The Impact of Web Search Result Quality on Decision-Making

Jan Heinrich Merker[1][0000−0003−1992−8696], Lena Merker[2][0009−0003−0523−5167],
and Alexander Bondarenko[3,1][0000−0002−1678−0094]

[1] Friedrich-Schiller-Universität Jena, Germany
{heinrich.merker,alexander.bondarenko}@uni-jena.de
[2] Martin-Luther-Universität Halle-Wittenberg, Germany
[3] Leipzig University, Germany

Abstract People often search the Web for answers to comparative questions like "Is pasta healthier than pizza?" to inform everyday decisions. However, web search engines sometimes may return biased or low-quality results. Still, previous research has not considered the impact of varying search result quality, relevance, or stance on the users' decision-making process. To close this gap, we conducted a user study on quality, relevance, and stance assessments of 120 Google search results retrieved for eight comparative questions. We asked study participants about their decision and confidence before and after seeing the top-4 search results and which results influenced their decision. Our study showed that (1) high-quality search results are more likely to influence a user's decision, (2) topical relevance and search result quality have a similarly strong impact on decision-making, and (3) search results are more likely to influence decisions for factual comparative questions than for subjective questions.

1 Introduction

Decision-making is an integral part of everyday life when weighing pros and cons for simple questions like "Should I eat sandwiches or cereal for breakfast?" or more critical questions like "Is buying a house better than renting?" [3, 23]. Nowadays, decisions are not only supported by prior knowledge and experience [26] but also by facts and arguments retrieved from the Web, e.g., when comparative questions are used as queries in web search [8]. While many studies have analyzed various kinds of web search biases and their impacts on the users [35, 38, 29, 5, 11, 12], still only little is known about the impact of the web search result quality on the users' decisions. In this paper, we close this gap by conducting a systematic quality assessment of Google's results for comparative questions, followed by a study on the impact of the search result quality on the users' decisions.

To this end, we developed a set of evaluation criteria grounded in previous research to assess the quality, relevance, and stance of 120 documents retrieved by Google for 30 comparative questions. The individual documents' quality scores were combined to determine the average search result quality for each comparative question (Section 3). We further conducted a follow-up user study on decision-making with eight selected comparative questions with search results of varying quality (Section 4). In the study, we asked participants to decide on either of the comparison options (e.g., buying a house vs. renting) before and after seeing the search results, and to rate their decision confidence and the influence of individual search results on their decision. After collecting 554 responses from 442 participants, we enriched the user study data with the quality, relevance, and stance scores from the previous quality assessments and tested six hypotheses:

H1 Comparative questions on subjective topics lead to less confident decisions than questions on factual topics. (Intuition: Factual comparative questions (e.g., "Does cider or beer contain more calories?") are often "better" answered by search engines than subjective comparative questions (e.g., "Should I study philosophy or psychology?") [8]. Subjective questions are also more prone to cognitive biases.)

H2 Comparative questions with low-quality results lead to less confident decisions than questions with high-quality results. (Intuition: People seek to make the best decision based on the known information [26]. Accordingly, comparative questions with low-quality search results would be harder to answer, and high-quality results would be more likely to be used in the decision-making.)

H3 The higher a search result's quality, the more likely it influences the decision-making. (Intuition: Same as for Hypothesis H2.)

H4 Users who are more confident in their decision before searching are less influenced by low-quality search results. (Intuition: Same as for Hypothesis H2.)

H5 The quality of a search result has a higher impact on the decision-making process than its relevance. (Intuition: While relevance depends on the topic at hand, our search result quality criteria are topic-independent. We hypothesize a higher impact on decision-making than relevance.)

H6 Documents that take a stance towards one of the compared options have a higher impact on the decision. (Intuition: Relevant documents can take different stances towards the compared options, favoring either option [7]. We assume that documents that do not take a stance are less helpful in making a decision.)

The significance tests indicate no significant difference between user confidence after seeing the search results for factual and subjective topics; thus, H1 cannot be confirmed. Similarly, we found no statistically significant evidence to confirm H2 that low-quality search results lead to less confident decisions than high-quality results. On the other hand, higher-quality results are still more likely to influence decisions; H3 is confirmed. We could also confirm H4 that more confident users in the decision before using web search are less influenced by low-quality search results. While our tests do not confirm H5 that the search result quality is more important than relevance in decision-making, combining both factors has a higher impact on the decision-making process than each factor alone. Finally, H6 is confirmed that search results that take a stance towards the compared options have a higher impact on the decision.

