to a lack of time resources and participants. Equally, alternative question forms were not designed as these can easily lead to fatigue and a longer questionnaire (Saunders, Thornhill & Lewis, 2019). The longer the questionnaire, the more the participants tend to respond with identical answers, affecting the measurement of reliability through alternative question forms (Herzog & Bachman, 1981).

4.7.3 Ethical principles

The research community has shown interest in discussing the ethics of academic works. Whether researchers have used legitimate techniques during the research process has become an important aspect of research reporting since 1960, after unethical experimental studies threatened participants' lives to test hypotheses (Toepoel, 2015; Aguinis & Henle, 2002). In the case of survey-based experiments, the experimenter must have ethical obligations to the respondents by living up to three principles: The

principle of beneficence, the principle of justice, and the principle of respect for others (Aguinis & Henle, 2002).

Principle of beneficence

This principle refers to the researchers providing a beneficial outcome with their study, minimizing harm to the participants, and deciding whether benefits measure up the costs (Aguinis & Henle, 2002). In this case, the participants were treated in their preferred settings, and the treatments did not require the survey respondents to engage in any physical or psychic harming activity. Furthermore, participation was voluntary, and respondents were able to abandon the experiment if they did not want to continue participating. Hence, as no harm has occurred, the principle of beneficence can be reported.

Principle of justice

The principle of justice refers to finding a balance between those benefiting the research and those bearing the burden. For example, if a group of individuals participates in an experiment but a different group benefits from the outcomes, the principle of justice cannot be reported (Aguinis & Henle, 2002). In this case, the use case referred to Google, and therefore the experiment was conducted with participants that fall under the category of Google users, who will benefit from the experimental outcomes as they generate more knowledge on their preferred SE. The results of this experiment are open to the research community and organizations that want to leverage the outcomes of this paper. Hence, there are no conflicting interests that make the findings of this paper more beneficial for a particular group of individuals than others.

Principle of 'respect for others'

This principle refers to respecting the integrity of the individuals. It is concerned with treating all individuals equally, with respect and courtesy (Sieber, 1993). In order to ensure that all individuals were treated equally, anonymity was ensured throughout the whole procedure. Hence, even though demographical data had to be shared from the participant, this did not impact how they were considered during participation. All participants also received the same information, and it was ensured before beginning the experiment that they had gotten enough details about the survey. Participation was also voluntary, and any participant could abandon the experiment.

5. Results

This section reports the outcomes of the survey-based experiment and goes through all the steps necessary for an adequate assessment of the established hypotheses in chapter 3.

5.1. Experimental observations

The final dataset consisted of 195 observations. Of these, 175 were considered for further analysis as 20 participants reported that Google was not their preferred SE to perform searches online. 27 respondents did not answer the filter question correctly. Hence, the considered sample for the data analysis consisted of 148 observations distributed in the four different experimental groups.

As the data analysis required the groups to be equally populated, observations had to be dropped from the control group, T1 and T3 to equal the observations in T2, which was the least populated group with 34 observations. The final analysis was conducted with 136 observations in total.

Distribution of observations in the experimental groups		
Group	**Observations before equally populating the groups**	**Final observations considered for analysis**
Control	42 observations	34 observations
T1	35 observations	34 observations
T2	34 observations	34 observations
T3	37 observations	34 observations
Total	**148 observations**	**136 observations**

Table 10. *Number of recorded observations in each experimental group after sorting out invalid responses.*

5.1.1 Demographics in the experimental groups

Demographic data was collected from the participants regarding their age, education level, and gender to gain a deeper understanding on the samples.

5.1.1.1 Age

Most of the participants were part of the age group 21-30, with a total of 93 participants, accounting for values between 54,3% and 74,3% of the participants in each group. The least populated age group was 61-70, with no respondents in any groups, followed by the +70 group with one participant allocated in T1, and the 41-50 group, with a total of 5 participants.

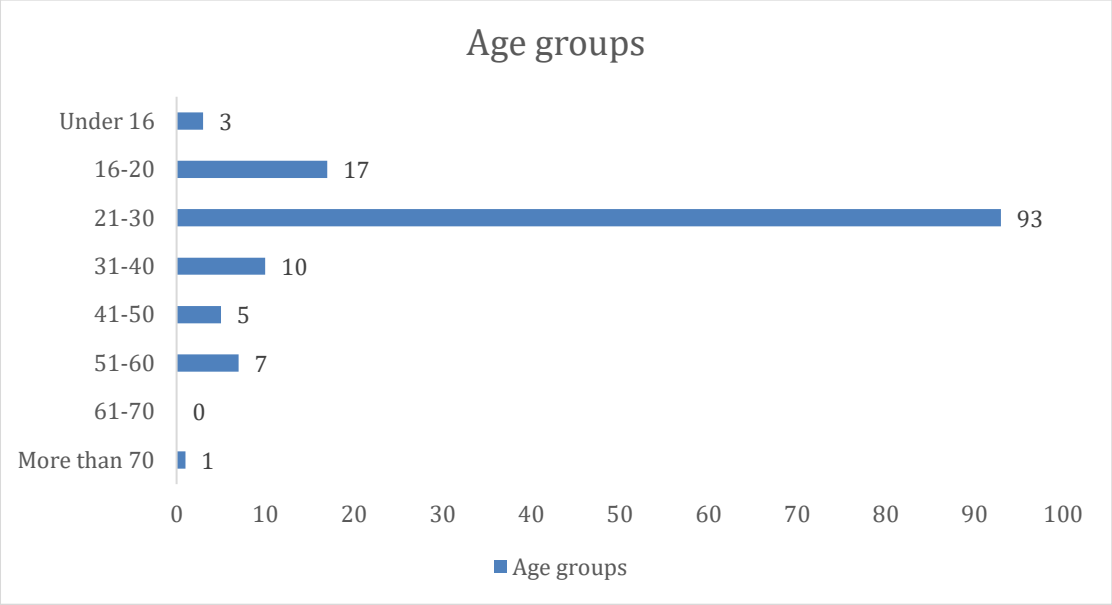


Figure 16. *Distribution of all participants' age groups.*

5.1.1.2 Gender

Genders were also not equally represented in the experimental groups, and females accounted for a higher proportion of the sample in all four groups. The control group was the group with most females, accounting for 71,4% of the participants, and the most equally distributed groups were T1 and T3, both with 57,1% female participants. In absolute terms, 84 participants were female, and 52 were male. No participants reported themselves in the Non-binary/third gender category.

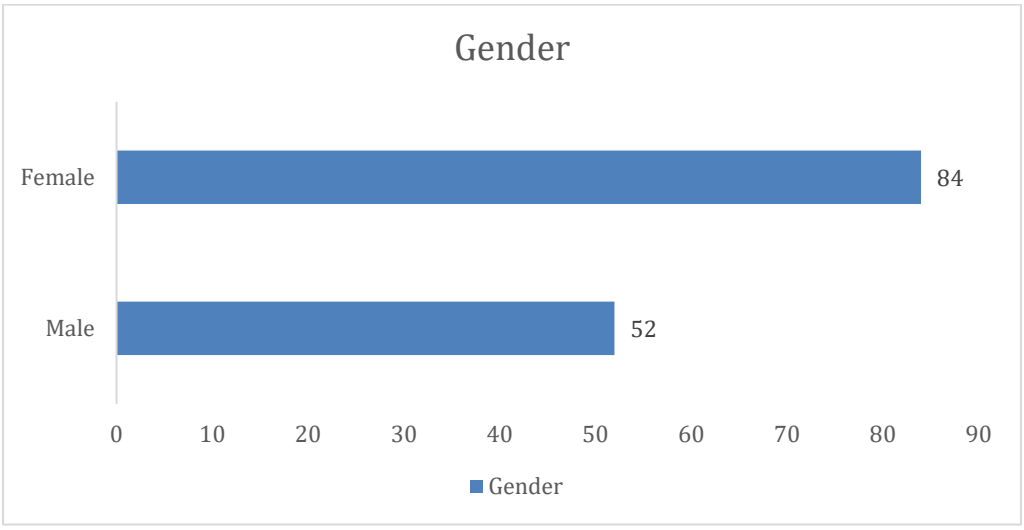


Figure 17. *Distribution of all participants' gender.*

5.1.1.3 Education

Participants were neither equally distributed when it comes to their education level. 58 participants stated that they had a bachelor's degree or professional degree, forming the most populated education group, followed by 46 people with a master's degree. 3 people reported that they had finished primary school as their highest education and were equally distributed between T1, T2, and T3. 0 participants said that they did not have any formal education. Some groups received a higher number of individuals with a Doctorate/Ph.D. For example, T2 received 5 participants, T1 received 1, whereas the control group and T3 received none.

