Featured Snippets and their Influence on Users' Credibility Judgements

Markus Bink
University of Regensburg
Regensburg, Germany
markus.bink@student.ur.de

Steven Zimmerman
University of Essex
Essex, United Kingdom
szimme@essex.ac.uk

David Elsweiler
University of Regensburg
Regensburg, Germany
david.elsweiler@ur.de

ABSTRACT

Search engines often provide featured snippets, which are boxed and placed above other results with the aim of directly answering user queries. To learn about how users judge the credibility of such results and how they influence search outcomes, a controlled web-based user study (N = 96) was conducted. Using resources made available by scholars in the community, we study featured snippets in a medical context with participants being tasked with determining whether a named treatment is helpful for a specified medical condition both before and after viewing the search results. Experimental conditions varied the presence and credibility of featured snippets. Our findings indicate that participants tend to overestimate the credibility of information in featured snippets. Featured snippets are, moreover, shown to often change users' opinion about a topic, especially if they are uncertain. Showing correct information inside featured snippets helped participants make more accurate decisions, whereas incorrect or contradicting information led to more harmful outcomes.

CCS CONCEPTS

• **Information systems → Search interfaces**.

KEYWORDS

Featured Snippets, Answer Module, Credibility, Web Search, Question Answering, Search Behaviour

ACM Reference Format:
Markus Bink, Steven Zimmerman, and David Elsweiler. 2022. Featured Snippets and their Influence on Users' Credibility Judgements. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22), March 14–18, 2022, Regensburg, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3498366.3505766

1 INTRODUCTION

Featured snippets are search results which are boxed and presented above other results with the aim of answering user queries directly [39] (see Figures 2 and 4 for examples). These snippets, which appear in 13% to 27% of searches [34], have been subject to much research [2, 5, 39, 40, 50] and are known to draw more attention than regular snippets [50], reduce the time spent searching [50], and increase user satisfaction [2, 50]. Other work has studied featured snippets in relation to the quality of information provided. For example, we know that featured snippets also reduce user interaction times when poor information is shown [50] and that users demonstrate a tendency to implicitly trust the information provided by featured snippets [2]. These findings suggest that including featured snippets on the results page may have consequences for user judgements, as well as search outcomes, but to date, our knowledge of these aspects is limited.

User search behaviour can be biased in many ways [1] and modern web search engines can amplify such biases via the imbalance of results returned [27, 47], the content contained within results [7, 47] and how the content is presented [4, 52]. Moreover, users have been shown to be both unreliable and uncertain when identifying credible information on the web [18, 35], a situation not helped by misconceptions about how rankings are created [16, 44]. Although past research has studied biases and credibility assessments in the context of SERPs, this work has tended to focus on vanilla presentations of results (analogous to '10 blue links') and does not necessarily reflect that of modern search engines [5, 50]. Accordingly, we build on past research to study how featured snippets are assessed, the potential biases they may have on users' search behaviour, and what this means for the outcome of searches. In particular we address the following research questions:

- **RQ1:** Are featured snippets more credible than regular snippets?
- **RQ2:** Do featured snippets influence the credibility judgements of regular result snippets?
- **RQ3:** How do featured snippets influence search outcomes?

To address these questions we perform a controlled study where participants provide credibility judgements in the context of a medical task. Before describing the study and results in detail, the following section summarises important related work.

2 RELATED WORK

The presented work builds on and is influenced by a large body of previous research. We summarise the most relevant in three sections. The first reviews how featured snippets are used and how they influence searches; the second summarises research relating to user credibility judgements; finally, we review research on how the search process, including user judgements, can be biased.

2.1 Direct Answers and Featured Snippets

As search engines have evolved, user expectations have progressed from being satisfied with the provision of relevant web pages to expecting answers to be supplied directly [38]. Providing users with

CHIIR '22, March 14–18, 2022, Regensburg, Germany
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9186-3/22/03…$15.00
https://doi.org/10.1145/3498366.3505766

direct answers can lead to the satisfaction of information needs without the user clicking to review any of the presented results [6, 48]. Therefore, much research has investigated which types of answers to display [5, 49, 54] as well as how to best present these [50, 54]. Providing direct answers, however, leads to a trade-off "between convenience and user experience on the one hand, and accuracy and retrievability on the other" [28] and some situations are better suited to this solution than others. For example, factoid questions lend themselves to the provision of direct answers [28] and these have been studied in diverse contexts ranging from web-search in desktop [2, 5, 50] and mobile [21, 49] settings to voice-controlled assistants [28, 41].

Providing answers directly changes both how users interact with the system and how they perceive the interaction. For example, they have been shown to enhance the user's perceived search experience [2, 50], shorten the time to complete tasks [50] and reduce user engagement with the SERP [5]. Moreover, eye-tracking studies have revealed that featured snippets generally attract more attention than regular result snippets, which changes how users interact [50].

People can and do use direct answers in the form of featured snippets for purposes beyond one-off question answering. For example, answers can be monitored for changes (as is the case for weather forecasts) and in such cases, the answers are often triggered by identical query phrasing [5]. Queries leading to direct answers are typically short (less than ten words) [39] and are often formulated as a grammatically correct question. The answers to these queries are typically sourced from prominent resources, such as Wikipedia and are most often presented in the form of a paragraph [40].

Yet, the quality of presented answers has been shown to vary. Person-related questions (e.g. "What kind of singer is Ice-T?") result in higher quality answers than other categories such as thing, organisation or event [54]. Further, questions starting with the interrogative word 'where' yield higher quality answers than those with 'who', 'what' or 'how'. This is worrying since identifying trustworthy information has been shown to be troublesome for users and varying quality makes this process even more difficult.

2.2 Credibility Judgements

While there is a lot of useful information on the internet, there are also many web pages that propagate misleading and erroneous information. Past research has shown that users are often both uncertain and inaccurate when judging which online content they can trust [18, 35]. Many search engine users believe search engines to be "a fair and unbiased source of information" [30] and have been shown to treat result ranking as a credibility indicator [16, 44].

Credibility judgements are influenced by a slew of different factors, ranging from the site's aesthetics and perceived professionalism [12] to the user's reading skills [15] and propensity to risk [22]. Different users base their judgements on different credibility cues, which can lead to wildly different judgements [18]. Moreover, even the same cue can be interpreted differently by different users depending on their experience, expertise and political views [18].

