Confirmation Bias: Roles of Search Engines and Search Contexts

Completed Research Paper

Varol Onur Kayhan
University of South Florida St. Petersburg
140 7th Ave. South, St. Petersburg, FL, 33701, USA
vkayhan@usfsp.edu

Abstract

Prior work shows that confirmation bias, defined as the tendency to seek confirming evidence, is prevalent on the Web as well. While this has been attributed to individuals' psychological needs or cognitive limitations, the roles of search engines and search contexts have largely been neglected. The goals of this study are to examine how search contexts may change the composition of search results, and how – if at all – search engines may contribute to confirmation bias. Results of two studies show that search engines may exacerbate confirmation bias by generating results that consist only of confirming evidence for search contexts where disconfirming evidence is identified using different terms or phrases. This induces individuals to make biased decisions. Findings of this study deepen our understanding of the ways in which confirmation bias unfolds on the Web when individuals use search engines.

Keywords: Confirmation bias, search engine, confirming evidence, experimental design

Introduction

The Internet is providing access to everything from entertainment to scientific knowledge. It has become so prominent in our lives that we sometimes have too much faith in what we read online regardless of who the author is or where it is published (Eysenbach and Köhler 2002; Kakol et al. 2013). Compounding this issue are the search strategies we employ on the Web that are, more often than not, geared toward seeking information that confirms our existing beliefs (Feufel and Stahl 2012; Huang et al. 2012; White 2013) – a phenomenon referred to as confirmation bias (Nickerson 1998).

Individuals' tendency toward confirmation bias has long been established in the offline world (see Klayman 1995; Nickerson 1998; Schulz-Hardt et al. 2000). Studies that examine online behaviors report that confirmation bias is also prevalent in the online world when individuals search for information using search engines (see Feufel and Stahl 2012; Huang et al. 2012; Keselman et al. 2008; Lau and Coiera 2007c; White 2013). While these studies mostly cite our cognitive limitations or psychological needs as the major reasons behind confirmation bias, we still have little to no insight into how confirmation bias unfolds on the Web, and whether search contexts have any role in individuals' engagement in confirmation bias. Therefore, the goal of this study is to examine how search contexts influence the composition of search results, and whether they induce search engines to exacerbate confirmation bias as individuals search for information on the Web about the validity of a statement.

This is important, because if we can uncover how – if at all – search engines and search contexts play a role in confirmation bias, we can not only shed more light on the process of making decisions, but also devise intervention techniques. Such an endeavor can also help us avoid costly consequences: in the context of healthcare alone, confirmation bias has been linked to incorrect diagnoses, unneeded tests, and unnecessary treatments – all of which contribute to increasing costs (Broom 2005; Kale et al. 2011; Markoff 2008; Meisel and Pines 2012; Rabin 2012; Wagner et al. 2001).

Background

In an effort to understand how confirmation bias transpires on the Web, we turn to the process of information search. An examination of the existing frameworks shows that web-based information search occurs in three high-level steps: 1) submission of a search query to a search engine; 2) analysis of results; and 3) the use of the results to make a decision, form a judgment, or perform a task (see Hodkinson and Kiel 2003; Kulviwat et al. 2004; Lueg et al. 2003; Marchionini and White 2007).

This process is relatively flawless for navigational searches, where the intent is to reach a desired website. For example, a user submits a set of keywords to a search engine – such as "American Airlines" – and completes the search process by clicking on the relevant link in the results ("http://www.aa.com"). However, the flow of events may have unexpected consequences for informational queries, where the intent is sense-making, because individuals may fall prey to different types of cognitive biases during the search process (see Ariely 2008; Tversky and Kahneman 1974). One of these biases is confirmation bias, defined as the tendency to seek information that validates the topic being searched (Klayman 1995; Nickerson 1998).

Confirmation bias generally operates at the subconscious level, which makes it difficult for people to realize or even prevent it (Gilovich 1991). Even though confirmation bias may be helpful in certain contexts – where it enables individuals to make faster and easier decisions through heuristics (Gigerenzer and Todd 1999) – it causes more harm than good in other contexts since it leads to unwarranted confirmation (see Nickerson 1998). Extant literature on confirmation bias paints a fragmented picture with no consensus on its types, underlying reasons, or consequences (Klayman 1995); however, it is generally agreed that confirmation bias manifests itself in two specific ways: selective search and biased interpretation (Park et al. 2013). While selective search induces individuals to specifically search for confirming information, biased interpretation makes them discredit any disconfirming information and rely heavily on confirming information. The underlying reasons behind these two biases have been attributed to three factors: 1) individuals' need for self-enhancement – i.e., their desire to hold a positive view of themselves (Taylor and Brown 1988); 2) individuals' need for consistency – i.e., their desire to avoid cognitive dissonance (Festinger 1957; Swann et al. 1987); and 3) individuals' need to minimize cognitive effort – i.e., their desire to use minimum cognitive resources during search tasks (Nickerson 1998).