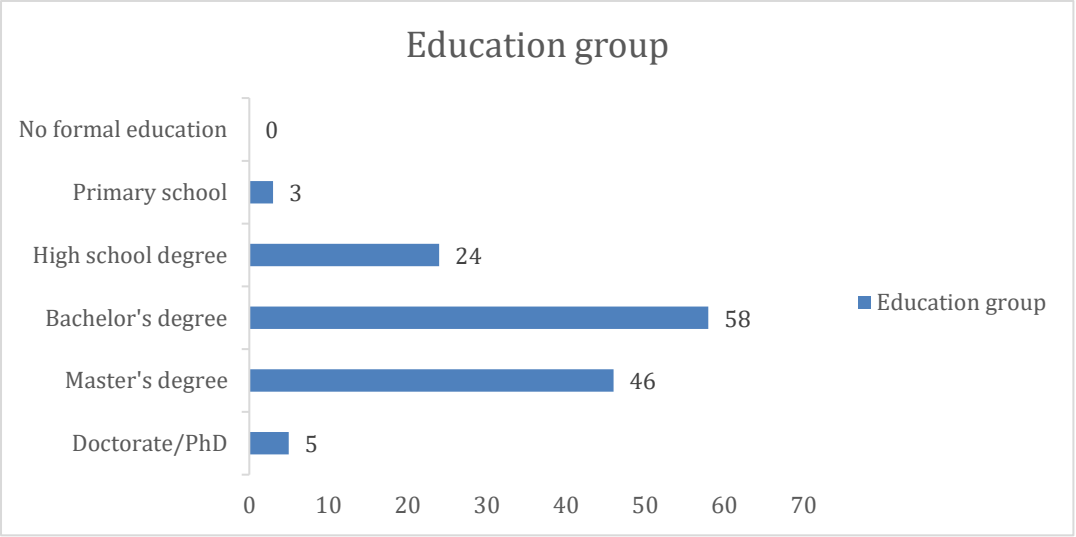


Figure 18. *Distribution of all participants' education groups.*

5.1.2 Reporting the User Experience (UX)

The mean overall UX score was assigned to each participant and treated as the response variable of interest in this research. This mean overall UX score was treated as a continuous variable based on Lachner et al. (2016) and was obtained by getting the average out of the 45 item scores that were assessed through a 5-point Likert scale. This score acts as a real number ranging between values of 1 and 5. The objective with analyzing the overall UX score as a response variable is to determine whether its value fluctuates significantly depending on the manipulation of SB and FSs, which are the two discrete variables needed to perform the two-way ANOVA (Kent, 2015). The terms "mean overall UX score" and "response variable" are used as synonyms in the following lines.

5.1.3 Basic summary statistics

The summary statistics on the response variable are reported for each experimental group to show tendencies in the collected data. The mean ($\bar{x}$), standard deviation (σ), minimum value (min), and maximum value (max) are shown in the following table for each experimental group:

Summary statistics				
	Control	**T1**	**T2**	**T3**
$\bar{x}$	**4.1147***	**3.4856***	4.0210	3.9273
σ	**0.5039***	0.5770	0.6318	0.6590
min	3.1500	2.3170	2.5142	2.0697
max	5.0000	4.6829	5.0000	4.7674

Table 11. *Summary statistics in the different experimental groups*

The experimental group that showed a higher UX mean was the control group ($\bar{x}$ = 4.11), whereas the lowest mean score was shown in T1 ($\bar{x}$ = 3.48). According to the mean values, all treatments had a negative impact on the UX, given that the highest mean was reported in the group that did not receive any treatment.

There is a difference of 0.1551 between the group with the highest and lowest standard deviation, meaning that the UX scores in the observations tended to fluctuate between 0.503 and 0.659 points around the mean**.** The standard deviation in the control group was lower than in the different treatment groups (σ = 0.503). This can be interpreted as participants responding very differently to the treatments in relation to when they are faced with unaltered rankings, where participants' mean overall UX score tended to be more consistent and closer to the mean.

The minimum values started at least at 2 points in regard to the Likert-scale, with the lowest being 2.06 in T3, followed by 2.31 in T1, which are the two groups containing SB. Two groups (control and T2) showed a maximum value of 5, which were the experimental treatments not containing any SB.

Contextualizing the results

There are no studies to the knowledge of the researcher that allows for a direct comparison with the leveraged research instrument, as the used instrument is adapted to the application domain based Lallemand & Koenig (2017) and Bernhaupt & Pirker (2013). Despite that, comparable questionnaires are found in the literature, which also leveraged 5-point Likert scales to measure the UX of a given product.

However, these studies have not focused on measuring the UX on SEs, and the dimensions and items differ from the ones of this study. Thus, it should be accounted for that different measurement instruments are being compared, and different products are being assessed.

In their questionnaire (N=15) consisting of 18 items and 4 dimensions, Hussain et al. (2017) measured the UX with a 5-point Likert scale and reported a higher mean than the ones reported in the experimental groups of this present study ($\bar{x} = 4.17$). However, their standard deviation differs from the reported in this experiment ($\sigma = 0.91$). Hassenzahl (2008) also quantified UX with an instrument containing 9 items and 3 dimensions leveraging a 5-point Likert scale (N=52). In this case, the reported mean was lower than any mean in the experimental groups in this research ($\bar{x} = 2.81$), whereas the standard deviation was higher ($\sigma = 1.21$). Fronemann & Peissner (2014, p. 733) concluded that when the $\bar{x} > 3.0$, the UX can be labeled as "positive," which is the case for both this current study and the one of Hussain et al. (2017).

The standard deviation reported for the two studies above is also higher than the ones noted in all four experimental groups for this study. A lower standard deviation is interpreted that most data points are closest to the mean (Toepoel, 2015). Therefore, the measurement instrument of this study produced results that are less sparse than previous UX measurements with 5-point Likert scales.

5.2. Data preparation

The data analysis was performed on the four different data frames created with the pandas library in Python used to generate the data structures. These data frames contained binary information on the treatments created for each experimental group. This step was necessary to perform a two-way ANOVA, which combines all the data frames to perform the analysis. Furthermore, statistical tests were run to ensure that the data suited the assumptions of a two-way ANOVA.

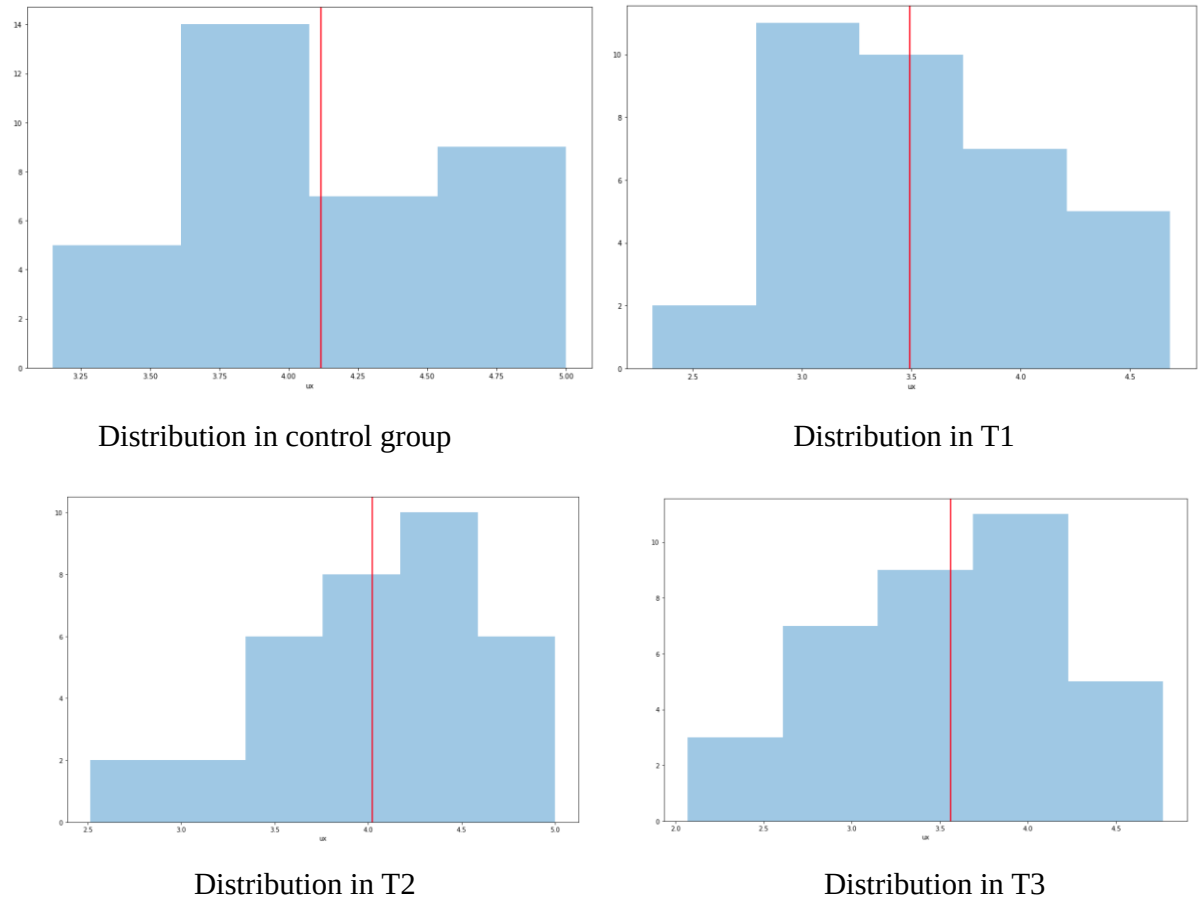
5.2.1 Results of the Shapiro-Wilk test

The Shapiro-Wilk test was performed on the four experimental groups to check for normality in the distributions. A $p > 0.05$ should be reported to ensure normality in the data (Ghasemi & Zahediasl, 2012). All the results supported the $H_{0,}$ and it can be assumed that the data in all four groups is normally distributed.