Thus, an ongoing research area is how to mitigate this issue and help users in their decision-making. One approach is to automatically estimate the credibility of web pages using machine learning approaches, which can be used to alter the ranking or nudge user choices in other ways [55]. Several predictive features have been tested with varying success ranging from those derived from a page's in and out links [25, 37] or signs of commercial interest [53] to the presentation [14, 25] or actual content of the text [23, 37]. For the latter, modern neural models have proved particularly effective [11].

A second approach is to give extra information about web pages to help the user make choices. For example, providing information about a page's PageRank and popularity has been shown to positively influence the accuracy of credibility judgements [35]. Creative presentation of such information, e.g. via a radar chart, has been shown to help topically familiar users in particular [51].

2.3 Biases in Web Search

A growing body of research highlights the multitude of biases influencing user search behaviour. A prime example is the ranking of results since users typically only examine a few high ranking listings [17]. Result lists, themselves, bias outcomes by typically containing more positive results and results favouring a particular view point [46], as well as being created in such a way that promotes already popular results [7, 26].

The composition and presentation of results are a further source of bias. For example, information scent theory and related empirical work have shown that the listings influence user click behaviour [4] and the properties of listings such as missing snippets, short snippets, missing query terms in titles or snippets and complex URLs negatively impact the probability of results being viewed [7].

Ranking and presentation biases are compounded by user biases, including confirmation bias, where users tend toward and seek information that confirms their prior beliefs and disregard contradicting information [19, 24]. The way search queries are formulated influences the results returned [20]. Lastly, anchoring bias has been shown to affect user interactions with documents [36]. Anchoring bias [43], in this context, is where users judge the credibility of results with respect to their impression of the first result they see.

2.4 Summary

The reviewed literature has highlighted that credibility judgements are challenging for users and biased in several ways. Direct answers and featured snippets have become a pervasive means to present results, yet the link between these aspects has not been studied. The presented research examines exactly this, looking at how the credibility of featured snippets are judged, what impact this has on how other listings are viewed, as well as what effect this has on overall search outcomes.

3 METHODOLOGY

To study the effect featured snippets have on users' credibility judgements, a controlled web-based experiment was conducted. Biases in SERPs are introduced by changing correctness of the information in a SERP's featured snippet. This resulted in one control condition, where 10 regular results were shown and three experimental conditions, where the featured snippet was manipulated by either showing a correct, incorrect or a contradicting answer (to the participants starting opinion) followed by 9 regular results.

Table 1: Medical treatments used throughout the study. Additionally, efficacy labels are provided based on the Cochrane Review. Further, participants answer distributions with respect to the medical treatment are shown.

T	Medical Treatment	Efficacy	Distribution of answers		
			Yes	No	Unsure
T1	Do antioxidants help female subfertility?	Unhelpful	14.3%	32.1%	53.6%
T5	Do sealants prevent dental decay in the permanent teeth?	Helpful	25.0%	14.3%	60.7%
T8	Does melatonin help treat and prevent jet lag?	Helpful	17.9%	21.4%	60.7%
T10	Does traction help low back pain?	Unhelpful	25.0%	14.3%	60.7%

A within-groups design was employed with every participant being exposed to all conditions. A Graeco-Latin Square was used to balance the order and combination of experimental conditions and treatments, as well as mitigate learning-effects.

3.1 Documents

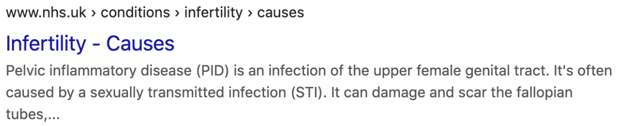
For our experiments we made use of a document collection in the medical domain made available by Zimmerman et al. [55]. We utilise these, since a large portion of the population uses the web to access health-related information [13] and the quality of health information may vary from site to site [9]. Different search engines such as Bing, Yahoo and Google were used to collect relevant documents for 10 medical treatments. The documents were labelled as either correct or incorrect based on the treatments' Cochrane review, which is *"a systematic review that synthesises the clinical evidence and informs clinical decision making"* [27]. A correct document contains information in line with the truth, whereas an incorrect document contradicts the truth or, in some cases, shows adverse side effects or harms of the treatment.

3.2 Topic Selection

A pre-study was conducted to identify controversial topics, i.e. those where the correct answer was not common knowledge and participants were uncertain about their answer. Using convenience sampling, 28 participants were recruited through a University mailing list. Topics related to whether a given treatments helped or not for a specified ailment. Participants were asked about 10 topics and could answer with either *Yes*, *No* or *Unsure*. Other than the treatment in question form (see Table 1), no further information was presented. 4 of the 10 topics were selected where the majority of participants' answers leaned towards *Unsure*. These topics along with the participant responses are given in Table 1.

3.3 Experimental Conditions

With these topics selected, 4 different configurations of SERPs were created:



www.nhs.uk › conditions › infertility › causes

Infertility - Causes

Pelvic inflammatory disease (PID) is an infection of the upper female genital tract. It's often caused by a sexually transmitted infection (STI). It can damage and scar the fallopian tubes,...

Figure 1: Regular snippet as used in the study. It consists of a URL on top, the page title below it in larger font size and blue text colour and the snippet description in medium grey on the bottom.

- **BASE:** Ten regular search results, analogous to "10 blue links".
- **FS_COR:** A correct featured snippet and 9 regular results.
- **FS_INCOR:** An incorrect featured snippet and 9 regular results.
- **FS_CONTR:** A featured snippet contradicting the participant's pre-task opinion and 9 regular results.

The last condition, in particular, examined the influence a contradicting featured snippet might have on users' decisions. For example, if users answered "Yes" to the question *"Does melatonin help and prevent jet lag?"* before seeing the SERP, the featured snippet showed an answer which suggests it does not help. To prevent bias in either direction, snippets were balanced to show 5 correct and 5 incorrect snippets and the order was randomly assigned.