Regardless of its types or underlying reasons, it has been reported that confirmation bias usually generates biased decisions during online information search (Feufel and Stahl 2012; Kayhan 2013; Lau and Coiera 2007a; Lau and Coiera 2009). Search engines have also been shown to contribute to confirmation bias. For example, individuals' selection of links are influenced by the ranking of links, attractiveness of captions used for links, and the domains of links provided in search results (Craswell et al. 2008; Ieong et al. 2012; Yue et al. 2010). Also, search engines have been shown to favor links that provide a "yes" answer for the question being searched – when the search is answered by a "yes" or "no" (White 2013). Therefore, it is difficult to make unbiased decisions while searching for information on the Web given our tendency toward confirmation bias and the other confounds introduced by search engines.

However, this contradicts not only the conventional wisdom, but also extant work since the Web continues to help us make better decisions (Hostler et al. 2005). Then, the question, and what has not been articulated in extant work, is: given existing search technologies and our search strategies we employ on the Web, when are we more likely to obtain a biased set of search results, and ultimately make biased decisions? To examine this, we offer a revised process of information search – discussed next.

A Revised Process of Web-based Information Search

Combining extant work on confirmation bias and the process of information search, we conceptualize a revised process of Web-based information search as shown in Figure 1. Accordingly, we suggest that an information seeker's familiarity with the search task as well as the way a search task is framed can influence his/her search queries submitted to search engines. These search queries can generate biased results contingent on the search context, and may provide the information seeker with only confirming evidence depending on the ways in which disconfirming evidence can be identified. The results, biased or

not, are further susceptible to individuals' selective search and interpretation bias while being used to make a decision or form a judgment.

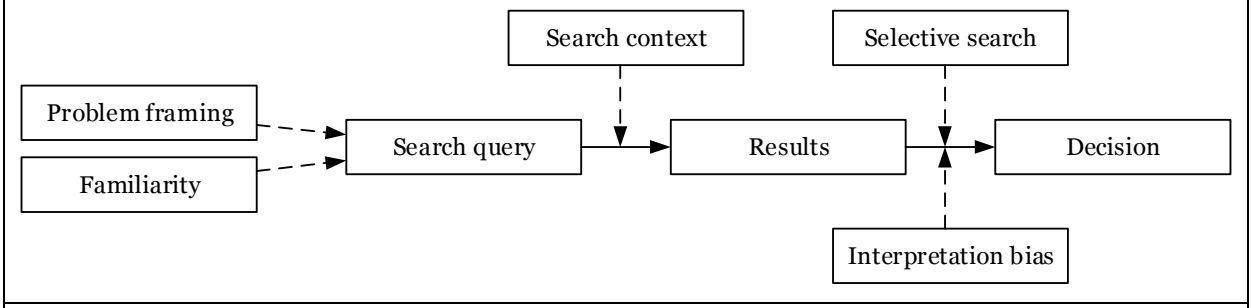

Figure 1. Revised Framework of Web-based Information Search

In an effort to shed more light on this process, we conducted two studies. Even though our major focus is on understanding the role of search context, we investigate individuals' selective search and interpretation bias as well. The first of these studies is discussed next.

Study 1

The main motivation behind this study is to understand how problem framing and familiarity influence search queries for a typical search context, and therefore, examine how information search unfolds on the Web with possible sources of bias. To this end, we provided participants with a statement about a relationship between coffee consumption and hypertension, and asked them to test the validity of this statement in a controlled online environment – described in detail below. In doing so, we manipulated the framing of this statement: half of the participants were asked to test the validity of the statement "there is a link between coffee consumption and hypertension," and the other half was asked to test the validity of the negatively framed statement: "there is no link between coffee consumption and hypertension." This manipulation helped us compare the composition of search results as a result of the search queries being used, and thus understand the use of search results during decision making.