Shapiro-Wilk test results	
Group	**P-value**
Control	0.380
T1	0.834
T2	0.313
T3	0.712

Table 12. *Results of the Shapiro-Wilk test*

The distribution of the data was visually represented using seaborn, a library used for statistical representation on Python. A distribution plot is reported with a red line symbolizing the mean value.



Distribution in control group

Distribution in T1

Distribution in T2

Distribution in T3

Figure 19. *Distributions of the experimental groups with the highlighted mean. A higher definition image is found in Appendix 3.*

5.2.2 Results of Bartlett's test

Bartlett's test is performed with the objective of checking whether the variances between experimental groups are equal. Bartlett's test showed all variances to be equal, given that $p > 0.05$ (Wu & Wong, 2003).

The test is performed by comparing the four experimental groups, and the hypothesis is validated with p = 0.575. Hence, an equality in the variances is ensured. After this test, it can be concluded that the sampled data is appropriate for performing the data analysis.

5.3. Experimental findings

The following lines will explore the results of the two-way ANOVA test, and report whether simple effects and interaction effects can be reported. The significance of the overall model and the effects sizes of the significant effects are also assessed.

5.3.1 Two-way ANOVA

The two-way ANOVA checking the effects of stance bias (SB) and featured snippets (FSs) on the mean score of UX is performed to report both simple effects in reference to the control group and the interaction between the two independent variables representing the experimental treatments.

An interaction plot is represented to gain a visual understanding of the interactions between the two independent variables. An interaction effect happens when the mean in the experimental response variable is affected by the combination of the two independent variables explored in the two-way ANOVA (Gamst, Meyers & Guarino, 2008).

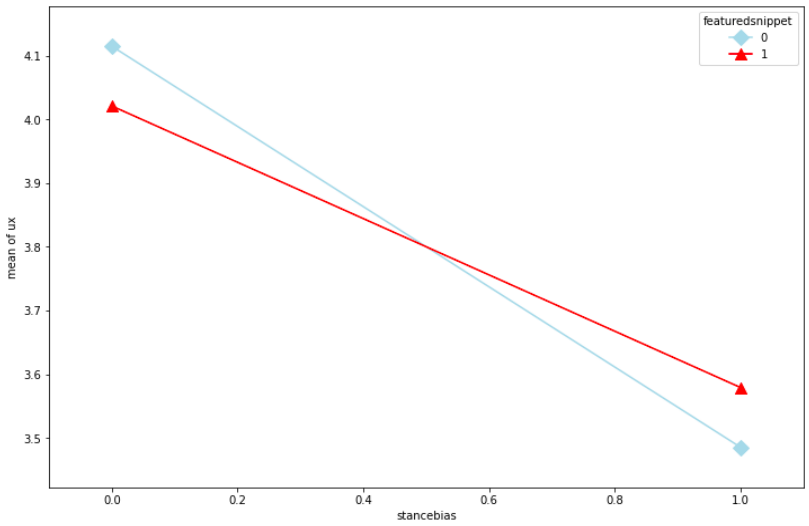


Figure 20. *Interaction plot between the two independent variables*

In order to detect a potential interaction effect, it is crucial that the two lines symbolizing the treatments are not parallel. In this case, the lines are crossing, which symbolizes that an interaction effect could be present, but can only be reported with a p-value that shows statistical significance (Gamst, Meyers &

Guarino, 2008). It is seen that the blue line, indicating the absence of FSs, reaches the lowest measured UX mean when SB is present (T1), staying under the red line symbolizing the presence of FSs and SB (T3). The blue line also stays over the red line when SB equals 0 (i.e., absent), indicating a higher UX when FSs appeared without the presence of SB (T2).

In order to evaluate whether significant simple effects and interaction effects can be reported, the p-values are explored:

OLS regression & two-way ANOVA summary table		
	coef.	p-value
Intercept	4.1147	-
T1 (SB)	-0.6216	0.000
T2 (FSs)	-0.0947	0.506
T3 (SB~FSs)	0.1634	0.415

Table 13. *OLS & two-way ANOVA results, highlighting the effect size and the significance.*

The result of the statistical tests leads to report a simple effect between the independent and response variable, but no interaction effects between the independent variables. The significant simple effect reported from the experimental procedure is the effect of SB on rankings, indicating that SB is associated with different levels of UX ($p = 0.000$), reducing the UX on average by -0.6216 points. On the contrary, the presence of FS is not related to different levels of UX ($p = 0.506$). Also, the relationship between SB and UX does not depend on the presence of FS ($p=0.415$). Therefore, the combination of SB and FSs on rankings does not affect the UX (i.e., no interaction effect is reported). Through these findings, the hypotheses resulting from the two-way ANOVA test are assessed:

Null hypotheses in the two-way ANOVA test	
H_0-1. There is no difference in the mean overall UX score given the presence of SB (simple effect)	Not supported
H_0-2. There is no difference in the mean overall UX score given the presence of FSs (simple effect)	Supported
H_0-3. The effect of SB on the mean overall UX score does not depend on the presence of FSs (interaction effect)	Supported

Table 14. *Null hypotheses assessment synthesis through the two-way ANOVA.*

5.3.2 Model significance

To perform the two-way ANOVA, it became a prerequisite to perform an OLS regression that determined the intercept of the model. In this experimental procedure, the intercept equals the mean of the control

group (4.11), as its value represents the response variable's value when the independent variables equal 0. The adjusted R^2 reports the fit of the model. This statistic explains in a value ranging between 0 and 1 how well the model predicts the dependent variable's values from the independent variable (Sharda, Delen & Turban, 2016).

The reported adjusted R^2 is relatively low, with a value of 0.163, considering the threshold of 0.3 for a good model in social sciences as proposed by Sharda, Delen & Turban (2016). Taking the reported figure into consideration, the model poorly predicts the response variable based on the two independent variables (SB & FSs). To improve the model, it is suggested to include new variables or remove existing ones to better explain the dependent variable (Sharda, Delen & Turban, 2016; Argyrous, 2011). In particular, the R^2 explains that other factors determine the UX on rankings, and that SB and FSs should not be used as the only predictors for determining the levels of UX.

5.3.3 Significance of the reported simple effect

The reported partial η2 reporting the effect of SB on the UX showed that 17.7% of the variance in the mean overall UX score can be explained through the presence of SB, given the partial η2 = 0.177. As the other two treatments were not shown to be significant, the partial η2 is not reported.

This figure can be reported as a medium to large effect according to Cohen's benchmarks, where a value close to 10% symbolizes a medium effect and a value close to 25% symbolizes a large effect (Cohen, 1968). If the outcome being studied is taken into consideration as suggested by Vacha-Haase & Thompson (2004), this effect size also tells the researcher that the substance of the content on rankings gains significance for the participants, but that other elements are also impacting the UX simultaneously. In the case of the current experimental settings, this could be the search users are involved with and the substance domain (Vuong et al., 2019), the layout presentation of the results on rankings (e.g., Ong et al., 2017), or the proposed URLs (Cutrell & Guan, 2007).

5.3.4 Hypothesis assessment

Based on these statistical results, the hypotheses established in chapter 3 are assessed. H1 is supported and H2 and H3 are not supported with the collected data.

H1. Rankings with top results leaning towards a particular stance influenced the UX compared to rankings without SB nor FSs on top results. The presence of SB on rankings showed a reduction of the

UX with -0.6216 points on average when assessing the means of the overall UX score. This was the only significant difference resulting from the analysis, and, hence, H1 is supported based on the collected data.

H2. The presence of FSs on rankings did not show to have a significant effect on the UX compared to rankings with neither FSs nor SB. The reported effect (-0.0947) is not interpreted, as p = 0.506. Hence, the given data does not allow to support H2.

H3. Combining both SB and FSs on rankings did not show to affect the UX neither in the experimental procedure in relation to rankings without SB and FSs, with an effect that cannot be interpreted (0.1634) as p = 0.415. H3 cannot be supported with the collected data.