Our results entail several implications for web search. As quality and relevance are significantly correlated (high-quality results are also more often used to make decisions), it is important to consider document quality in document ranking. Since documents with a stronger stance have a higher impact on the users' decisions, the stance should also be considered a ranking signal. Moreover, our results show that high-quality documents are especially important to form decisions on high-stake subjective comparisons. Thus, search engines should potentially first identify whether a comparative question is subjective.

2 Related Work

How people decide on one or another option has been well studied by psychologists [33, 3, 26, 23]. Decisions are made either intuitively or analytically [33], and can be influenced by prior knowledge or research made ad hoc [26]. Web search engines have become a common means for collecting facts, opinions, and arguments that guide decisions, with at least three percent of web search queries being comparative questions [8]. While factual questions (e.g., "Does cider or beer contain more calories?") can often be answered analytically based on facts, subjective comparative questions (e.g., "Should I study philosophy or psychology?") may require arguments that discuss the pros and cons of possible options [8]. With an increasing trend towards direct answers [28], web search engines became tools for rather intuitive or ready-to-use solutions than analytical decision-making.

This intuitive decision-making intensifies four types of cognitive biases [5] affect the decision-making: First, users are more likely to examine results that confirm their own prior beliefs, expectations, or hypotheses known as a *confirmation bias*, and adapt their search patterns accordingly [24, 39, 38, 16, 40, 27]. Second, despite the diversity of viewpoints on the Web, search engines often favor results representing a particular point of view [35, 12]. This *viewpoint bias* affects searchers' attitudes [11, 31]. Third, users overestimate the trustworthiness of web search engines and their ranking models [34, 13, 37, 16, 22, 6]. This *trust bias* is less pronounced for more experienced users [32]. Last, the *position bias* describes the tendency to prefer web pages placed at the top of the returned search results [27, 32, 34, 17, 2, 28].

Furthermore, prior research on the impacts of search result quality is focused on *system-centered* evaluations [30, 1, 4]. So far, the impact of search result quality on the users' decision-making process has not been studied in detail. Our work contributes to a better understanding of the quality of search results for comparative questions and their impact on decision-making, by analyzing results retrieved by Google, and hence, is a first step to increasing the accountability of major search engines to their users' decisons [25, 14].

3 Assessing Search Result Quality

To assess the search result quality for comparative questions, we (1) manually selected 30 questions from the 100 topics of the Touché shared task on comparative argument retrieval [9, 10], (2) for each question, retrieved the top-4 results with Google, and (3) asked ten volunteer assessors to rate the quality of each search result following a set of predefined quality criteria.

Table 1. Quality, relevance, and stance criteria, their aspects, and answer options. The 'Score' column indicates points/multipliers for each choice and the criterion's weight in the aggregated quality. Agreement is Fleiss' κ; aspects without agreement are not used.

Aspect		Score
A	*Content*	×4
A1	Completeness ($\kappa < 0.00$)	
A2	Scope ($\kappa = 0.24$)	
	scarce	+1
	precise	+2
	appropriate	+3
	very detailed	+4
	excessive	±0
A3	Language ($\kappa = 0.32$)	
	objective / factual	+2
	entertaining	+1
	judgmental	±0
	promotional	±0
B	*Usability*	×4
B1	Media types ($\kappa < 0.00$)	
B2	Structure ($\kappa = 0.25$)	
	unstructured	±0
	roughly structured	+2
	well structured	+4
	very well structured	+6
C	*Credibility*	×2
C1	Source ($\kappa = 0.52$)	
	news portal	+4
	public institution	+4
	Q&A platform	±0
	encyclopedia	+2
	corporate website	+2
	blog	±0
C2	Author ($\kappa = 0.30$)	
	qualified author	+2
	unqualified author	+1
	generated	±0
	unknown	+1
C3	Truthfulness ($\kappa = 0.29$)	
	yes	×1
	no	×0
	partially	×0.5
	unknown	×1
C4	Verifiability ($\kappa < 0.00$)	
D	*Up-to-dateness*	×1
D1	Date ($\kappa = 0.40$)	
	outdated	+2
	up to date	+4
	timeless	+4
D2	Updates ($\kappa = 0.15$)	
	at least one update	+1
	no updates	±0
E	*Relevance*	
E1	Topical relevance ($\kappa = 0.19$)	
	not relevant	
	relevant	
	highly relevant	
F	*Stance*	
F1	Referral ($\kappa < 0.00$)	
F2	Emphasis ($\kappa < 0.00$)	
F3	Direction ($\kappa < 0.00$)	
F4	Magnitude ($\kappa = 0.51$)	
	strong	
	weak	
	no stance	
	neutral	