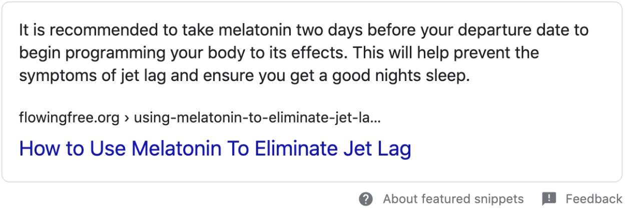
3.4 Snippet Creation

Snippets were generated from the raw HTML documents with all non-text content, such as HTML tags (e.g. *<script>*, *<style>* or *<div>*) being removed in a pre-processing stage. This was achieved with Python[1] and the Beautiful Soup library[2]. The final result listings consisted of the page title (extracted using the HTML *<title>* tag), its respective URL, and a short summary.

The summaries for the standard results were generated using a text-to-text transformer [31], specifically the $t5 - large$ model with its summarisation pipeline[3], on the pre-processed text for each document (see Figure 1). The summaries for featured snippets had to be generated in such a way that they answered the question (see Figure 2). To achieve this we utilised BERT (Bidirectional Encoder Representations from Transformers), specifically the *bert-large-uncased-whole-word-masking-finetuned-squad model* [10] available through the Hugging face community[4]. The specified model is pre-trained on the English language and fine-tuned on The Stanford Question Answering Dataset (SQuAD) [32], which improves reading comprehension of models.

Extracting answers that provided incorrect information proved to be especially difficult since some of the documents labelled as incorrect did not always contain false information and, at times, information that did not directly support or oppose the truth [27]. To remedy this issue, 2 correct and 2 incorrect answers were manually selected for each topic. 4 independent judges rated each answer, from a pool selected by the model in a previous step, based on

[1] https://www.python.org
[2] https://pypi.org/project/beautifulsoup4
[3] https://huggingface.co/t5-large
[4] https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad



It is recommended to take melatonin two days before your departure date to begin programming your body to its effects. This will help prevent the symptoms of jet lag and ensure you get a good nights sleep.

flowingfree.org › using-melatonin-to-eliminate-jet-la...

How to Use Melatonin To Eliminate Jet Lag

Figure 2: Featured snippet as used in the study. It consists of the snippet description in medium grey on top, the URL below it and the page title in larger font size and blue text colour on the bottom. Most noticeably is also the medium grey border around the snippet and the *About featured snippets* and *Feedback* buttons below the box.

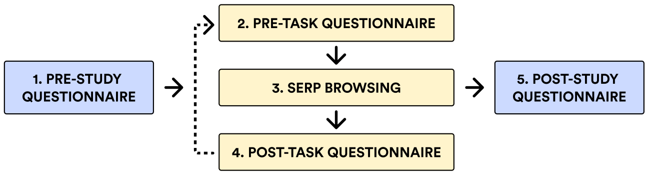
its usefulness with respect to answering the medical question. To ensure that the medical treatment in question was fully understood, definitions for both health issue and treatment were provided. These were initially selected by [27] from sites like Merriam-Webster[5] or the Mayo Clinic's[6] medical dictionaries.

Judges were instructed to rate the usefulness of each snippet independently from one another on a 7-point Likert scale ranging from *Useless* to *Useful*. A snippet would be rated useful if it contained information that would answer the medical question. For each answer, usefulness ratings were accumulated and those with the highest average usefulness rating were selected for the final study. This process resulted in 2 featured snippet summaries for each topic: one correct and one incorrect.

3.5 Experimental System

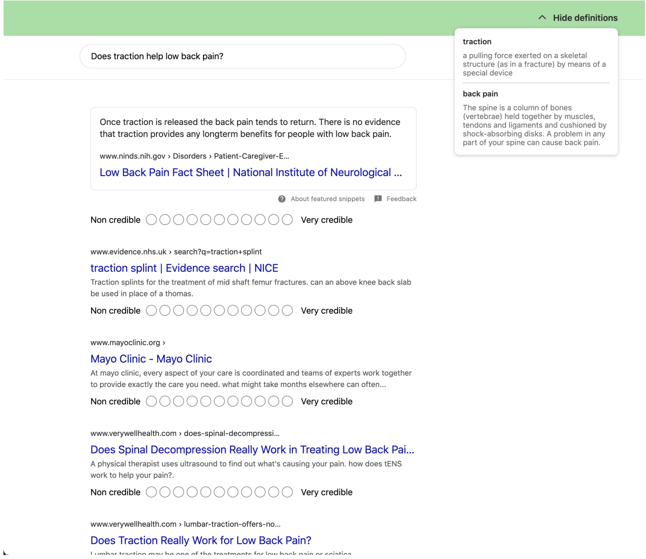
The main study flow can be seen in Figure 3. Participants were first provided with a pre-study questionnaire, where they completed an informed consent form and provided information on their prior knowledge and use of featured snippets. The main study included four topical search tasks. In a pre-task questionnaire, participants answered questions related to the topic, such as what they believed the answer to be and level of confidence in their belief. During the main task, a SERP appropriate to the condition was displayed where participants judged all results in terms of credibility (see Figure 4). Upon completion of the main task, a post-task questionnaire was completed with the same questions used in the pre-task questionnaire about answer beliefs and confidence in their answer. After completion of the main study (four tasks), a post-study questionnaire was used to capture participant demographics and impressions of the study.

The SERPs were designed to be similar to those commercially available e.g. Google or Bing (see Figure 4). In the baseline condition, 10 regular result snippets were shown. In the three experimental conditions, a SERP *with* 1 featured snippet on top and 9 regular snippets below was shown (see Figure 4), with the featured snippet containing a correct or incorrect answer as appropriate. Below each result listing, an 11-point Likert scale ranging from "Non-credible" to "Very credible" allowed participants to submit their judgement of the credibility of each SERP result. This allows the participant to judge results intuitively and offers the opportunity to split participants into smaller groups post-hoc during data analysis. Inspired by the method applied by similar studies in the literature, participants were provided with medical definitions of the key terms (see [27] and [55]). These were placed in a dropdown field on the top right-hand side of the screen. Thus, if users were not familiar with specific medical terms, they were given the opportunity to look them up in this small panel. However, if users did not wish to see the panel, they could also hide it by clicking an arrow. To ensure this information was accessible even when scrolling the results, the definitions were made 'sticky', i.e. they keep their position on the screen as users scroll such that the search results themselves were not obscured.



1. PRE-STUDY QUESTIONNAIRE
2. PRE-TASK QUESTIONNAIRE
3. SERP BROWSING
4. POST-TASK QUESTIONNAIRE
5. POST-STUDY QUESTIONNAIRE

Figure 3: Main flow of the user study.