Hypotheses assessment	
H1. The presence of SB on rankings affects the UX	Supported
H2. The presence of FSs on rankings affects the UX	Not supported
H3. The presence of both SB and FSs on rankings affects the UX	Not supported

Table 15. *Hypotheses assessment synthesis.*

6. Discussion

The findings in this research add value to existing academic theories on SEs. Managers interested in strengthening their online presence can also leverage the results to their advantage. This discussion provides a deeper understanding on how the UX is shaped depending on the configuration of SE rankings. As noted in the methodology chapter, this discussion contributes to theory-building and theory-modification by using an abductive approach to theory-development. A pragmatist philosophy is also adopted at this stage to analyze the findings to propose optimal solutions and paths for both academia and businesses. Emphasis is put on answering the RQ: *How does the presence of stance bias and featured snippets on rankings affect the user experience on search engines?*

This chapter also includes a critical insight into the limitations of this study to guide the reader in interpreting the results. Proposals of future directions for the research community are also provided.

6.1 Theoretical implications

The findings in this current research contribute to establishing a novel link between Google's choices about the configuration of SE rankings and UX. A relationship has been explored between the concepts of SEB and rich snippets with the UX. Whether this relationship exists is uncovered through experimentation, a research strategy that has been frequently used in digitized settings to find causal relationships by ensuring a random assignment to groups. This random assignment guarantees that the experimental groups are populated with participants that do not differ significantly between groups, and, therefore, possible effects of an alternative explanation that could be caused by differences in the group samples are removed to ensure causality (Kohavi, Henne & Sommerfield, 2007; Saunders, Thornhill & Lewis, 2019; Montgomery, 2017).

Users can show skepticism towards the content appearing on rankings

This study shows that the presence of SB on rankings decreases the UX ($p = 0.000$; coef. = -0.6216), whereas the presence of FSs on rankings, or an interaction of both FSs and SB, does not have any effect on the UX. This means that only H1 is supported through the results, with an effect of 17,7% of variance explained (partial $\eta 2 = 0.177$) In contrast, H2 and H3 cannot be supported based on the data. It should be noted that none of the experimental groups reported a decidedly negative UX when being treated, as a threshold of 3.0 was surpassed in the mean scores of each group. This value determines whether a product delivers a positive or negative experience when assessed with a 5-point Likert scale (Fronemann & Peissner, 2014). These findings come with theoretical implications, as they contribute to build on previous theories established in the state-of-the-art literature.

The user's tendency to rely blindly on a ranking's ability to sort information has been a reported oftentimes in academic works, and the algorithmic ideology perpetuated by Google has been a major cause for concern among experts in both sociology and IS, given that users are unconsciously willing to accept Google's rules (Mager, 2012; Beer, 2017). One of the most cited works exploring the psychological implications of Google's rankings is the research proposed by Pan et al. (2007), which suggests that SE users blindly trust Google's ability to rank websites. This is still true today as it has been found in more updated studies that "users strongly trust Google" (Schultheiß & Lewandowski, 2021, p. 1). These theoretical statements should be revised by contrasting it to the data of this work. The experiment shows that users are picky when faced with Google's rankings and that the results provided do not necessarily offer relevance to the user in terms of the substance of the content. The UX decreases when users are faced with a highlighted stance on rankings, which challenges the idea of rankings and algorithmic

outcomes being tools that the user would unconsciously rely on for decision-making. The fact that the UX would worsen when users were faced with documents containing stances in them suggests that the users have the ability to question the outcomes of the algorithm, and do not rely blindly on the SE.

It should be noted that most experiment participants shared demographic characteristics. Taking these characteristics into consideration is crucial to interpret the results as SE algorithms affect people with various demographic backgrounds in different ways (Mehrotra et al., 2017). The recorded observations came from Google users that were mainly 16 and 30 years old (83.08% of the participants), women (62.5%), and having at least a bachelor's degree (80.14%). The sample is comparable to the participants of Pan et al.'s (2007) study, which consisted of undergraduate students with an average age of 20.4. This similarity between participants indicates a possible paradigm shift, as younger users in 2007 showed high levels of trust bias and believed in the SEs ability to rank results neutrally. In contrast, the current findings suggest that users show skepticism towards Google's rankings ability to provide relevant results, as they see the UX reduced when the results lean towards a particular stance.

Users give up their "romanticized view" towards Google

The skepticism towards Google can also be interpreted given the low measurements in the visual branding dimension. During the design of the research of this study, the measurement instrument was adapted to fit the domain of interest (Lallemand & Koenig, 2017; Bernhaupt & Pirker, 2013). Even though the instrument was built on deductive premises (Hinkin, 1998), various dimensions were omitted to suit the experimental procedure. Therefore, the UX was measured based on 5 different dimensions instead of the 9 suggested by Lachner et al. (2016). One of these dimensions chosen for the experimental procedure, visual branding (VB), raised validity concerns during the CFA. 18 items of the dimension had to be omitted in the data analysis to ensure validity, accounting for half the items of a total of 36.

The VB dimension made the user assess the brand behind the tested product, in this case, Google. The brand showed to be, in half of the circumstances, an invalid measurement of UX, as much lower mean scores were reported compared to the measurement of other dimensions. For example, in the VB items of T1, mean scores under 3.0 were reported in 7 out of 9 items. Considering 3.0 as a threshold for a positive experience (Fronemann & Peissner, 2014), the VB received a negative score when participants were asked whether they trusted Google or felt it is an honest and safe brand. This was not observed in any other items in the different dimensions. These findings suggest a paradigm shift and challenge views established by authors such as Goldman (2005, p. 189), reporting that users have a "romanticized view" towards SEs.

Also, more recent research still reports that a large group of users still show high levels of trust regarding Google as a SE (Schultheiß & Lewandowski, 2021). Instead, it is more appropriate to conclude that users are also able to show skepticism towards Google as a brand when evaluating algorithmic outcomes.

The enigmatic role of featured snippets

To the knowledge of the researcher, existing literature has not focused on understanding how FSs in isolation of other rich snippets impact the user. This experiment has attempted to do so to fill in a theoretical gap. The results show that FSs do not affect the UX, as no significant difference was reported between experimental groups containing FSs and those omitting them.

It is still unknown what implications FSs have on users navigating rankings. No answers can be provided without further research, but some hypotheses can be established based on the findings in this paper.

It is a theme of interest why the UX did not decrease in the experimental group where participants faced rankings containing both SB and FSs (T3), given that participants in T1 saw their UX decreased when seeing SB in isolation. A plausible explanation is that a change in the layout as provided with FSs might distract the user from focusing solely on the content. Further investigation is necessary for this hypothesis to be confirmed.

Users enjoy researching

The search tasks the participants encountered during experimentation were framed as informational, with an individual intention of gaining knowledge about a topic in three different substance domains (Vuong et al., 2019). The fact that participants saw their UX decreased when encountering SB leaves room for interpretation regarding their preferences during a search session. Results indicate that users are willing to do their own research on topics they have not formed an opinion on, as participants showed more interest in rankings with sites that gave them information accounting for both stances where they could explore a topic from a neutral perspective. This interest can be interpreted as users having a predisposition to extend the search session to explore documents that could provide deeper insights on a topic, which becomes an eye-opening finding for managers discussed in the next section.

6.2 Managerial implications

The experimental findings are also relevant for professionals interested in comprehending the synergy between SE rankings and the UX. The results give new possibilities for orienting business strategies that account for the variables explored in this experiment, which affect business performance given their impact on end-users.

Google's algorithm has impacted the economic order and dictated how companies are marketed online, conditioning both the exposure and the content produced globally (Bilić, 2016; Latzer et al., 2016). Google's algorithm favors a system of punishments and rewards, and organizations with an online presence tend to try to "please" Google (Röhle, 2009, Bar-Ilan, 2007b, 163). SEO is no longer seen as an option. Instead, it has become a fundamental marketing tool for ensuring a website's visibility (Schultheiß & Lewandowski, 2020). This study presents a new approach to SEO consisting of living up to the algorithm's requirements while simultaneously considering users' preferences. To ensure a website's presence on rankings, it is crucial to keep up to date with Google's algorithmic decisions and ranking factors. These ranking factors tend to have one thing in common: They rarely assess the semantical substance of the content (Ziakis et al., 2019). Managers interested in revising or creating new SEO and content strategies can use this to their advantage by making editorial choices that are shown to possibly impact the UX.

Managers can impact experiences online by considering the role of stances in their content. An informed SEO strategy uses meta title tags and meta descriptions that do not lean towards any particular stance and takes into consideration that Google interprets poor title and description tags as a low-quality website (Ziakis et al., 2019). Improving the presentation of sites on rankings by removing stances is likely to attract visitors based on the findings of this study.