3.1 Data, Criteria, and Methodology

Data. Out of 100 Touché topics (each consisting of a comparative question, a description of the information need, and the relevance criterion) [9, 10], we manually selected 30 topics that contain exactly two comparison options (we discarded, e.g., superlative questions like "What are the best dish detergents?").[4] Since the assessors were native German speakers, each topic's question and relevance criterion were translated into German. For each selected question, we retrieved the top-4 search results with Google, the most popular search engine in Germany,[5] excluding videos or PDFs, and using anonymous browsing to prevent personalization. In total, we collected 120 search results in German.

Criteria. Prior quality assessment frameworks for web documents (WebQual [20], 2QCV3Q [21], AIMQ [18], Touché [10]) do not directly apply to search results for comparative questions. Therefore, we developed a set of four search result quality criteria (content, usability, credibility, and up-to-dateness), which we complemented with relevance and stance. Each criterion is further narrowed to one or more aspects (Table 1): (1) *Content* quality is determined by a document's completeness, scope, and rhetoric style. High-quality documents cover the comparative information need comprehensively and provide reasoning supported by solid evidence [36]. (2) *Usability* hinges on the document structure and readability. Offering the same content in more than one media type (e.g., text, tables,

[4] Code and data available online: https://github.com/webis-de/CLEF-24

[5] Retrieved on May 4–5, 2022. Archived results available online (see Footnote 4).

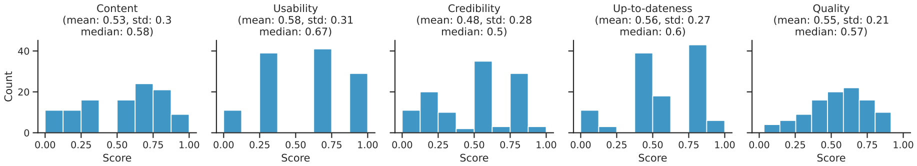

Figure 1. Quality distributions for each quality criterion and aggregated quality.

figures) improves the accessibility of a document. High-quality documents are well-structured, do not contain disruptive content (e.g., advertisements), and are easy to read [19]. (3) *Credibility* is assessed by the document's source (e.g., newspaper or government), author's qualification, truthfulness, and verifiability [36]. Thus, credible documents come from reputable sources, are written by qualified authors, only contain truthful information, and provide references to their sources. (4) The *up-to-dateness* [19] describes whether a document is up-to-date (e.g., at most 40 days old [19]) and whether it has been updated at least once. Since a publication date was not always available, we considered a document up-to-date if it was not outdated or if it indicated that it was based on recent studies.

Relevance and stance were included to support the analysis of the hypohtheses H5 and H6 (Section 1) but were not used for measuring the quality of the search results. For the topical *relevance*, the topic narratives defined in the Touché shared task [9, 10] were used and adapted to binary relevance labels (relevant or irrelevant). For the *stance*, we asked the assessors to judge which of the comparison options was mentioned in the document, which was discussed in more detail, and whether the document took a stance towards one of the options (e.g., "pro Pepsi" for the question of "Which is better, Pepsi or Cola?"). We also considered the stance magnitude, where direct recommendations indicate a strong stance and indirect supportive statements still indicate a weak stance.

Methodology. The 120 Google search results were assessed by ten volunteer assessors (German university students; 7 studied media science, 3 computer science). All assessors were provided with the codebook[4] and, for an initial pilot study, assessed the top-4 results of the same search query (topic 19, randomly selected) in random order. Table 1 shows the agreement (Fleiss' κ) for each evaluation aspect. The six aspects without agreement ($\kappa < 0.00$; i.e., A1, B1, C4, F1, F2, and F3) were removed from further analysis. In a follow-up video call, assessors discussed questions regarding the criteria and conflicting assessments. Afterwards, each assessor was given 12 search results for three queries to assess.