Does traction help low back pain?

traction

a pulling force exerted on a skeletal structure (as in a fracture) by means of a special device

back pain

The spine is a column of bones (vertebrae) held together by muscles, tendons and ligaments and cushioned by shock-absorbing disks. A problem in any part of your spine can cause back pain.

Once traction is released the back pain tends to return. There is no evidence that traction provides any longterm benefits for people with low back pain.

www.ninds.nih.gov › Disorders › Patient-Caregiver-E...

Low Back Pain Fact Sheet | National Institute of Neurological ...

Non credible ○○○○○○○○○○○ Very credible

www.evidence.nhs.uk › search?q=traction+splint

traction splint | Evidence search | NICE

Traction splints for the treatment of mid shaft femur fractures. can an above knee back slab be used in place of a thomas.

Non credible ○○○○○○○○○○○ Very credible

www.mayoclinic.org ›

Mayo Clinic - Mayo Clinic

At mayo clinic, every aspect of your care is coordinated and teams of experts work together to provide exactly the care you need. what might take months elsewhere can often...

Non credible ○○○○○○○○○○○ Very credible

www.verywellhealth.com › does-spinal-decompressi...

Does Spinal Decompression Really Work in Treating Low Back Pai...

A physical therapist uses ultrasound to find out what's causing your pain. how does tENS work to help your pain?.

Non credible ○○○○○○○○○○○ Very credible

www.verywellhealth.com › lumbar-traction-offers-no...

Does Traction Really Work for Low Back Pain?

Lumbar traction may be one of the treatments for low back pain or sciatica

Figure 4: Search Engine Results Page used in the study.

To distribute the study, the crowdsourcing platform Prolific was used. An attention check was added to the study, as suggested by [29], to ensure paid participants took participation seriously. Further, a pre-test was run to determine the duration of the study and thus compensate participants adequately [7]. To ensure that language barriers were not an issue and that the provided information would be understood, only those living in the United Kingdom who are fluent in English were eligible to partake in the study.

[5] https://www.merriam-webster.com

[6] https://www.mayoclinic.org

[7] Participants were paid 1.88 GBP for an expected duration of 15 minutes.

3.6 Participants

A total of 96 participants, aged between 18 and 68 ($M = 37.30, SD = 13.47$) were acquired via the crowdsourcing platform Prolific[8]. To ensure that neither instructions nor page contents were misunderstood, only participants from the UK, fluent in English, were recruited for the study. 55 identified as male, 40 as female and 1 as diverse. 34 participants had a bachelor's degree and 10 a master's degree, 47 had obtained a high school diploma, 5 described themselves as having no high school diploma or had some other form of education. 17 of the participants were students, while others had wide ranging employment, from accountants to software engineers and beyond (69). 10 reported being retired or currently unemployed.

Participants reported using search engines to look up health-related information ($M = 6.2, SD = 1.07$), and tend to be more trustworthy of the information shown inside featured snippets ($M = 4.57, SD = 1.15$). These constructs were measured on a 7-point Likert scale ranging from "Strongly disagree" to "Strongly agree". They also stated that they know about featured snippets in search engines ($\sum(Yes) = 90$; $\sum(No) = 6$). When asked about their average featured snippet usage, 6 participants reported using featured snippets multiple times per day, 5 use them once per day, 38 multiple times per week, 36 once per week and 11 mentioned never using them.

Before seeing the Search Engine Result Page (SERP), participants were asked how familiar they were with the presented topics. Measurements were done using a 7-point Likert scale ranging from "Not familiar" to "Very familiar". Generally, their prior knowledge was rather low across all topics (T1: $M = 1.78, SD = 1.34$; T5: $M = 2.56, SD = 1.68$; T8: $M = 3.0, SD = 1.85$; T10: $M = 2.49, SD = 1.6$).

4 RESULTS

The following section describes the results obtained to answer the outlined research questions. The data was analysed using the statistical software R [42]. Statistical significance of results was assumed at $p = .05$.

4.1 Credibility of Featured Snippets (RQ1)

To analyse to what extent featured snippets are perceived as more credible than regular snippets, participants' credibility judgements of featured snippets were compared to those of regular snippets in the same position (i.e. the first result on the SERP). Judgements were recorded on an 11-point Likert scale ranging from *Non-credible* to *Very credible*. Judgements above 7 were treated as Credible, judgements below 5 as Non-Credible and in between as Unsure. Using this split, subjects classified 60.8% of the featured snippets as Credible, 12.5% Non-Credible and 26.7% as Unsure. To compare judgements of featured snippets with regular snippets in the same position, judgements in the experimental conditions were averaged over all conditions for each participant. This resulted in two credibility values per user, one for the baseline snippet and one for the featured snippets. Due to non-normality of data, a Wilcoxon rank-sum test was conducted, with featured snippets ($M = 7.8, SD = 1.78$) being judged significantly more credible than regular snippets ($M = 6.49, SD = 3.24$) in the top-most position ($W = 5497, p = .021$).

Post-hoc analyses compared the credibility of featured snippets containing correct answers ($M = 6.83, SD = 2.59$) to those containing incorrect and contradicting answers. In both cases, featured snippets containing correct information were judged significantly less credible than incorrect featured snippets ($M = 8.53, SD = 2.47, W = 2783.5, p < .001$) and contradicting featured snippets ($M = 8.03, SD = 2.41, W = 3357, p < .01$).

4.2 Judgement Accuracy (RQ2)

Table 2 presents the relationship between participant credibility judgements and ground-truth document assessments as in [18]. The accuracy of participant credibility judgements was analysed for all four conditions. For consistency with the binary ground-truth labels of the documents, participant credibility judgements were converted to categorical labels. Judgements with a credibility rating above 7 were labelled as *Credible*, judgements below 5 as *Non-Credible* and judgements in between as *Unsure*. A judgement is rated as correct if it aligns with the ground-truth. Accuracies across conditions were relatively consistent (BASE: 33.8%, FS_COR: 32.8%, FS_INCOR: 32.8%, FS_CONTR: 32.4%). Removal of the judgements labelled as *Unsure* increased accuracy, but was still rather low (BASE: 47.6%, FS_COR: 48.6%, FS_INCOR: 47.7%, FS_CONTR: 47.8%). Participants consistently rated snippets more credible (42.19%) than non-credible (26.51%), with the remaining snippets rated as unsure (31.30%).