Enhancing the UX and profoundly considering the user sitting at the other side of the screen results in an innovative way of optimizing websites and ensuring online exposure. This study suggests that professionals should enhance the UX by using user-centric approaches to SEO. According to the findings, a good online presence allows the end-user to do their own research on the topics they are searching for. Therefore, experts in the field of content marketing can leverage this to their interest by not presenting clear viewpoints about a topic on the parts of a website that could be exposed on ranking snippets. If presenting stances is in the interest of the business strategy, it is recommendable to give the user a sense of cognitive effort and offer viewpoints that seem logical in their information-seeking process.

To the advantage of managers, FSs did not significantly impact the UX. This is a positive finding for this industry, considering that the appearance of FSs is an uncontrollable factor that depends on the algorithm (Strzelecki & Rutecka, 2020).

6.3 Limitations

The experimental results should be analyzed considering this research's limitations. Whenever an experiment is conducted, some factors become uncontrollable, and these tend to be the significant limitations for validating the outcome of experiments (Montgomery, 2017). Hence, this section presents the considerations that should be made when reading through this study's results. In this case, the detected uncontrollable factors have been the searching behavior, the chosen device and UI, and the user variability. Limitations associated with the regression model are also highlighted here.

Users behave differently depending on the subtask they are engaged with. In this case, the experiment has focused on understanding the UX solely on the result scanning step. The omission of other searching subtasks might raise external validity concerns, as the experiment is missing behavioral patterns that are characteristic of other steps of the information-seeking process that might affect the UX (Baskaya, Keskustalo & Järvelin, 2013). An important aspect when assessing the UX on SEs are the so-called feedback loops (Spink & Saracevic, 1998), which are the user-initiated loops consisting of users inputting a query and interpreting the appropriateness of the content returned. This current research does not collect any data allowing for the interpretation of this loop due to the omission of the querying step. Also, no information is gathered on what pages the participants were willing to click on.

The query formulation step was omitted intentionally, as its disadvantages outweigh the advantages. For example, it is known that users pose a different number of queries for similar information needs (Bailey et al., 2015), which would frustrate users who expect to pose more than one query in the experimental procedure. The possibility of posing several queries for an information need was not an option, as the participants would only have access to a limited number of rankings. Furthermore, it would make the experimental procedure much longer than planned, leading the participants to fatigue. Also, when users interact with systems, they sometimes have the misconception that they are being tested on their ability to interact with the technology, leading to a change in their behavioral patterns and assessment (Lauesen, 2005).

Search behaviors are also device-dependent (Ong et al., 2017). Even though the experiment was designed for participants to only see rankings on mobile screens, some uncontrollable factors regarding the environment of use might condition experimental results. No information was collected on whether participants went through the questionnaire from a desktop or mobile device, and neither the effect of having desktop users see mobile rankings. It is known that mobile users are more susceptible to ranking effects (Ghose, Goldfarb & Han, 2013), but it is unclear whether this depends on the effect that the UI layout has on the user or whether it is affected by the device used. It also remains unknown whether the same experimental procedure would have provided other measures for UX if the FSs were used in the context of voice search, as it is common with FSs (Strzelecki & Rutecka, 2020).

Changes in the surrounding environment during experimentation could also affect the results. Participants filling the questionnaire in a situation where all their cognitive load had been focused on the experimental procedure might have led to responses different from those in a situation where the focus was put on other activities simultaneously (Harvey & Pointon, 2017). As it was not possible to control under which circumstances the participants performed the experiment, it should be noted that the UX might have been measured under very different circumstances. It is expected that differences between groups had been ruled out through randomization to avoid that this became an alternative explanation to the results, but this cannot be guaranteed as no data exists on the participants' environmental circumstances (Saunders, Thornhill & Lewis, 2019). Still, it should be noted that reproducing the experiment with users ensured to be in the same contextual settings might alter the results.

The UI presented to the participants also impacts the experimental procedure. The experimental layout was created by ensuring that only the manipulated factors (SB and FSs) were considered during ranking assessment. Therefore, participants were faced with a ranking UI that was not identical to natural search settings, which might impact the response variable. For example, it is known that when the snippet length increases, users tend to pay more attention to the snippet than the URL (Cutrell & Guan, 2007). However, the URL was omitted in this experiment so it would not become an assessment factor that could influence participants' perceptions towards SB and FSs. This was done as well for other SERP elements such as paid results or multimedia (Bailey et al., 2010). These changes might raise questions towards the study's external validity.

The configuration of the sample might also have affected the results. The lack of a sampling frame led to adopt convenience sampling, which can provide limitations that should be highlighted for the proper

interpretation of the results. In this case, young users were predominant with an educational level of at least a bachelor's degree. However, the use of a SE varies from user to user, and this user variability might affect UX scores. For example, younger users behave differently than older users on SEs (Mehrotra et al., 2017; Wolfram & Xie, 2002; Carterette, Kanoulas & Yilmaz, 2012). Therefore, the reported results might be very user-dependent, and reproducing the study with predefined user segments might lead to different measures of UX.

Another limitation is that a moderate amount of data on the participant's demographics were collected. The objective of minimizing data collection was to reduce fatigue during experimentation. One of the variables that were not reported is the participants' geographic location or nationality. The lack of this information becomes a limitation as SEs affect people from different cultures differently (Madankar, Chandak & Chavhan, 2016). It is unknown whether cultural factors influenced UX measurements during experimentation.

When performing the OLS regression prior to interpreting the two-way ANOVA results, some limitations must be highlighted in relation to the reported adjusted R^2, which was considerably low ($R^2 = 0.163$) in relation to the thresholds accepted in social sciences ($R^2 = 0.3$) (Sharda, Delen & Turban, 2016). In other words, the model is poorly fit, and does not properly predict the outcomes of a response variable given two independent variable measures. This statistic should be considered when interpreting the results of this work and seen as an indicator highlighting that the variables explaining the UX on rankings should be reconsidered. The following section, focusing on future research, proposes directions to overcome this issue.

6.4 Future research

New research directions have emerged during the discussion of the results. Building on the findings of this work would contribute to a richer interpretation of the experimental outcomes. This section proposes new approaches for data collection, experimentation, and novel methods that can be leveraged to provide a deeper understanding on the relationship between rankings and the UX, together with other variables that fall outside the scope of this study. Furthermore, some limitations highlighted in this research can be overcome with further investigations attempting to perfect measurements of UX to achieve better predictions.

This research focused on understanding both simple effects and interaction effects of SB and FSs on rankings when they appeared in combination with result snippets containing a title and a description. However, the reported adjusted R^2 raised concerns towards SB and FSs being good enough predictors of the UX. In order to propose a new model that better explains the UX on rankings, it is suggested to add more variables to gain richer insights on this relationship (Sharda, Delen & Turban, 2016; Argyrous, 2011). Therefore, future works can focus on employing similar statistical techniques such as multiple regression models that consider other elements such as paid results or multimedia to interpret how ranking configurations affect the UX. Including new variables in a similar study might help improve the adjusted R^2 and provide a richer explanation on the relationship between ranking elements and UX investigated in this academic work.

An inconsistency found in SE literature was detected when comparing the findings of this research with those of Pan et al. (2007) and Schultheiß & Lewandowski (2021), showing that Google users often experience trust bias. In contrast, this current research indicates that participants rated Google as a brand with much lower scores than items in other UX dimensions, which questions the user's trust towards the SE. It remains unclear what the current level of trust towards Google is, and what factors are conditioning their perceptions towards the SE. Despite this apparent lack in trust reported in this research, Google is currently the most used SE, accumulating 86.6% of SE users worldwide by February 2021 (StatCounter, 2021). Hence, future studies might focus on exploring whether a paradigm shift regarding trust bias can be detected by exploring user's information consumption habits in their daily routines through observational studies based on Liu & Li's (2016) theory.

A theme worth researching to build on existing findings is whether SE users are bias-aware. Even though a lower UX was reported by participants seeing rankings containing SB, this current research does not provide answers to why this effect is reported. In fact, previous research supports that the effects of algorithms such as Google go unnoticed by users (Beer, 2017). Therefore, qualitative user research based on self-reported judgements as suggested by Mehrotra et al. (2017), becomes a valuable tool to build on the existing findings.

Confirmation bias (CB) has also been a recurrent topic in SE literature but was not included in the scope of this study. Despite that, theory suggests that introducing CB in a similar experimental procedure would alter the UX outcome if users faced stances confirming their preconceived ideas about a topic. This argumentation is made based on Meppelink et al.'s (2019) findings, suggesting that users prefer to see

information that aligns with their prior beliefs. Therefore, such research could bring new perspectives of SB on SE rankings, trying to investigate whether the presence of stances on rankings increases if users are faced with belief-consistent information.