3.2 Evaluation

To analyze the quality of search results, we calculate quality scores for each quality criterion (content, usability, credibility, and up-to-dateness) based on their aspects, excluding four aspects without sufficient agreement. Scores for each criterion were calculated as the sum of answer points to its aspects (see the 'Score' column in Table 1). One exception is the truthfulness aspect, where the score is multiplied to account for the potential misinformation harmfulness. The resulting scores are normalized to a 0–1 range, 1 indicating a perfect score. If an aspect was not assessed (i.e., n/a was selected), we did not calculate a quality score for the corresponding criterion. Due to this filtering, 14 documents were excluded from the content quality score and seven documents from the credibility score. An aggregated score is then calculated as the weighted sum of the individual quality scores, where the weights (see the 'Score' column in Table 1) represent a media scientist's rated importance of the criteria. The weighted sum is again normalized to a 0–1 range. Documents that lack a score for at least one of the criteria are exempt from the aggregated score computation, leaving 103 documents.

The distributions of the quality scores are shown in Figure 1. Usability tends to be the "easiest" criterion to fulfill (24% of the documents achieve a perfect usability score), whereas credibility is the "hardest" (median 0.5). Quality scores of all criteria vary largely and are not normally distributed, indicating a potential selection bias due to only selecting the top-4 results. The aggregated quality scores, however, are approximately normally distributed, with an average score of 0.55 and a median of 0.57. No correlation was found between the document ranks on the result page and their quality (Kendall's $\tau = 0.07$, $p = 0.37$, $\alpha = 0.05$). Additionally, we measured topical relevance and stance. Like quality, the topical relevance is not correlated to ranks ($\tau = -0.09$, $p = 0.29$), but relevance and quality have a significant positive rank correlation ($\tau = 0.21$, $p = 0.01$).

4 User Study

4.1 Data and Methodology

Data. To characterize and select queries for the user study, we compute the average quality score and standard deviation across all documents retrieved for each query. The 10 queries where at least one result's quality could not be assessed were excluded. From the remaining 18 topics, we first removed topics that would require extensive prior knowledge. Then, we manually selected eight topics that cover a wide range of the topic-wise average quality and standard deviations within the top-4 retrieved results by Google.[6] For example, topic 12 has a high quality and low standard deviation among the retrieved documents, topic 24 has a high std. deviation and high average quality, topic 22 has a consistently average-level quality, and topics 28 and 20 have deficient overall quality.

Methodology. After selecting the topics for the user study, we archived their top-4 search results and created a questionnaire for each topic.[4] Participants were asked to imagine the situation described in the topic and then reported whether they had prior knowledge of the topic. Before seeing the search results, they decided on one of the comparison options and indicated their confidence in their decision (1–6 rating scale). Then, after they were shown the top-4 search results (screenshot, title, and source), they were asked to decide again, to report their confidence, and to indicate which of the documents shown influenced their

[6] Topics, results, and questionnaire: https://github.com/webis-de/CLEF-24

Table 2. Contingency tables of the change of the users' decisions, the decision confidence after seeing search results, and change in confidence due to seeing the results, w.r.t. topic background or avg. search result quality for the topic. Significance marked bold (Pearson's χ^2 tests, $\alpha = 0.05$, Bonferroni correction). Changes to expected frequencies in grey font. Quality threshold: 0.57, confidence threshold: 5.

Predictor	**Decision change**			**Final decision confid.**			**Decision confid. change**			
	Equal	Changed	$\sum$	Low	High	$\sum$	Decr.	Equal	Incr.	$\sum$
Background	$\chi^2(1) = 0.45,\ p = 0.502$			$\chi^2(1) = 3.29,\ p = 0.070$			$\boldsymbol{\chi^2(2) = 18.76,\ p < 0.001}$			
factual	173^{+4}	101^{-4}	274	93^{-11}	181^{+11}	274	46^{-1}	98^{-22}	121^{+23}	265
subjective	168^{-4}	112^{+4}	280	117^{+11}	163^{-11}	280	50^{+1}	146^{+22}	78^{-23}	274
$\sum$	341	213	554	210	344	554	96	244	199	539
Quality	$\chi^2(1) = 5.59,\ p = 0.018$			$\chi^2(1) = 0.03,\ p = 0.859$			$\chi^2(2) = 4.81,\ p = 0.090$			
low quality	154^{-14}	119^{+14}	273	105^{+2}	168^{-2}	273	49^{+1}	132^{+11}	87^{-12}	268
high quality	187^{+14}	94^{-14}	281	105^{-2}	176^{+2}	281	47^{-1}	112^{-11}	112^{+12}	271
$\sum$	341	213	554	210	344	554	96	244	199	539

Table 3. Contingency tables of the self-assessed agreement with five statements about the decision-making w.r.t. topic background or avg. search result quality for the topic. Significance marked bold (Pearson's χ^2 tests, $\alpha = 0.05$, Bonferroni correction). Changes to expected frequencies in grey font. Quality threshold: 0.57.