Experimental Setup

For this study, we focused on the controversial relationship between coffee consumption and hypertension: while certain studies support this relationship, others refute it (see Nurminen et al. 1999). Using the abstracts of several peer-reviewed studies, we created four bogus abstracts that reported that there was a link between coffee consumption and hypertension, and another four that reported that there was no link. The creation of these abstracts was an arduous process (interested readers can refer to Appendix A to find more information about the development as well as the authenticity and believability of these abstracts). Then, we created a separate web page for each abstract: each web page included a title, an abstract, bogus author details, and bogus journal details (please see Figure B1 of Appendix B for an example web page).

In addition to these eight web pages, we created 30 more web pages to act as noise – these pages did not concern the link between coffee consumption and hypertension, but concerned other issues about either coffee consumption or blood pressure (but not both). The web pages created for this study were hosted on a web server and included several PHP scripts to track participants' online activities.

In order to enable participants to conduct keyword searches, we used Google's Custom Search Engine service (http://www.google.com/cse/). This service helps create a custom search engine using Google's proprietary algorithm and index a specific set of web pages so that users can conduct searches only within those pages. The custom search engine created using this service indexed only the 38 pages discussed earlier.

Participants and Procedure

A total of 40 participants were recruited from Amazon Mechanical Turk (AMT). Our decision to use AMT was motivated, in part, by the reliability of findings obtained from AMT's participant pool (see Steelman et al. 2014). All participants were anonymous and all study procedures, including the experimental task, were in line with the Institutional Review Board (IRB) rules and regulations.

In order to recruit participants, we created a project in AMT and provided the link of the web server that hosted our web pages. When participants clicked on this link, a script running on the server assigned each participant to one of the two sets of instructions in a round-robin fashion. The difference between the instructions was framing: one set of instructions asked participants to test the validity of the statement "there is a link between coffee consumption and hypertension," and the other set asked them to test the negatively framed statement ("there is no link between coffee consumption and hypertension"). An example set of instructions used in the experiment is provided in Figure B2 of Appendix B.

The instructions page also provided the link of the search engine to conduct keyword searches (see Figure B3 of Appendix B for the search interface). Note that, the instructions did not indicate the number or nature of the web pages indexed by this search engine. The search engine's search results were the same as a typical Google search – except there were no ads or sponsored links (see Figure B4 of Appendix B for a sample search result). Upon examining the results, participants went back to the instructions page to indicate whether the statement was valid, invalid, or neither (i.e., "other").

It is important to note that we measured participants' initial familiarity about the link between coffee consumption and blood pressure using a 7-point Likert scale before they were shown the instructions of the experiment. Overall, participants were not very familiar with this relationship (mean familiarity score was 2.95 with a standard deviation of 2.03). Further, there was no statistical difference between the familiarity scores of the two groups (means were 2.85 versus 3.05, p=0.76).

Results

Composition of Search Results

Overall, participants conducted a total of 47 searches (1.2 searches per participant). Twenty-five of these searches (53%) were conducted by participants who received the negatively framed statement. Recall that one of the motivations of this study was to identify the differences – if there were any – between the search queries written by participants across the two groups. To test this, we captured the result set generated for each query submitted by participants. Within each result set we identified the number of pages that supported the link between coffee consumption and hypertension (referred to as "Coffee & Hypertension" hereafter), and the number of pages that did not support this link (referred to as "Coffee & No hypertension" hereafter).

We conducted a multivariate analysis of covariance (MANCOVA) test, where the dependent variables were the number of pages that entertained "Coffee & Hypertension", and the number of pages that entertained "Coffee & No hypertension", the independent variable was framing, and the covariate was participant's familiarity with the scenario.

The results suggested that neither framing ($F(2,43)=0.12$, $p=0.89$), nor familiarity ($F(2,43)=0.46$, $p=0.63$) had any effect on the composition of search results. Univariate analyses suggested that search queries that tested the positively framed statement, on average, returned more pages that entertained "Coffee & Hypertension" in search results than those that tested the negatively framed statement, but the difference was not statistically significant (means were 2.46 versus 2.07 respectively, $p=0.34$). Similarly, search queries that tested the positively framed statement, on average, returned more pages that entertained "Coffee & No hypertension" in search results than those who tested the negatively framed statement, but the difference, again, was not statistically significant (means were 3.27 versus 3.04 respectively, $p=0.52$)

Pages Downloaded by Participants

Even if the search engine did not generate different results for the two groups of participants, we examined the types of pages downloaded by participants from search results to see whether they selectively downloaded pages that confirmed the statement provided to them – the selectivity assertion of confirmation bias.