The role of FSs gives fresh opportunities for future research, as this study could only conclude that they had no effects on the user. It is suggested to get guidance on exploring FSs with a more interpretivist approach by obtaining qualitative data on how FSs impact the user. During experimentation, it was registered that there were no interaction effects between FSs and SB. In contrast, a simple effect of SB on rankings was reported when it appeared in isolation of FSs. Therefore, it is also convenient to explore why these results are reported as such by using self-reported judgements.

7. Conclusion and final words

At the beginning of this research the reader met Richard Gilmore. In the described scene, it became implicit in his behavior that he found immediacy and joy in using Google to research his wife. In other words, it seemed that Google was offering Richard Gilmore a positive UX. However, for the time being, state-of-the-art literature has paid little attention to how users like Richard react when they meet with Google's offerings and how that affects the UX. As reported in chapter 2, IR evaluation has tended to leave the user aside when evaluating SEs (Moffat et al., 2017).

A lot has changed since 2003, when the described scene was aired. The Google Richard Gilmore interacted with is very different from the one users interact with today. Even though the PageRank algorithm appeared as a technologically superior product already in the 1990s (Bilić, 2016), Google has invested energy in updating the algorithm year after year to an improved user solution to avoid deprecation (Ziakis et al., 2019; Bilić, 2016). This importance Google has given this algorithmic ideology has come with its effects on HCI (Mager, 2012). For instance, the SE has favored particular types of content in front of others, introducing bias to results on rankings (e.g., Goldman, 2005; Gao & Shah, 2020). Furthermore, the UI Richard Gilmore interacted with looks different than the SERPs cluttered with new information layouts that people are used to in 2021. Despite all these changes, the research agenda has not focused on uncovering whether there is a relationship between all these rich elements and inclinations appearing on rankings and the UX. Hence, a need for updated explanatory research that could establish a one-way relationship between these algorithmic outcomes and the UX emerged.

The data resulting from the survey-based experiment in this research could support 1 out of the 3 established hypotheses in this study. A negative effect on the UX was reported when users encountered SB on rankings. Despite that, no impact on the UX was found when FSs or a combination of SB and FSs appeared on rankings. These findings update state-of-the-art literature with a deeper understanding on how ranking elements affect the UX, showing that users raise questions towards retrieved documents and show skepticism towards the role of Google as a brand despite using it as their preferred SE.

Furthermore, it gives information to managers engaging in SEO and content marketing activities. As a starting point, users value that they are treated as individuals capable of doing their own research and document exploration instead of being told the mindset they are encouraged to have on a particular topic based on the stance appearing on rankings. In other words, marketers should put the user in the center and optimize content on rankings with titles and descriptions that treat the user as an individual capable of making informed choices.

What does the future hold?

Due to the evolving nature of SEs, and the changes in perceptions of the users behind the screens, this field of research becomes particularly attractive to study new phenomena and patterns in the users' expectations towards SEs. Google has traditionally made the algorithm condition the economic order and the user's behavior on SEs, but Google's interest in enhancing the UX might indicate that users are potentially able to shape the algorithm and have the power to make this rule go the other way around. The findings in this study have shown a growing trend of fussy users craving a more trustworthy SE. Whether Google will catch up with these user demands, or competitors will trigger disruptive market rules is unknown. Despite that, this study has been an attempt to show a cross-sectional snapshot on the UX on Google to help orient research towards exploring whether the future might be holding a paradigm shift regarding information consumption on SEs.

REFERENCES

Ageev, M., Lagun, D., & Agichtein, E. (2013). Towards task-based snippet evaluation: preliminary results and challenges. In MUBE, SIGIR, pages 1–2, 2013

Aggarwal, C. C. (2018). Information retrieval and search engines. In Machine Learning for Text (pp. 259-304). Springer, Cham.

Aguinis, H., & Henle, C. A. (2002). Ethics in research. Handbook of research methods in industrial and organizational psychology, 5, 34-56.

Al-Maskari, A., Sanderson, M., & Clough, P. (2007, July). The relationship between IR effectiveness measures and user satisfaction. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 773-774).

Anderson, N. H. (2001). Empirical direction in design and analysis. Psychology Press.

Anderson, V. L., & McLean, R. A. (2018). Design of Experiments: A Realistic Approach. CRC Press.

Argyrous, G. (2011). Statistics for Research: With a Guide to SPSS (3rd ed.). SAGE Publications.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. Methodology, 6, 128-138

Azzopardi, L., Kelly, D., & Brennan, K. (2013, July). How query cost affects search behavior. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 23-32).

Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2004, March). Query recommendation using query logs in search engines. In International conference on extending database technology (pp. 588-596). Springer, Berlin, Heidelberg.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54-61.

Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A., & Tahaghoghi, S. M. (2010, July). Evaluating whole-page relevance. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 767-768).

Bailey, P., Moffat, A., Scholer, F., & Thomas, P. (2015, August). User variability and IR system evaluation. In Proceedings of The 38th International ACM SIGIR conference on research and development in Information Retrieval (pp. 625-634).

Bar-Ilan, J. (2007a). Google bombing from a time perspective. Journal of Computer-Mediated Communication, 12(3), 910-938.

Bar-Ilan, J. (2007b). Manipulating search engine algorithms: the case of Google. Journal of Information, Communication and Ethics in Society, Vol. 5 No. 2/3, pp. 155-166.

Baskaya, F., Keskustalo, H., & Järvelin, K. (2013, October). Modeling behavioral factors in interactive information retrieval. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 2297-2302).

Beer, D. 2017. The Social Power of Algorithms. Information, Communication & Society, 20(1), pp. 1-13.

Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. MIS quarterly, 3-16.

Bernhaupt, R., & Pirker, M. (2013, September). Evaluating user experience for interactive television: towards the development of a domain-specific user experience questionnaire. In IFIP Conference on Human-Computer Interaction (pp. 642-659). Springer, Berlin, Heidelberg.

Bilić, P. (2016). Search algorithms, hidden labour and information control. Big Data & Society, 3(1), 1-9.

Blaikie, N., & Priest, J. (2019). Designing social research: The logic of anticipation. John Wiley & Sons.

Blair-Early, A., & Zender, M. (2008). User interface design principles for interaction design. Design Issues, 24(3), 85-107.

Boudreau, M. C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. MIS quarterly, 1-16.

Brown, J. D. (2000). What issues affect Likert-scale questionnaire formats. Shiken: JALT Testing & Evaluation SIG Newsletter, 4(1).

Burrell, G., & Morgan, G. (1979). Sociological paradigms and organisational analysis (Vol. 248). Aldershot: Gower Publishing Company Limited.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. Information processing & management, 31(2), 191-213.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Cardinal, R. N., & Aitken, M. R. F. (2013). ANOVA for the Behavioral Sciences Researcher. Taylor and Francis.

Carterette, B., Kanoulas, E., & Yilmaz, E. (2012, October). Incorporating variability in user behavior into systems based evaluation. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 135-144).

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. Journal of Economic Behavior & Organization, 81(1), 1-8.

Child, D. (2006). The essentials of factor analysis. (3rd ed.). New York, NY: Continuum International Publishing Group.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological bulletin, 70(6p1), 426.

Creswell, J. W. (2003). Research design: Qualitative and quantitative approaches. Thousand Oaks, CA: Sage publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297-334.

Crotty, M. (1998). The foundations of social research: Meaning and perspective in the research process. Sage.

Cutrell, E., & Guan, Z. (2007, April). What are you looking for? An eye-tracking study of information usage in web search. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 407-416).

Dawson, J. (2016). Analysing Quantitative Survey Data for Business and Management Students. SAGE Publications.

Dempsey, L., & Heery, R. (1998). Metadata: a current view of practice and issues. Journal of documentation, Vol. 54 No. 2, pp. 145-172.

Edwards, A., & Kelly, D. (2017, August). Engaged or frustrated? Disambiguating emotional state in search. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 125-134).

Ellis, D. (1989). A behavioural approach to information retrieval system design. Journal of Documentation, Vol. 45 No. 3, pp. 171-212.

Epstein, R., & Robertson, R. E. (2015). 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.

Evaluating page experience for a better web. **(2020, May 28).** Google Search Central Blog. Retrieved May 08, 2021, from https://developers.google.com/search/blog/2020/05/evaluating-page-experience

Field, A. P. (2005). Discovering Statistics Using SPSS. Sage Publications Inc.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. ACM Transactions on Information Systems (TOIS), 23(2), 147-168.

Fronemann, N., & Peissner, M. (2014, October). User experience concept exploration: user needs as a source for innovation. In Proceedings of the 8th nordic conference on human-computer interaction: Fun, fast, foundational (pp. 727-736).

Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). Analysis of Variance Designs: A Conceptual and Computational Approach with SPSS and SAS. Cambridge University Press.

Gao, R., & Shah, C. (2020). Toward creating a fairer ranking in search engine results. Information Processing & Management, 57(1), 102138.

Gezici, G., Lipani, A., Saygin, Y., & Yilmaz, E. (2021). Evaluation metrics for measuring bias in search engine results. Information Retrieval Journal, 1-29.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism, 10(2), 486.