Pred.	**Conf. opinion**			**Better decis.**			**Did not help**			**Learned sth.**			**Contd. search**		
	No	Yes	$\sum$	No	Yes	$\sum$	No	Yes	$\sum$	No	Yes	$\sum$	No	Yes	$\sum$
Backg.	$\chi^2(1)=6.10$ $p=0.014$			$\boldsymbol{\chi^2(1)=18.32}$ $\boldsymbol{p<0.001}$			$\chi^2(1)=2.16$ $p=0.141$			$\chi^2(1)=3.49$ $p=0.062$			$\chi^2(1)=0.20$ $p=0.658$		
factual	199^{-13}	75^{+13}	274	171^{-23}	103^{+23}	274	227^{+7}	47^{-7}	274	127^{-11}	147^{+11}	274	208^{-3}	66^{+3}	274
subj.	229^{+13}	51^{-13}	280	222^{+23}	58^{-23}	280	217^{-7}	63^{+7}	280	153^{+11}	127^{-11}	280	218^{+3}	62^{-3}	280
$\sum$	428	126	554	393	161	554	444	110	554	280	274	554	426	128	554
Qual.	$\chi^2(1)=0.53$ $p=0.467$			$\boldsymbol{\chi^2(1)=16.71}$ $\boldsymbol{p<0.001}$			$\boldsymbol{\chi^2(1)=13.57}$ $\boldsymbol{p<0.001}$			$\boldsymbol{\chi^2(1)=25.18}$ $\boldsymbol{p<0.001}$			$\boldsymbol{\chi^2(1)=10.96}$ $\boldsymbol{p<0.001}$		
low	215^{+4}	58^{-4}	273	216^{+22}	57^{-22}	273	201^{-18}	72^{+18}	273	168^{+30}	105^{-30}	273	193^{-17}	80^{+17}	273
high	213^{-4}	68^{+4}	281	177^{-22}	104^{+22}	281	243^{+18}	38^{-18}	281	112^{-30}	169^{+30}	281	233^{+17}	48^{-17}	281
$\sum$	428	126	554	393	161	554	444	110	554	280	274	554	426	128	554

decision. We also asked the participants whether they agreed with five statements regarding the confirmation of the prior opinion, the helpfulness, the knowledge gained, and the necessity to do further research. Participants were allowed to skip reading documents they felt were irrelevant (as search engine users would normally do [15]) but reported which documents they read. The user study was conducted with 442 volunteer participants (German university students). They were given a link which randomly redirected to an online questionnaire corresponding to one of the eight topics. At the end of the questionnaire, the participants could volunteer to continue with another topic. A total of 554 submissions were received (1.25 submissions per participant on avg.; 69 per topic, min. 63, max. 80).

4.2 Evaluation

The majority of the participants (45%) did not change their decision after seeing the search results, while only 38% did (Table 2). The remaining 17% did not decide on either comparison option before or after seeing the results. Participants were already confident in their decisions before seeing the results (53% rated their confidence as 5/6 or 6/6) and further increased after seeing the results (64% rated 5/6 or 6/6). For 45% of the participants, their confidence did not change after seeing the results. For 37%, confidence increased, and for 18%, it decreased. Only 35% of the documents were reported to have influenced the users' decisions (Table 4) and only 29% of the participants reported that they could make a better decision based on the search results while 23% would continue their search (Table 3). Yet, half of the participants (49%) stated that they had learned something new about the topic and only 20% found the search results unhelpful. To verify each of the six hypotheses (Section 1), we perform significance tests (Pearson's χ^2, $\alpha = 0.05$, Bonferroni correction) on the contingency tables.

H1: Comparative questions on subjective topics lead to less confident decisions than questions on factual topics. No significant differences in the final users' confidence after seeing the search results were observed between factual and subjective topics (Table 2). However, the confidence increased significantly more for factual than subjective topics, resulting in higher final decision confidence. Participants reported they could make a better decision significantly more often for subjective than factual topics (Table 3). None of the remaining statements about the users' decision-making yielded significant differences between factual and subjective topics. Hence, we discard the hypothesis, even though factual topics lead to increased confidence more often than subjective topics.