To this end, we captured the pages downloaded by each participant as well as the approximate time the participant spent on each document. To capture the time spent on a page, we obtained the time difference between two consecutive downloads of each participant. We assumed that a participant read a downloaded page if he/she spent more than two seconds on the page before downloading the next page.

Similar to our initial analysis, we conducted a MANCOVA, where the dependent variables were the number of downloaded pages that entertained "Coffee & Hypertension", and the number of downloaded pages that entertained "Coffee & No hypertension", the independent variable was framing, and the covariate was participant's familiarity with the scenario. For this analysis, our level of analysis was a single participant – 38 participants were included into the analysis (out of a total of 40), since two participants (one from each group) did not download any pages from their results. No elimination was performed as a result of the two-second rule. (Using a five-second rule did not lead to any eliminations either. On average, a participant spent 19 seconds on each page.)

The results suggested that neither framing ($F(2,34)=1.63$, $p=0.21$), nor familiarity ($F(2,34)=0.75$, $p=0.48$) had any effect on the types of pages downloaded by participants. Those who tested the positively framed statement, on average, downloaded more pages that entertained "Coffee & Hypertension" than those who tested the negatively framed statement, but the difference was not statistically significant (means were 1.84 versus 1.37 respectively, $p=0.28$). Further, those who tested the positively framed statement, on average, downloaded fewer documents that entertained "Coffee & No hypertension" than those who tested the negatively framed statement, but the difference was, again, not statistically significant (means were 2.05 versus 2.37 respectively, $p=0.42$).

Decisions Made by Participants

Finally, we examined the decisions made by participants to see whether participants engaged in biased interpretation of results – another assertion of confirmation bias. From a total of 40 participants, we eliminated two from the analysis (one from each group), since they provided answers without downloading any pages from their search results. The answers of the remaining 38 participants (19 in each group) were as follows.

Of the 19 participants who tested the positively framed statement, two (or 10%) selected the "other" option, indicating that they could not make a decision due to conflicting evidence; seven (or 37%) indicated that the statement was valid; and the remaining 10 (53%) indicated it was not valid. Among the other 19 participants who tested the negatively framed statement, two (or 10%) selected the other option (due to conflicting evidence); 11 (or 58%) indicated that the statement was valid; and the remaining 6 (or 32%) indicated that it was not valid. The breakdown of the answers are provided in Table 1.

Answers	Positive framing (Coffee & Hypertension)	Negative framing (Coffee & No hypertension)
Valid	7 (37%)	11 (58%)
Not valid	10 (53%)	6 (32%)
Other	2 (10%)	2 (10%)
Total	19 (100%)	19 (100%)

Table 1. Breakdown of answers

In an effort to further identify the determinants of decisions, we conducted a logistic regression using participants' answers as the dependent variable. Results (shown in Table 2) showed that the number of downloaded documents was influential in decisions: the number of downloaded pages that entertained "Coffee & Hypertension" increased participants' likelihood of indicating that the link between coffee consumption and hypertension was valid, while the number of downloaded pages that entertained "Coffee & No hypertension" decreased this likelihood. Control variables, framing, the nature of the first

downloaded page (a binary variable: 1 = Coffee & Hypertension; 0 = Coffee & No hypertension), and initial familiarity, did not have any effect on decisions.

Logistic regression:
Dependent variable: Agreement with 'Coffee & Hypertension'
(χ^2= 11.58, p=0.04)

Variables	**B**	**p-value**	**Exp(β)**
No. of downloads for 'Coffee & Hypertension'	0.95	0.05	2.58
No. of downloads for 'Coffee & No hypertension'	-0.68	0.05	0.51
Framing	-0.40	0.65	0.67
First download is 'Coffee & Hypertension'	1.11	0.31	3.02
Initial familiarity	0.39	0.10	1.47

Table 2. Results of Logistic Regression

Note that we re-conducted the same logistic regression to check whether submitting more than one query influenced participants' decisions. To this end, we added another binary variable into the existing model: whether a participant conducted more than one search or not. The results showed that neither the model (χ^2= 11.75, p=0.07), nor the new variable (p=0.67) was significant. The effects of the remaining variables were nearly identical to the ones shown in Table 2.