These results contradict previous findings (see [27]) that biasing search results towards correct or incorrect does impact user judgements. This discrepancy motivated the following post-hoc analyses.

4.2.1 Accuracy of Document Source. The documents can be grouped into two categories, those with a credible or trustworthy source and those without a credible source. For instance, documents from credible sources were taken from institutions like the "National Health Service" or the "National Center for Biotechnology Information", whereas documents from non-credible sources were from pages such as "ebay.co.uk" or "canstockphoto.co.uk". A grand mean of credibility judgements was calculated for each specific page across all users.

Sources with an average credibility judgement of below 6.5 were grouped in the non-credible category, whereas everything above 6.5 were assigned to the credible category.

Using this approach, credible source documents were consistently rated as more credible regardless of its ground-truth correctness (Credible: $M = 7.67$, $SD = 2.74$; Non-Credible: $M = 4.98$, $SD = 2.86$). Conducting a Welch's test, this difference was found to be statistically significant, $t(3455.2) = -29.345, p < .001, d = -0.96$.

4.2.2 User Confidence. Judgements were next analysed based on participants' confidence in their answer [9].

Participant judgements were split into two groups (confident / unconfident) based upon a mid-point split of 7-point Likert scale pre-task response for answer confidence.

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

[9] A general trend of how users' pre-answer confidence influences the accuracy of users' credibility judgements can be seen in Figure 5a, where user confidence is positively linked to accuracy.

Table 2: Relationship between participant judgements and objective judgements across conditions. Judgements above 7 were treated as credible, judgements below 5 as non-credible and judgements in-between as unsure. Each cell provides the amount of agreement between the ground-truth judgements and ones provided by participants, i.e. the cell in the second row of the Baseline column shows that 23.12% non-credible snippets were judged as credible.

Ground truth	Judgement	BASE	FS_COR	FS_INCOR	FS_CONTR	Overall
Non-credible	Non-credible	142 (14.79%)	133 (13.85%)	130 (13.54%)	134 (13.96%)	539 (14.04%)
Non-credible	Credible	222 (23.12%)	216 (22.5%)	233 (24.27%)	223 (23.23%)	894 (23.28%)
Non-credible	Unsure	116 (12.08%)	131 (13.65%)	117 (12.19%)	123 (12.81%)	487 (12.68%)
Credible	Non-credible	134 (13.96%)	117 (12.19%)	112 (11.67%)	116 (12.08%)	479 (12.47%)
Credible	Credible	182 (18.96%)	182 (18.96%)	185 (19.27%)	177 (18.44%)	726 (18.91%)
Credible	Unsure	164 (17.08%)	181 (18.85%)	183 (19.06%)	187 (19.48%)	715 (18.62%)
$\sum$		960	960	960	960	3,840

Table 3: Accuracy of participants' judgements of SERP snippets based on the confidence of users' answer certainty and topical familiarity.

Group	BASE	FS_COR	FS_INCOR	FS_CONTR
Confident User	35.5%	37.7%	33.4%	34.0%
Non–confident User	34.0%	30.4%	32.4%	32.3%
Topically Familiar	32.5%	37.5%	30.0%	28.7%
Topically Unfamiliar	34.7%	33.1%	34.2%	33.2%

Accuracy for both groups was calculated and is depicted in Table 3. Judgements at the median value (i.e. unsure rating) were excluded from the accuracy calculation. As the Table shows, accuracy in both groups is relatively similar. Although confident users' accuracy follows the featured snippets direction (i.e. higher accuracy when a correct answer is shown, lower when an incorrect is shown), this difference between the two groups is not significant. Only the difference in the FS_CORRECT condition was rendered significant ($\chi^2(1) = 4.12, p = .042$).

4.2.3 Topic Familiarity. Accuracy as it relates to a participants' subjective topic familiarity prior to each search task was also explored. A visual trend of how participants' self-assessed topical knowledge influences the accuracy of participants' credibility judgements is provided in Figure 5b, where self-reported topical knowledge is negatively correlated with answer accuracy.

Similar to the previously seen analysis, two groups were formed based on the median split of participants' topic familiarity. People with topic familiarity below 4 were grouped as unfamiliar and those above 4 as familiar. Table 3 depicts the resulting accuracy for each condition. While unfamiliar users generally outperform the familiar group, a χ^2-test found no statistically significant differences.

4.3 Search Outcome Accuracy (RQ3)

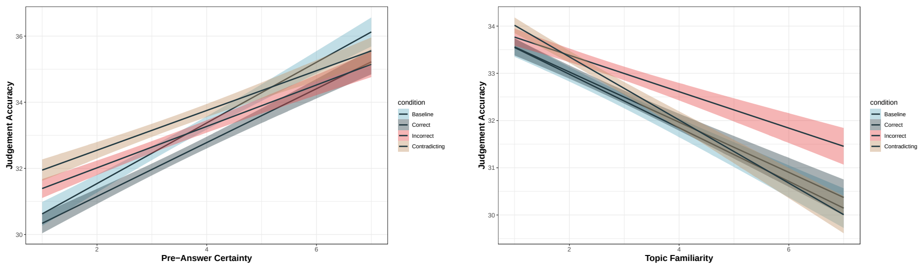
To understand the accuracy of users' search outcomes, participants were asked before and after seeing the SERP whether the provided treatment helped or not. Given the categorical nature of the dependant variable the McNemar Chi-square test was used for analyses, with results provided in Table 4.When participants were presented with a SERP that contained a correct featured snippet (FS_COR), accuracy increased by 14.6%. However, showing participants featured snippets with contradictory information (FS_CONTR) resulted in a decrease of 6.2%. Incorrect featured snippets (FS_INCOR) produced a strong negative effect, with a 30.2% decrease in accuracy of participant responses. Though the baseline SERP (BASE) had a decrease in accuracy, the magnitude of this difference is minimal compared to experimental variants. As the analyses are statistically significant, featured snippets (most notably correct and incorrect snippets) appear to impact how a user answers important medical questions.

Table 4: Accuracy of Pre and Post-SERP answers (given as percentage correct) across experimental conditions. Results of McNemar Chi-square test and Cohen's g effect sizes are included.