H2: Comparative questions with low-quality results lead to less confident decisions than questions with high-quality results. To compare decision confidence w.r.t. search result quality, we first define low-quality topics as topics with an avg. document quality score below the median quality score of 0.57. High-quality topics have an avg. quality score of at least 0.57. Study participants changed their decision slightly more often for low-quality topics than for high-quality topics, and high-quality results led to a slightly increased decision confidence but none of the changes were significant (Table 2). Regarding the self-assessment of the users' decision-making, Table 3 shows that for high-quality topics, users more often reported that they could make a better decision and felt they had learned something new. For low-quality topics, users stated more often that the search results did not help and that they would continue the search. Due to the partially contradicting results for decision confidence and helpfulness in topics with different result quality, we discard the hypothesis.

H3: The higher a search result's quality, the more likely it influences the decision-making. For this hypothesis, we asked participants to report which documents had influenced their decision. Only documents that were at least partially read were considered for this analysis. We again consider documents with a quality score of less than 0.57 as low quality and documents with a quality score of at least 0.57 as high quality. Table 4 shows that low-quality documents influence decisions significantly less often than high-quality documents. A document's rank also significantly affects a document's influence on the decision. However, a position bias can be ruled out as the ranks were not correlated to quality (Section 3.2). Hence, the hypothesis can still be confirmed.

Table 4. Contingency tables of the influence of documents on the users' decisions w.r.t. result quality, relevance, quality and relevance, stance magnitude, initial confidence for low-quality documents. Significance marked bold (Pearson's χ^2 tests, $\alpha = 0.05$). Changes to expected frequencies in grey font. Quality threshold: 0.57, confidence threshold: 5.

Predictor	**Document influence**		
	No influence	Influence	$\sum$
Quality	$\boldsymbol{\chi^2(4) = 44.49, p < 0.001}$		
low quality	648 $^{+67}$	252 $^{-67}$	900
high quality	492 $^{-67}$	374 $^{+67}$	866
$\sum$	1140	626	1766
Relevance	$\boldsymbol{\chi^2(1) = 41.77, p < 0.001}$		
not relevant	649 $^{+65}$	255 $^{-65}$	904
relevant	491 $^{-65}$	371 $^{+65}$	862
$\sum$	1140	626	1766
Quality × *relevance*	$\boldsymbol{\chi^2(8) = 79.48, p < 0.001}$		
not rel., low qual.	390 $^{+65}$	113 $^{-65}$	503
not rel., high qual.	259 $^{\pm 0}$	142 $^{\pm 0}$	401
rel., low qual.	258 $^{+2}$	139 $^{-2}$	397
rel., high qual.	233 $^{-67}$	232 $^{+67}$	465
$\sum$	1140	626	1766

Predictor	**Document influence**		
	No influence	Influence	$\sum$
Stance strength	$\boldsymbol{\chi^2(2) = 26.76, p < 0.001}$		
no stance	273 $^{-6}$	237 $^{+6}$	510
weak stance	111 $^{+26}$	45 $^{-26}$	156
strong stance	75 $^{-20}$	99 $^{+20}$	174
$\sum$	459	381	840
Init. confid.	$\boldsymbol{\chi^2(1) = 4.51, p = 0.034}$		
low confidence	302 $^{-15}$	138 $^{+15}$	440
high confidence	346 $^{+15}$	114 $^{-15}$	460
$\sum$	648	252	900
Ranking position	$\boldsymbol{\chi^2(3) = 10.62, p = 0.014}$		
rank 1	286 $^{-26}$	197 $^{+26}$	483
rank 2	302 $^{\pm 0}$	166 $^{\pm 0}$	468
rank 3	281 $^{+7}$	144 $^{-7}$	425
rank 4	271 $^{+19}$	119 $^{-19}$	390
$\sum$	1140	626	1766

H4: Users who are more confident in their decision before searching are less influenced by low-quality search results. We further analyze the influence of the users' prior decision confidence by filtering low-quality documents. Of the 900 low-quality documents that were at least partially read, 252 documents influenced the decision. In Table 4, we consider low-quality documents with below-median initial confidence and high initial confidence separately. The significance test reveals that low-quality documents influenced the decision significantly more often if the initial confidence was low, confirming the hypothesis.