Ghose, A., Goldfarb, A., & Han, S. P. (2013). How is the mobile Internet different? Search costs and local activities. Information Systems Research, 24(3), 613-631.

Gillespie, T. (2014). The relevance of algorithms. In: Gillespie T, Boczkowski P and Foot KA (eds) Media Technologies: Essays on Communication, Materiality, and Society. Cambridge, MA; London: MIT Press, pp. 167–193.

Goertzen, M. J. (2017). Introduction to quantitative research and data. Library Technology Reports, 53(4), 12-18.

Goldman, E. (2005). Search engine bias and the demise of search engine utopianism. Yale JL & Tech., 8, 188.

Grossman, D. A., & Frieder, O. (2012). Information retrieval: Algorithms and heuristics (Vol. 15). Springer Science & Business Media.

Haim, M., Graefe, A., & Brosius, H. B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. Digital journalism, 6(3), 330-343.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. Proceedings of the National Academy of Sciences, 112(8), 2395-2400.

Harman, D. (2012, January). TREC-style evaluations. In PROMISE Winter School (pp. 97-115). Springer, Berlin, Heidelberg.

Harvey, M., & Pointon, M. (2017, March). Perceptions of the effect of fragmented attention on mobile web search tasks. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (pp. 293-296).

Haslwanter, T. (2016). An Introduction to Statistics with Python: With Applications in the Life Sciences. Springer International Publishing.

Hassan, A., Jones, R., & Klinkner, K. L. (2010, February). Beyond DCG: user behavior as a predictor of a successful search. In Proceedings of the third ACM international conference on Web search and data mining (pp. 221-230).

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. Behaviour & information technology, 25(2), 91-97.

Hassenzahl, M. (2008, September). User experience (UX) towards an experiential perspective on product quality. In Proceedings of the 20th Conference on l'Interaction Homme-Machine (pp. 11-15).

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. Public opinion quarterly, 45(4), 549-559.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. Organizational research methods, 1(1), 104-121.

Holden, M. T., & Lynch, P. (2004). Choosing the appropriate methodology: Understanding research philosophy. The marketing review, 4(4), 397-409.

Huang, H., Zhang, B. (2009). Text Indexing and Retrieval. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

Hussain, A., Mkpojiogu, E. O., Musa, J. A., & Mortada, S. (2017, October). A user experience evaluation of amazon kindle mobile application. In AIP conference proceedings (Vol. 1891, No. 1, p. 020060). AIP Publishing LLC.

International Standards Organization. (2010). Ergonomics of Human-System Interaction—Part 210: Human Centred Design for Interactive Systems, ISO 9241-210.

Jafarzadeh, H., Abedin, B., Aurum, A., & D'Ambra, J. (2019). Search Engine Advertising Perceived Effectiveness: A Resource-Based Approach on the Role of Advertisers' Competencies. Journal of Organizational and End User Computing (JOEUC), 31(4), 46-73.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2017, August). Accurately interpreting clickthrough data as implicit feedback. In ACM SIGIR Forum (Vol. 51, No. 1, pp. 4-11). New York, NY, USA: Acm.

Johnson, K., & Goldwasser, D. (2016, November). Identifying stance by analyzing political discourse on twitter. In Proceedings of the First Workshop on NLP and Computational Social Science (pp. 66-75).

Kammerer, Y., & Gerjets, P. (2010, March). How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (pp. 299-306).

Kayhan, V. (2015). Confirmation bias: Roles of search engines and search contexts. Paper presented at Thirty Sixth International Conference on Information Systems, Fort Worth 2015.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Now Publishers Inc.

Kelly, D., & Azzopardi, L. (2015, August). How many results per page? A study of SERP size, search behavior and user experience. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 183-192).

Kent, R. A. (2015). Analysing Quantitative Data: Variable-based and Case-based Approaches to Non-experimental Datasets. SAGE Publications.

Khoo, M., & Hall, C. (2012, September). What would 'google'do? users' mental models of a digital library search engine. In International Conference on Theory and Practice of Digital Libraries (pp. 1-12). Springer, Berlin, Heidelberg.

Kim, J., & Carvalho, V. R. (2011, April). An analysis of time-instability in web search results. In European Conference on Information Retrieval (pp. 466-478). Springer, Berlin, Heidelberg.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H. J. (2017, March). What snippet size is needed in mobile web search?. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (pp. 97-106).

Kohavi, R., Henne, R. M., & Sommerfield, D. (2007, August). Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 959-967).

Lachner, F., Naegelein, P., Kowalski, R., Spann, M., & Butz, A. (2016, October). Quantified UX: Towards a common organizational understanding of user experience. In Proceedings of the 9th Nordic conference on human-computer interaction (pp. 1-10).

Lagun, D., Hsieh, C. H., Webster, D., & Navalpakkam, V. (2014, July). Towards better measurement of attention and satisfaction in mobile search. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 113-122).

Lallemand, C., & Koenig, V. (2017, September). How Could an Intranet be Like a Friend to Me? Why Standardized UX Scales Don't Always Fit. In Proceedings of the European Conference on Cognitive Ergonomics 2017 (pp. 9-16).

Latzer, M., Hollnbuchner, K., Just, N., & Saurwein, F. (2016). The Economics of Algorithmic Selection on the Internet. Zurich: University of Zurich.

Lauesen, S. (2005). User interface design: a software engineering perspective. Pearson Education.

Law, E., Roto, V., Vermeeren, A. P., Kort, J., & Hassenzahl, M. (2008). Towards a shared definition of user experience. In CHI'08 extended abstracts on Human factors in computing systems (pp. 2395-2398).

Law, E. L. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009, April). Understanding, scoping and defining user experience: a survey approach. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 719-728).

Law, E. L. C., & Van Schaik, P. (2010). Modelling user experience–An agenda for research and practice. Interacting with computers, 22(5), 313-322.

Lewandowski, D., & Sünkler, S. (2013). Representative online study to evaluate the revised commitments proposed by Google on 21 October 2013 as part of EU competition investigation AT. 39740-Google Report for Germany.

Lewandowski, D. (2017). Is Google responsible for providing fair and unbiased results?. In The responsibilities of online service providers (pp. 61-77). Springer, Cham.

Lewis, B. R., Snyder, C. A., & Rainer Jr, R. K. (1995). An empirical assessment of the information resource management construct. Journal of Management Information Systems, 12(1), 199-223.

Liu, A., & Li, T. M. (2016). Develop habit-forming products based on the Axiomatic Design Theory. Procedia CIRP, 53, 119-124.

Ma, W., Wang, Z., Zhang, M., Qian, J., Luan, H., Liu, Y., & Ma, S. (2019, October). Stance Influences Your Thoughts: Psychology-Inspired Social Media Analytics. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 685-697). Springer, Cham.

Madankar, M., Chandak, M. B., & Chavhan, N. (2016). Information retrieval system and machine translation: a review. Procedia Computer Science, 78, 845-850.

Mager, A. (2012). Algorithmic ideology: How capitalist society shapes search engines. Information, Communication & Society, 15(5), 769-787.

Marcos, M. C., Gavin, F., & Arapakis, I. (2015, September). Effect of snippets on user experience in web search. In Proceedings of the XVI International Conference on Human Computer Interaction (pp. 1-8).

McKinney, W. (2011). "pandas: a foundational Python library for data analysis and statistics." Python for High Performance and Scientific Computing 14, no. 9: 1-9.

McNamee, R. (2003). Confounding and confounders. Occupational and environmental medicine, 60(3), 227-234.

Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., & Yilmaz, E. (2017, April). Auditing search engines for differential satisfaction across demographics. In Proceedings of the 26th international conference on World Wide Web companion (pp. 626-633).

Meppelink, C. S., Smit, E. G., Fransen, M. L., & Diviani, N. (2019). "I was right about vaccination": Confirmation bias and health literacy in online health information seeking. Journal of health communication, 24(2), 129-140.

Mitchell, V. (1996). Assessing the reliability and validity of questionnaires: an empirical example. Journal of Applied Management Studies, 5, 199-208.

Modave, F., Shokar, N. K., Peñaranda, E., & Nguyen, N. (2014). Analysis of the accuracy of weight loss information search engine results on the internet. American Journal of Public Health, 104(10), 1971-1978.

Moffat, A., Thomas, P., & Scholer, F. (2013, October). Users versus models: What observation tells us about effectiveness metrics. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 659-668).

Moffat, A., Bailey, P., Scholer, F., & Thomas, P. (2017). Incorporating user expectations and behavior into the measurement of search effectiveness. ACM Transactions on Information Systems (TOIS), 35(3), 1-38.

Montgomery, D. C. (2017). Design and analysis of experiments. John Wiley & Sons.