Condition	Pre	Post	χ^2	p	g
BASE	69.8%	65.6%	6.34	.01	-0.08
FS_COR	60.4%	75.0%	56.83	<.001	0.21
FS_INCOR	62.5%	32.3%	225.73	<.001	-0.39
FS_CONTR	60.4%	54.2%	6.69	<.01	-0.06

4.4 Additional Analyses

4.4.1 Answer Transition. Since the previously mentioned results suggest that featured snippets influence participant decision making, further analysis was conducted to determine how this change occurred. To analyse this, participants were split into two groups based on their pre-SERP answer certainty, which was measured on a 7-point Likert scale, ranging from *Not Confident* to *Confident*. Participants with a pre-answer certainty of less than 4 were treated as non-confident and participants with a pre-answer certainty of above 4 as confident users. Similar to the previous section, this was also done based on participants' topic familiarity. Change in question-answer was determined by comparing their pre-SERP answer with the post-SERP answer. If the post-SERP answer was different from the pre-SERP one, the participant changed their opinion. Table 5 shows changes in per cent. Across all conditions, users



(a)

(b)

Figure 5: Judgement accuracy based on participants' answer confidence (5a) and topic familiarity (5b).

Table 5: Fraction of changes comparing pre and post-SERP answers about the efficacy of a treatment.

Group	Baseline	Correct	Incorrect	Contradicting	Overall
Confident participant	16.1%	19.4%	25.0%	42.9%	26.4%
Non-confident participant	36.2%	50.0%	49.0%	66.7%	50.5%
Topically Familiar	37.5%	18.8%	29.4%	40.0%	31.2%
Topically Unfamiliar	23.5%	40.8%	41.7%	57.3%	41.3%

Table 6: Fraction of harmful decisions comparing their post SERP answer with the Cochrane answer.

Group	Baseline	Correct	Incorrect	Contradicting	Overall
Confident Participant	38.7%	29.0%	43.8%	48.6%	40.3%
Non-confident Participant	34.0%	21.7%	80.4%	45.8%	46.4%
Topically Familiar	50.0%	37.5%	47.1%	33.3%	42.2%
Topically Unfamiliar	32.4%	19.7%	75.0%	46.7%	43.7%

that had lower confidence in the prior answer, i.e. describing unconfident users, changed their opinion more than confident users. This difference is significant ($\chi^2(1) = 185.52, p < .001$). According to Cramer [33], this effect was moderate ($v = .24$). The same is true for topically unfamiliar users ($\chi^2(1) = 21.52, p < .001$) obtaining a negligible effect ($v = .08$). They also change their answer more than topically familiar users. However, changes in the baseline condition were greater for topically familiar users.

Similar to [27] and [55], the proportion of harmful decisions was analysed. A harmful decision is defined as a post-SERP answer that contradicts the ground-truth Cochrane Answer. Results are depicted in Table 6. Non-confident participants made more harmful decisions, most noticeably when an incorrect featured snippet was shown. This is also the case for topically unfamiliar people. Generally, harmful decisions were higher when an incorrect answer was shown. The opposite is true when a correct answer was shown, which decreases the number of participants' harmful decisions. Comparing harmful decisions overall, the number of harmful decisions differ significantly between confident and non-confident participants ($\chi^2(1) = 11.2, p < .001$). This effect was negligible ($v = .06$). However, harmful decisions do not differ significantly between topically familiar and unfamiliar users ($\chi^2(1) = 0.43, p < .51$).

Table 7: Confidence in participant answer pre and post seeing the SERP for different conditions. Further, results of a Wilcoxon signed-rank test are depicted. The last column depicts the calculated effect size using Cohen's d [8].

Condition	Pre Certainty		Post Certainty		V	p	d
	M	SD	M	SD			
BASE	3.61	1.67	4.41	1.57	177910	<.001	0.48
FS_COR	3.54	1.76	4.89	1.46	247660	<.001	0.83
FS_INCOR	3.36	1.88	4.89	1.38	240865	<.001	0.77
FS_CONTR	3.53	1.85	4.57	1.41	207515	<.001	0.55

4.4.2 Answer Confidence. Since the previous results suggest that participant answers are influenced by the presented SERP and featured snippet, an analysis regarding participant confidence was conducted. In this case, the participants' pre-answer certainty was compared to their post-answer certainty, meaning after being exposed to different SERP conditions. Table 7 shows the results of this analysis. Participant confidence after being exposed to SERPs increased in all conditions, no matter whether a correct, incorrect or contradicting snippet was shown. Wilcoxon signed-rank tests confirm these results as significant.

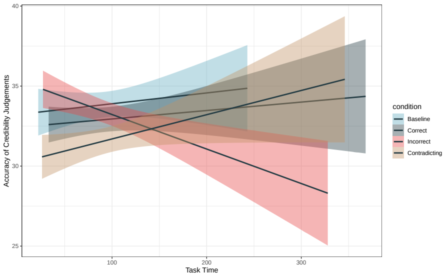


Figure 6: Relationship of task time and participants' accuracy in judging the credibility of search results.

4.4.3 Task completion time. Considering that users spend less time finding a suitable answer when a featured snippet is present [50], the time (measured in seconds) to complete each search task was also analysed. A one-way ANOVA was conducted to examine the effect of different conditions on participant task completion time. The results show that the task completion time did not differ significantly between different types of conditions (BASE: $M = 93.1, SD = 46.4$; FS_COR: $M = 96.5, SD = 59.9$; FS_INCOR: $M = 95.0, SD = 51.4$; FS_CONTR: $M = 97.5, SD = 51.8$; $F(3, 3836) = 1.27, p > .28$.

Further analysis was done to identify if judgement accuracy was influenced by how long participants took their time. Figure 6 shows the time for each condition and the resulting accuracy. For conditions BASE, FS_COR and FS_CONTR the graph shows a slight positive slope indicating that the longer participants took to assess the credibility of search results, the better their accuracy was. However, in the FS_INCOR condition, which showed a wrong answer, the slope is negative and much steeper, indicating that participant accuracy got worse the longer they took. Splitting participants into two groups based on the median task time and comparing their accuracy reveals no significant difference ($\chi^2(1) = 0.26, p = .61$).