H5: The quality of a search result has a higher impact on the decision-making process than its relevance. Even though a document's topical relevance and quality are conceptually different, our quality assessments revealed that both are highly correlated (Section 3). Compared to the significant influence of search result quality on decision-making, Table 4 also highlights a significant influence of topical relevance with only a slightly smaller effect than for result quality. Hence, the hypothesis cannot be confirmed. However, the tests also show that combining both factors has a higher impact on decision-making than the factors alone.

H6: Documents that take a stance towards one compared option have a higher impact on the decision. Table 4 examines the impact of the stance magnitude in either direction. Documents with a strong stance (e.g., containing direct recommendations) influenced the decision significantly more often than those with a weak (e.g., indirect statements) or no stance, confirming the hypothesis.

4.3 Limitations

Our study results have several limitations. First, all participants were German university students, which might not represent the general population. Second, the study was conducted using a single search engine; thus results might not be generalizable to other search engines. Even though we used comparative questions from prior work claimed to represent real user information needs, for more robust findings, a larger study (more participants and questions) might be needed.

5 Conclusion

We evaluated the quality of web search results and the quality's impact on the decision-making for questions comparing two options. We derived guidelines to manually assess four quality criteria (content quality, usability, credibility, and up-to-dateness). The evaluation of the 120 assessed documents (top-4 results for 30 comparative questions) w.r.t. the search result quality, topical relevance, and stance showed substantial heterogeneity in the search result quality, significant correlation between relevance and quality, but no correlation of either quality or relevance with their ranks on Google's result page. Our quality assessments also highlighted that individual quality criteria on their own are not representative of a document's overall quality, motivating more systematic quality measurements for evaluation. Our criteria could serve as a starting point to design formal measures. Based on the quality assessments, we selected eight queries with varying result quality for a user study examining the search results' impact on user decisions. In the study, the participants were asked about their decisions and confidence before and after seeing the search results, which documents influenced their decision, and if they agreed with five statements about the decision-making process.

Our results showed that the quality of search results has a significant impact on being used in the decision-making process (H3) but not on the confidence of user decisions (H2). Quality can thus be considered an important factor in the search result ranking for comparative questions. As documents with a stronger stance also have a higher impact on the users' decisions (H6), we suggest that the stance magnitude should also be considered for ranking. Even though no significant difference was found between the confidence after seeing the search results of factual and subjective questions (H1), the topic background still significantly influences the change in decision confidence. Users gained confidence in their decisions significantly more often for factual than for subjective questions. Improving search result quality, especially for subjective questions, could thus help users to make more confident decisions. The user study also showed that users with initially high decision confidence are less likely to be influenced by low-quality results (H4). Last, we observed a similarly pronounced impact of both quality and relevance on the decision-making process (H5) and that combining the two factors has a higher impact on decision-making than any factor alone. Because current evaluation merely considers relevance or quality on its own [10], combining both factors in future evaluations of comparative queries is worthwhile.

Acknowledgments. This work was partially supported by the DFG (German Research Foundation) through the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (project 376430233) as part of the priority program "RATIO: Robust

Argumentation Machines" (SPP 1999) and by the Stiftung für Innovation in der Hochschullehre under the "freiraum 2022" call (FRFMM-58/2022).

Bibliography

[1] Abualsaud, M.: The effect of queries and search result quality on the rate of query abandonment in interactive information retrieval. In: CHIIR 2020

[2] Abualsaud, M., Smucker, M.D.: Exposure and order effects of misinformation on health search decisions. In: ROME@SIGIR 2019

[3] Ajzen, I.: The Social Psychology of Decision Making (1996)

[4] Alanazi, A.O., Sanderson, M., Bao, Z., Kim, J.: The impact of ad quality and position on mobile serps. In: CHIIR 2020

[5] Azzopardi, L.: Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In: CHIIR 2021

[6] Bink, M., Schwarz, S., Draws, T., Elsweiler, D.: Investigating the influence of featured snippets on user attitudes. In: CHIIR 2023

[7] Bondarenko, A., Ajjour, Y., Dittmar, V., Homann, N., Braslavski, P., Hagen, M.: Towards understanding and answering comparative questions. In: WSDM 2022