Discussion

In this study, our main goal is to examine how information search unfolds on the Web by examining search strategies employed by information seekers and determining the prevalence of confirmation bias while testing the validity of a statement. According to the process of Web-based search, individuals' tendency to seek confirming evidence should have biased their search queries through framing or familiarity, and therefore, led search engines to generate biased results. Therefore, participants who received the positively framed statement should have had more links that supported the link between coffee and hypertension in their search results, while participants who received the negatively framed statement should have had more links that entertained "Coffee & No hypertension" in their search results. However, we were not able to provide much support for this argument. The compositions of search results were not statistically different between the two groups.

According to the selective-search assertion of confirmation bias, we also expected participants to download more pages that confirmed the statement provided to them in the instructions. Therefore, participants who received the positively framed statement should have downloaded more pages that supported the link between coffee and hypertension, while participants who received the negatively framed statement should have downloaded more links that failed to support this link (i.e., "Coffee & No hypertension"). However, we were not able to validate this either since there were no statistical differences between the types of downloaded pages across the two groups.

Finally, we examined whether participants engaged in biased interpretation across the two groups. According to the biased interpretation hypothesis, participants who received the positively framed statement should have exhibited more agreement with the link between coffee and hypertension, while participants who received the negatively framed statement should have exhibited more agreement with the statement about coffee and no hypertension. However, we were not able to see any discernable patterns, because participants' answers – after adjusting for framing – were not statistically different: 53% of participants in the positively framed group and 58% of participants in the negatively framed group indicated that there was no link between coffee consumption and hypertension (t=-0.32, p=0.75).

Study 2

The motivation behind Study 2 was to examine how the search context would influence search results and decisions. Therefore, we manipulated the search context by changing the nature of pages that disconfirm the link between coffee consumption and hypertension. Since the opposite end of hypertension is low blood pressure (see Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html), the link between coffee consumption and hypertension can be refuted by a link between coffee consumption and low blood

pressure – a phenomenon observed in peer-reviewed academic studies (see Appendix A for more details). Therefore, we repeated Study 1 by replacing the four abstracts that entertained "Coffee & No hypertension" with another set of four abstracts that reported a link between coffee consumption and low blood pressure. As a result, four abstracts supported a link between coffee consumption and hypertension, and another four supported a link between coffee consumption and low blood pressure. Once again we manipulated framing by asking half of the participants to test the validity of the statement "there is a link between coffee consumption and hypertension," and the other half to test "there is a link between coffee consumption and low blood pressure."

Experimental Setup

We used the same experimental setup described in Study 1 with only one exception: four abstracts that supported the link "Coffee & No hypertension" in Study 1 were replaced with four new abstracts that supported a link between coffee consumption and low blood pressure (referred to as "Coffee & Low blood pressure" hereafter). Therefore, the composition of the abstracts in the web server were as follows: four abstracts supported "Coffee & Hypertension", four abstracts supported "Coffee & Low blood pressure", and there were 30 more abstracts as noise. Similar to Study 1, we created a Google custom search engine to index these 38 pages.

Participants and Procedure

Another 40 participants were recruited from AMT for this study using the same procedure employed in Study 1. Half of these participants were assigned to the "Coffee & Hypertension" group, while the other half to the "Coffee & Low blood pressure" group in a round-robin fashion.

As in Study 1, we measured participants' initial familiarity with the link between coffee consumption and blood pressure before showing them the instructions of the experiment. Mean familiarity score was 3.15 with a standard deviation of 1.98. No statistical differences were observed between the two groups (with means of 3.00 versus 3.29, $p=0.66$), or between participants in this study and Study 1 ($F(1,78)=0.20$, $p=0.66$).

Results

Composition of Search Results

In this study, participants conducted a total of 45 searches (compared to 47 searches in Study 1). Twenty-one of these (47%) were conducted by participants who tested "Coffee & Hypertension". The MANCOVA test revealed that framing had a statistically significant effect on the composition of search results ($F(2,41)=27.75$, $p<0.001$), whereas initial familiarity did not ($F(2,41)=0.39$, $p=0.68$). Univariate analysis suggested that search queries that tested "Coffee & Hypertension" included more pages that supported "Coffee & Hypertension" in search results than those that tested "Coffee & Low blood pressure" with statistical significance (means were 2.80 versus 0.93 respectively, $p<0.001$). Similarly, queries that tested "Coffee & Low blood pressure" included more pages that supported "Coffee & Low blood pressure" in search results than those that tested "Coffee & Hypertension" with statistical significance (means were 2.88 versus 0.86 respectively, $p<0.001$).