More time, tools, and details on the page experience update. (2021, April 19). Google Developers. Retrieved May 05, 2021, from https://developers.google.com/search/blog/2021/04/more-details-page-experience

Mowshowitz, A., & Kawaguchi, A. (2002). Bias on the Web. Communications of the ACM, 45(9), 56-60.

Najork M. (2009). Web Crawler Architecture. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), 175-220.

Nunnally, J.C. (1978). Psychometric Theory. McGraw-Hill, New York.

O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Ong, K., Järvelin, K., Sanderson, M., & Scholer, F. (2017, August). Using information scent to understand mobile and desktop web search behavior. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 295-304).

Orlikowski, W. J., & Iacono, C. S. (2001). Research commentary: Desperately seeking the "IT" in IT research—A call to theorizing the IT artifact. Information systems research, 12(2), 121-134.

Palos-Sanchez, P., & Saura, J. R. (2018). The effect of internet searches on afforestation: The case of a green search engine. Forests, 9(2), 51.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. Journal of computer-mediated communication, 12(3), 801-823.

Pasquinelli, M. (2009). Google's PageRank algorithm: A diagram of the cognitive capitalism and the rentier of the common intellect. In: Becker K and Stalder F (eds) Deep Search. London: Transaction Publishers, 152–162.

Patten, M. L., & Galvan, M. C. (2019). Proposing Empirical Research: A Guide to the Fundamentals (6th ed.). Taylor and Francis.

Pichai, S. (2020, April 1). Today Google stops funding climate change deniers. Retrieved March 01, 2021, from https://agreenergoogle.com/

Poulton, E. C. (1973). Unwanted range effects from using within-subject experimental designs. Psychological Bulletin, 80(2), 113.

Rhodes, A. E., Lin, E., & Streiner, D. L. (1999). Confronting the confounders: the meaning, detection, and treatment of confounders in research. The Canadian Journal of Psychiatry, 44(2), 175-179.

Röhle, T. (2009). Dissecting the Gatekeepers : Relational Perspectives on the Power of Search Engines. In Deep Search : The Politics of Search beyond Google (pp. 117–132). Innsbruck, Wien, Bozen.

Rose, D. E., & Levinson, D. (2004, May). Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web (pp. 13-19).

Rosenberg, D. (2018). The business of UX strategy. Interactions, 25(2), 26-32.

Saunders, M., Thornhill, A., & Lewis, P. (2019). Research Methods for Business Students. Pearson Education, Limited.

Schultheiß, S., & Lewandowski, D. (2020). "Outside the industry, nobody knows what we do" SEO as seen by search engine optimizers and content providers. Journal of Documentation, Vol. 77 No. 2, pp. 542-557.

Schultheiß, S., & Lewandowski, D. (2021). Misplaced trust? The relationship between trust, ability to identify commercially influenced results, and search engine preference. arXiv preprint arXiv:2101.09159.

Schulte-Mecklenbeck, M. & Huber, O. (2003). Information search in the laboratory and on the Web: With or without an experimenter. Behavior Research Methods, Instruments & Computers, 35(2), 227–235.

Schultz, C. D. (2020). Informational, transactional, and navigational need of information: relevance of search intention in search engine advertising. Information Retrieval Journal, 23(2), 117-135.

Seabold, S., & Perktold, J. (2010, June). Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference (Vol. 57, p. 61).

Sen, R. (2005). Optimal search engine marketing strategy. International Journal of Electronic Commerce, 10(1), 9-25.

Sharda, R., Delen, D., & Turban, E. (2016). Business intelligence, analytics, and data science: a managerial perspective. Pearson.

Sherman-Palladino, A. (Writer), Palladino, D (Writer) & Moore, Tom. (Director). (2003, November 11). Die, Jerk (Season 4, Episode 8) [TV series episode]. In A. Sherman-Palladino, D. Palladino, G. Polone (Executive Producers), Gilmore Girls. Dorothy Parker Drank Here Productions; Hofflund/Polone; Warner Bros. Television

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences?. International Journal of Information Management, 52, Article 102061.

Sieber, J. E. (1993). Ethical considerations in planning and conducting research on human subjects. Academic Medicine, 68(9), S9–S13.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2020). Database system concepts (Seventh edition). New York: McGraw-Hill.

Spink, A., & Saracevic, T. (1998). Human-computer interaction in information retrieval: nature and manifestations of feedback. Interacting with computers, 10(3), 249-267.

StatCounter. (November 1, 2020). Percentage of mobile device website traffic worldwide from 1st quarter 2015 to 3rd quarter 2020 [Graph]. In Statista. Retrieved March 05, 2021, from https://www-statista-com.esc-web.lib.cbs.dk:8443/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/

StatCounter. (February 10, 2021). Worldwide desktop market share of leading search engines from January 2010 to January 2021 [Graph]. In Statista. Retrieved March 05, 2021, from https://www-statista-com.esc-web.lib.cbs.dk:8443/statistics/216573/worldwide-market-share-of-search-engines/

Steiner, M., Magin, M., Stark, B., & Geiß, S. (2020). Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. Information, Communication & Society, 1-25.

Strzelecki, A., & Rutecka, P. (2020). Featured Snippets results in Google web search: an exploratory study. In Marketing and Smart Technologies (pp. 9-18). Springer, Singapore.

Sukamolson, S. (2007). Fundamentals of quantitative research. Language Institute, Chulalongkorn University, Bangkok, Thailand, 1, 2-3.

Suzuki, M., & Yamamoto, Y. (2020, November). Analysis of Relationship between Confirmation Bias and Web Search Behavior. In Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services (pp. 184-191).

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 48(6), 1273-1296.

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. International Journal of Academic Research in Management, 5 (3) (2016), pp. 28-36

Taylor, Z. W., & Bicak, I. (2020). Buying search, buying students: how elite US institutions employ paid search to practice academic capitalism online. Journal of Marketing for Higher Education, 30(2), 271-296.

Thode, H. C. (2002). Testing for normality (Vol. 164). New York: Marcel Dekker, 2002.

Thompson, B. (2004). Exploratory and confirmatory factor analysis. American Psychological Association.

Toepoel, V. (2015). Doing Surveys Online. SAGE Publications.

Treen, K. M. D. I., Williams, H. T., & O'Neill, S. J. (2020). Online misinformation about climate change. Wiley Interdisciplinary Reviews: Climate Change, 11(5), e665.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. Journal of counseling psychology, 51(4), 473–481.

Van Maanen, J. (2011). Tales of the field: On writing ethnography. University of Chicago Press.

Vallat, R. (2018). Pingouin: statistics in Python. Journal of Open Source Software, 3(31), 1026.

Vuong, T., Saastamoinen, M., Jacucci, G., & Ruotsalo, T. (2019). Understanding user behavior in naturalistic information search tasks. Journal of the Association for Information Science and Technology, 70(11), 1248-1261.

Waters, C. K. (2007). The nature and context of exploratory experimentation: An introduction to three case studies of exploratory research. History and Philosophy of the Life Sciences, 29(3): 275–284.

White, R. (2013, July). Beliefs and biases in web search. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 3-12).

Wolfram, D. & Xie, H. (2002). Traditional IR for Web users: a context for general audience digital libraries, Information Processing & Management, Vol. 38 No. 5, pp. 627-48.

Wu, J., & Wong, A. C. M. (2003). A note on determining the p-value of Bartlett's test of homogeneity of variances. Communications in Statistics-Theory and Methods, 32(1), 91-101.

Wu, W. C., Kelly, D., Edwards, A., & Arguello, J. (2012, August). Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In Proceedings of the 4th information interaction in context symposium (pp. 254-257).

Wu, W. C., Kelly, D., & Sud, A. (2014, July). Using information scent and need for cognition to understand online search behavior. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 557-566).

Xie, H. (2004). Online IR system evaluation: online databases versus Web search engines. Online information review, 28(3), 211-219.

Yigit, S., & Mendes, M. (2018). Which effect size measure is appropriate for one-way and two-way ANOVA models? A Monte Carlo simulation study. Revstat Statistical Journal, 16(3), 295-313.

York, R., & Clark, B. (2006). Marxism, positivism, and scientific sociology: Social gravity and historicity. The Sociological Quarterly, 47(3), 425-450.

Zhang, J., & Dimitroff, A. (2005). The impact of webpage content characteristics on webpage visibility in search engine results (Part I). Information Processing & Management, 41(3), 665-690.

Zhang, C., Zhang, X., & Wang, H. (2018, November). A machine reading comprehension-based approach for featured snippet extraction. In 2018 IEEE International Conference on Data Mining (ICDM) (pp. 1416-1421). IEEE.

Zhou, B., & Yao, Y. (2010). Evaluating information retrieval system performance based on user preference. Journal of Intelligent Information Systems, 34(3), 227-248.

Ziakis, C., Vlachopoulou, M., Kyrkoudis, T., & Karagkiozidou, M. (2019). Important factors for improving Google search rank. Future internet, 11(2), 32.