[8] Bondarenko, A., Braslavski, P., Völske, M., Aly, R., Fröbe, M., Panchenko, A., Biemann, C., Stein, B., Hagen, M.: Comparative web search questions. In: WSDM 2020

[9] Bondarenko, A., Fröbe, M., Beloucif, M., Gienapp, L., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2020: Argument retrieval. In: CLEF 2020

[10] Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., Hagen, M.: Overview of Touché 2021: Argument retrieval. In: CLEF 2021

[11] Draws, T., Tintarev, N., Gadiraju, U., Bozzon, A., Timmermans, B.: This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In: SIGIR 2021

[12] Gezici, G., Lipani, A., Saygin, Y., Yilmaz, E.: Evaluation metrics for measuring bias in search engine results. Inf. Retr. J. **24**(2) (2021)

[13] Grimmelmann, J.: The google dilemma. NYL Sch. L. Rev. **53**(2) (2008)

[14] Gunning, D., Aha, D.W.: DARPA's explainable artificial intelligence (XAI) program. AI Mag. **40**(2) (2019)

[15] Joachims, T.: Optimizing search engines using clickthrough data. In: KDD 2002

[16] Knobloch-Westerwick, S., Johnson, B.K., Westerwick, A.: Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. J. Comput. Mediat. Commun. **20**(2) (2015)

[17] Lau, A.Y.S., Coiera, E.W.: Do people experience cognitive biases while searching for information? J. Am. Medical Informatics Assoc. **14**(5) (2007)

[18] Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y.: AIMQ: A methodology for information quality assessment. Inf. Manag. **40**(2) (2002)

[19] Lewandowski, D., Höchstötter, N.: Web Searching: A Quality Measurement Perspective (2008)

[20] Loiacono, E.T., Watson, R.T., Goodhue, D.L.: WebQual: A measure of website quality. In: AMA 2002

[21] Mich, L., Franch, M., Cilione, G.: The 2QCV3Q quality model for the analysis of web site requirements. J. Web Eng. **2**(1) (2003)

[22] Narayanan, D., De Cremer, D.: 'google told me so!' on the bent testimony of search engine algorithms. Philos. Technol. **35**(2) (2022)

[23] Newell, B.R., Lagnado, D.A., Shanks, D.R.: Straight Choices: The Psychology of Decision Making (2022)

[24] Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology **2**(2) (1998)

[25] Nyholm, S.: Attributing agency to automated systems: Reflections on human-robot collaborations and responsibility-loci. Sci. Eng. Ethics **24**(4) (2018)

[26] Peterson, M.: An Introduction to Decision Theory (2017)

[27] Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.L.A.: The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In: ICTIR 2017

[28] Potthast, M., Hagen, M., Stein, B.: The dilemma of the direct answer. SIGIR Forum **54**(1) (2020)

[29] Puschmann, C.: Beyond the bubble: Assessing the diversity of political search results. Digital Journalism **7**(6) (2019)

[30] Raiber, F., Kurland, O.: Using document-quality measures to predict web-search effectiveness. In: ECIR 2013

[31] Rieder, B., Sire, G.: Conflicts of interest and incentives to bias: A microeconomic critique of Google's tangled position on the web. New Media Soc. **16**(2) (2014)

[32] Schultheiß, S., Lewandowski, D.: Misplaced trust? the relationship between trust, ability to identify commercially influenced results and search engine preference. J. Inf. Sci. **49**(3) (2023)

[33] Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science **185**(4157) (1974)

[34] Unkel, J., Haas, A.: The effects of credibility cues on the selection of search engine results. J. Assoc. Inf. Sci. Technol. **68**(8) (2017)

[35] Vaughan, L., Thelwall, M.: Search engine coverage bias: Evidence and possible causes. Inf. Process. Manag. **40**(4) (2004)

[36] Wang, Y., Sarkar, S., Shah, C.: Juggling with information sources, task type, and information quality. In: CHIIR 2018

[37] Westerwick, A.: Effects of sponsorship, web site design, and google ranking on the credibility of online information. J. Comput. Mediat. Commun. **18**(2) (2013)

[38] White, R.: Beliefs and biases in web search. In: SIGIR 2013

[39] White, R.W.: Belief dynamics in web search. J. Assoc. Inf. Sci. Technol. **65**(11) (2014)

[40] Yom-Tov, E., Dumais, S., Guo, Q.: Promoting civil discourse through search engine diversity. Social Science Computer Review **32**(2) (2014)