Pages Downloaded by Participants

Next, we examined the nature of the pages downloaded by participants. Of the 40 participants, 32 were included to this analysis, because eight participants did not download any pages from their search results. No other eliminations were made based on the two-second rule. (Using a five-second rule did not lead to any eliminations either. On average, a participant spent 24 seconds on each page.)

Results of MANCOVA showed that framing had a statistical effect on the types of downloaded pages ($F(2,28)=12.25$, $p<0.001$), but familiarity did not ($F(2,28)=0.12$, $p=0.89$). Univariate analysis showed that those who tested "Coffee & Hypertension" downloaded more documents that supported "Coffee & Hypertension" than those who tested "Coffee & Low blood pressure" with statistical significance (means were 2.17 versus 0.78 respectively, $p=0.005$). Further, those who tested "Coffee & Low blood pressure"

downloaded more pages that supported "Coffee & Low blood pressure" than those who tested "Coffee & Hypertension" with statistical significance (means were 2.00 versus 0.67 respectively, p=0.003).

These results were expected because of the composition of search results: as seen in Table 3, the number of downloaded pages were highly correlated to the number of pages in search results for both "Coffee & Hypertension" and "Coffee & Low blood pressure." Therefore, participants' downloads were a true reflection of the composition of search results obtained from the search engine.

		1	2	3	4
1	Pages on "Coffee & Hypertension" (downloaded)	1			
2	Pages on "Coffee & Low blood pressure" (downloaded)	-0.06	1		
3	Pages on "Coffee & Hypertension" (in search results)	0.75*	-0.06	1	
4	Pages on "Coffee & Low blood pressure" (in search results)	-0.30	0.86*	-0.26	1

* p< 0.01

Table 3. Correlations between search results and downloads

Decisions Made by Participants

Finally, we examined the answers of 32 participants (out of 40), since eight did not download any documents from their search results. Of these participants, 18 were in the "Coffee & Hypertension" group, and the other 14 were in the "Coffee & Low blood pressure" group. Among the 18 who tested "Coffee & Hypertension", one (or 6%) selected the other option (indicating that there was conflicting evidence); 13 (or 72%) indicated that the statement was valid; and the remaining four (or 22%) indicated it was not valid. Among the 14 who tested "Coffee & Low blood pressure", three (or 21%) selected the other option (stating that there was conflicting evidence); 10 (or 71%) indicated the statement was valid; and one (or 7%) indicated it was not valid. The breakdown of answers are provided in Table 4.

Answers	**Coffee & Hypertension**	**Coffee & Low blood pressure**
Valid	13 (72%)	10 (71%)
Not valid	4 (22%)	1 (7%)
Other	1 (6%)	3 (21%)
Total	18 (100%)	14 (100%)

Table 4. Breakdown of answers

A cross-tabulation of participants' answers to the types of pages they downloaded provides insights into these answers (see Table 5). Among 18 participants who tested "Coffee & Hypertension", 13 (or 72%) downloaded only those pages that entertained "Coffee & Hypertension," and therefore, 12 of these (67%) indicated that the statement was valid. Only one out of 18 (6%) downloaded only those pages that entertained "Coffee & Low blood pressure," and consequently indicated that the statement was not valid. The remaining four participants (or 22%) downloaded both types of pages, and three of these (17%) indicated that the statement was invalid, while the other (6%) indicated that it was valid. See Table 5 for details.

	Statement tested: "Coffee & Hypertension" (N=18)		**Statement tested: "Coffee & Low blood pressure" (N=14)**	
Type of downloaded pages	**Valid**	**Not valid**	**Valid**	**Not valid**
Only "Coffee & Hypertension"	67% (12 of 18)	6% (1 of 18)	-	-
Only "Coffee & Low blood pressure	-	6% (1 of 18)	57% (8 of 14)	7% (1 of 14)
Both	6% (1 of 18)	17% (3 of 18)	14% (2 of 14)	21% (3 of 14)

Table 5. Breakdown of answers

Among the other 14 participants who tested "Coffee & Low blood pressure", nine (64%) downloaded only those documents that entertained "Coffee & Low blood pressure,", and eight of these (57%) indicated that the statement was valid. The remaining five participants (36%) downloaded both types of pages, and decisions were split: two (14%) indicated that the statement was valid, and three (21%) indicated that it was not valid. No participant downloaded only those pages that entertained "Coffee & Hypertension" in this group. See Table 5 for details.