5 DISCUSSION

In this section we summarise the findings and discuss these with respect to our research questions and what the answers mean for our community and the design of search engine result pages.

RQ1: Our findings suggest that the credibility of featured snippets is judged differently to regular snippets. Overall, our participants seemed to believe what they read in featured snippets, with participants rating these as credible in almost two thirds of cases. Featured snippets were judged to be more credible on average than regular snippets, even when comparing with the first (top) result in the baseline condition. The trends uncovered are quite worrying. In particular, featured snippets that provide inaccurate information - even when they contradict the participant's original opinion - were judged on average to be more credible than correct results.

RQ2: We found little evidence to suggest that the featured snippets, irrespective of the accuracy of the information they contain, influenced the credibility judgements for regular snippets. No statistical difference was found in the accuracy of judgements across the experimental conditions.

RQ3: Despite the lack of differences in credibility judgements across conditions, our findings show significant differences in terms of the outcome of searches. Whereas in the baseline condition no great change of opinion was found between before and after results were viewed - as would be expected with balanced results - when a correct featured snippet was shown, participants jumped from holding a correct view in 60.4% of cases, pre-task to 75% of cases, post-task. An even larger effect, in the opposite direction, was found when incorrect information was provided in featured snippets. In this condition, participants held the correct opinion 62.5% of the time before reviewing results, but only in 32.3% of cases after seeing the SERP with the incorrect featured snippet. A further troubling finding with respect to search outcomes was that participant confidence in their answer increased post-task, irrespective of condition or whether or not they were correct.

We were able to provide findings that complement our original research questions. For example, we discovered that regardless of experimental condition, participants were biased toward rating results as credible. This is again a troubling finding, which differs from more cautious behaviour reported in the literature [18]. Our results build on but can be distinguished from those by Pogacar et al. [27]. Whereas in [27], bias was identified when a large proportion of vanilla results were either correct or incorrect (i.e. in a 8/2 split), in our study, participants' judgements were shown to be biased by a single prominent snippet. This demonstrates the importance a featured snippet can have on users' search outcomes due to its salient position and separation from the rest of the results.

Moreover, participant confidence and topical familiarity seem to have impacted the judgements provided. Confident participants stuck to their original, pre-task opinions more often corroborating past findings [45]. Nevertheless, more than one third of these participants, changed answers when presented with contradicting information. A similar pattern could be observed for topic familiarity. Non-confident or topically unfamiliar participants seem to be strongly influenced by the answers presented in featured snippets and, as a result, chose a harmful outcome in 80% and 75% of the cases, respectively. This is, again, a worrying finding since the overwhelming evidence suggests that people fail to identify correct and incorrect information.

Our findings overall have clear implications for the use of featured snippets and direct answers in SERPs. Featured snippets seem to be a powerful means to alter user viewpoints. Confirmation bias is known to be a problem generally in information seeking [24] and changing user opinion to the correct answers has been shown to be difficult in the medical domain [45]. Nevertheless our findings highlight strong risks with featured snippets since they can shift user opinion, both positively and negatively. Consequently, featured snippets should only be shown when the system can be confident that the presented answer is factually correct.

If search engines are able to present a correct information in feature snippets, not only would this increase the likelihood of users ending searches with correct information, it could also help improve the search process since past work has shown featured snippets to reduce search time and increase user satisfaction [50].

Limitations:

Before presenting our conclusions, it is important to acknowledge that this work has limitations. These include that participants could only see search results as presented on SERPs and did not have the chance to click on individual sites to gain more information. The justification here was that our aim was to isolate possible effects originating from featured snippets. To achieve this and to make the findings comparable with past results (e.g. those presented in [18, 27, 55]), the experimental design employed followed the approaches in those credibility studies. A limitation associated with all of these studies is that participants did not address their own information needs or those associated with a simulated work task [3], but were provided with a simple question and asked to assess results on pre-determined results pages.

6 CONCLUSION AND FUTURE WORK

The presented study investigated how the use of direct answers in SERP featured snippets may influence user credibility judgements and bias answer outcomes. This was tested using a web-based experiment, in which participants were presented with different SERP variations. The SERP either displayed 10 regular search result snippets (as a baseline) or an experimental condition, where the top-most snippet was a featured snippet with either a correct, incorrect or contradicting answer. Participants were asked before and after seeing the SERP whether a particular medical treatment would be helpful or not for a given condition. The results show that participants based their answers on the information found in featured snippets and judge the information inside these snippets to be significantly more credible than that found in the top-ranked regular result. This finding is reinforced by how participants were influenced by the type of answer presented in the featured snippet. When a correct featured snippet was presented, participants post-task viewpoint was much more likely to align with the correct answer and the opposite was true when incorrect information was presented in the featured snippet. Showing contradicting information, i.e. opposing participant's pre-task belief, was shown to often change the participant's view. This effect was particularly prominent when participants were not confident in their initial view or lacked topically familiarity. On the other hand, featured snippets seem to have had a limited influence on the credibility judgements participants applied to individual results on the SERP. Participants who were confident in their answer or believe themselves to be topically familiar seem to be more influenced by the information presented in the featured snippet. This could hint that non-confident users or topically unfamiliar are more careful in their judgements, whereas confident or topically familiar people are not.

Results show that users implicitly trust the information shown inside featured snippets regardless of its correctness. This is very worrying since, in important contexts, such as the medical one studied here, correct information is vital and incorrect information can be potentially very harmful. Therefore, future research should aim to discover in which situations featured snippets may be beneficial with little to no risk attached and in which situations it would be better to show no featured snippet at all. Another possible direction is risk mitigation through phrasing answers in a less absolute way, conveying uncertainty or through providing multiple answers [50], which lets users compare results and helps them in their decision-making process.

REFERENCES

[1] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*. 27–37.

[2] Michael S. Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct Answers for Search Queries in the Long Tail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. 237–246.

[3] Pia Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research* 8, 3 (2003), 8–3.

[4] Stuart K. Card, Peter Pirolli, Mija Van Der Wege, Julie B. Morrison, Robert W. Reeder, Pamela K. Schraedley, and Jenea Boshart. 2001. Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. 498–505.

[5] Lydia B. Chilton and Jaime Teevan. 2011. Addressing People's Information Needs Directly in a Web Search Result Page. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. 27–36.

[6] Aleksandr Chuklin and Pavel Serdyukov. 2012. Good Abandonments in Factoid Queries. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. 483–484.

[7] Charles L. A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. 2007. The Influence of Caption Features on Clickthrough Patterns in Web Search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. 135–142.

[8] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

[9] Lubna Daraz, Allison S. Morrow, Oscar J. Ponce, Bradley Beuschel, Magdoleen H. Farah, Abdulrahman Katabi, Mouaz Alsawas, Abdul M. Majzoub, Raed Benkhadra, Mohamed O. Seisa, et al. 2019. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. *Journal of general internal medicine* 34, 9 (2019), 1884–1891.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Marcos Fernández-Pichel, David E Losada, Juan C Pichel, and David Elsweiler. 2021. Comparing Traditional and Neural Approaches for Detecting Health-Related Misinformation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 78–90.

[12] B. J. Fogg. 2003. Prominence-Interpretation Theory: Explaining How People Assess Credibility Online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. 722–723.

[13] Susannah Fox and Maeve Duggan. 2013. Health online 2013. *Health* 2013 (2013), 1–55.

[14] Alexandru L Ginsca, Adrian Popescu, and Mihai Lupu. 2015. Credibility in information retrieval. *Foundations and Trends in Information Retrieval* 9, 5 (2015), 355–475.

[15] Carolin Hahnel, Frank Goldhammer, Ulf Kröhne, and Johannes Naumann. 2018. The role of reading skills in the evaluation of online information gathered from search engine environments. *Computers in Human Behavior* 78 (2018), 223–234.

[16] Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas. 2010. Trust online: Young adults' evaluation of web content. *International journal of communication* 4 (2010), 27.

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Trans. Inf. Syst.* (April 2007), 7–es.

[18] Markus Kattenbeck and David Elsweiler. 2019. Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management* (2019).

[19] Joshua Klayman and Young-Won Ha. 1987. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review* 94, 2 (1987), 211.

[20] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. 417–432.

[21] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good Abandonment in Mobile and PC Internet Search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. 43–50.

[22] D. Harrison McKnight and Charles J. Kacmar. 2007. Factors and Effects of Information Credibility. In *Proceedings of the Ninth International Conference on Electronic Commerce (ICEC '07)*. 423–432.

[23] Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging Joint Interactions for Credibility Analysis in News Communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. 353–362.

[24] Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[25] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. 2013. Web Credibility: Features Exploration and Credibility Prediction. In *Proceedings of the 35th European Conference on Advances in Information Retrieval (ECIR'13)*.

557–568.

[26] Sandeep Pandey, Kedar Dhamdhere, and Christopher Olston. 2004. WIC: A General-Purpose Algorithm for Monitoring Web Information Sources. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30 (VLDB '04)*. 360–371.

[27] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. 209–216.

[28] Martin Potthast, Matthias Hagen, and Benno Stein. 2021. The Dilemma of the Direct Answer. *SIGIR Forum* 54, 1, Article 14 (2021), 12 pages.

[29] Prolific. 2021. Using attention checks as a measure of data quality. https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Using-attention-checks-as-a-measure-of-data-quality

[30] Kristen Purcell, Lee Rainie, and Joanna Brenner. 2012. Search engine use 2012. (2012).

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[32] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[33] Louis M. Rea and Richard A. Parker. 2014. *Designing and conducting survey research: A comprehensive guide*. John Wiley & Sons.

[34] Kristina Sam-Martin. 2020. Google's Featured Snippets in the Context of Strategic Content Marketing. (2020). https://sam-martin.at/downloads/featured-snippets-and-strategic-content-marketing-ebook-2020/

[35] Julia Schwarz and Meredith Morris. 2011. Augmenting Web Pages and Search Results to Support Credibility Assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. 1245–1254.

[36] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and Adjustment in Relevance Estimation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. 963–966.

[37] Parikshit Sondhi, V. G. Vinod Vydiswaran, and Cheng Xiang Zhai. 2012. Reliability Prediction of Webpages in the Medical Domain. In *Proceedings of the 34th European Conference on Advances in Information Retrieval (ECIR'12)*. 219–231.

[38] Sofia Stamou and Efthimis N. Efthimiadis. 2010. Interpreting User Inactivity on Search Results. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval (ECIR'2010)*. 100–113.

[39] Artur Strzelecki and Paulina Rutecka. 2020. Direct Answers in Google Search Results. *IEEE Access* 8 (2020), 103642–103654.

[40] Artur Strzelecki and Paulina Rutecka. 2020. Featured snippets results in Google Web search: An exploratory study. In *Marketing and Smart Technologies*. Springer, 9–18.

[41] Danny Sullivan. 2021. A reintroduction to Google's featured snippets. https://blog.google/products/search/reintroduction-googles-featured-snippets

[42] R Core Team et al. 2013. R: A language and environment for statistical computing. (2013).

[43] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.

[44] Julian Unkel and Alexander Haas. 2017. The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology* 68, 8 (2017), 1850–1862.

[45] Ryen White. 2013. Beliefs and Biases in Web Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. 3–12.

[46] Ryen W. White. 2014. Belief Dynamics in Web Search. *J. Assoc. Inf. Sci. Technol.* (2014), 2165–2178.

[47] Ryen W. White and Ahmed Hassan. 2014. Content Bias in Online Health Search. *ACM Trans. Web* (2014), 33 pages.

[48] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting Good Abandonment in Mobile Search. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. 495–505.

[49] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Is This Your Final Answer? Evaluating the Effect of Answers on Good Abandonment in Mobile Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 889–892.

[50] Zhijing Wu, Mark Sanderson, B. Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*. 1635–1644.

[51] Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing Credibility Judgment of Web Search Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. 1235–1244.

[52] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. 1011–1018.

[53] Wei Zha and H Denis Wu. 2014. The Impact of Online Disruptive Ads on Users' Comprehension, Evaluation of Site Credibility, and Sentiment of Intrusiveness. *American Communication Journal* 16, 2 (2014).

[54] Yiming Zhao, Jin Zhang, Xue Xia, and Taowen Le. 2019. Evaluation of Google question-answering quality. *Library Hi Tech* (2019).

[55] Steven Zimmerman, Alistair Thorpe, Chris Fox, and Udo Kruschwitz. 2019. Privacy Nudging in Search: Investigating Potential Impacts. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. 283–287.