Discussion

In this study, our goal is to see whether the search context can make individuals more susceptible to confirmation bias. We expected that when disconfirming evidence was identified using a different word or phrase, the search engine would generate a result set consisting mostly of confirming evidence, which, in turn, would lead to downloading confirming evidence only, and thus, making biased decisions. Our findings provided support for this expectation. Participants who received the relationship between coffee and hypertension had more links in their search results that confirmed this relationship. This led them to download confirmatory evidence, and thus agree with the validity of this relationship. Similarly, participants who received the relationship between coffee and low blood pressure had more links in their search results that supported the validity of this relationship, which in turn led them to download confirmatory evidence for this relationship. Therefore, these participants agreed that there was a relationship between coffee and low blood pressure.

General Discussion

Search Results

The tendency to engage in confirmation bias has long been established in extant work: when individuals test the validity of a statement or hypothesis, they subconsciously gravitate toward evidence that confirms this statement or hypothesis (Gilovich 1991; Klayman 1995; Nickerson 1998). Even though extant work that examines online search behaviors supports this tendency (see Feufel and Stahl 2012; Huang et al. 2012; Kayhan 2013; Lau and Coiera 2007a; Markoff 2008; White 2013), it provides little to no insight into the ways in which search contexts contribute to confirmation bias. We address this gap in literature using two studies.

Our findings suggest that the nature of a search context can create the sufficient conditions for a search engine to generate biased results, which further lead to biased decisions – discussed in detail in the next section. In an effort to formalize the ways in which biased results are generated, and thereby, understand how the process of confirmation bias may unfold on the Web, we need to differentiate the syntactical differences between statements using the concept of opposition.

In everyday language, creating oppositions is simple and intuitive. For example, one can create the opposition of the statement "I am drinking" using "I am *not* drinking," or the statement "it is good" using "it is bad." The theory of negation, which dates back to Aristotle, suggests that natural languages lend themselves to four types of opposition (Horn 1989): (1) *correlation* (such as double vs. half); (2) *contrariety* (such as good vs. bad); (3) *privation* (such as blind vs. sighted); and (4) *contradiction* (such as he sits vs. he does not sit). For the purposes of this study, the semantic differences between these oppositions are immaterial – interested readers can refer to the work of Horn (1989) for more information. More important for this study is the syntax of these oppositions. From a syntactic perspective, the types of oppositions can be categorized into two groups: *similar-term opposition*, (i.e., contradiction), where opposition is created using the 'not' operator; and *dissimilar-term opposition*, (i.e., correlation, contrariety, and privation), where opposition is created using a different word or phrase.

The distinction between these two types of oppositions are important in the online world, because they lead to different types of disconfirming evidence about a given statement or hypothesis. Consider a statement where two concepts, such as P and Q, have a relationship in the form of "if P then Q." For a similar-term opposition scenario, disconfirming evidence takes the form "if P then not Q" (as in "there is no link between coffee consumption and hypertension"), whereas for a dissimilar-term opposition scenario, it takes the form "If P then R" where R is the opposite of Q (as in "there is a link between coffee

consumption and low blood pressure"). Both types of disconfirming evidence invalidate the statement "if P then Q" due to their conflicting nature (Hempel 1945; Nickerson 1998).

The keyword-matching algorithm used in most search engines handle these two types of disconfirming evidence differently. For example, if individuals are asked to test the validity of the statement "if P then Q," they are likely to use both P and Q in their search queries to narrow the search space and generate the most relevant result set – doing otherwise may generate irrelevant results, and require more time, effort, and thus cognitive resources to sift through the results. After receiving this query (with P and Q as keywords), the keyword-matching algorithm used in a search engine is likely to return all pages that include both P and Q. If this is a similar-term opposition scenario, the search engine may return disconfirming pages as well – if there are any – since these pages also include the two keywords – as shown in Study 1. However, if the scenario is a dissimilar-term opposition scenario, then the search engine is likely to exclude the available disconfirming pages, because the term Q is nowhere to be found in them – as shown in Study 2.

In light of this, our studies posit that if queries are written toward statements or hypotheses where disconfirming evidence can be identified using dissimilar-term oppositions, one should expect more, and perhaps only, confirming evidence in search results, whereas if queries are written toward statements or hypotheses for which disconfirming evidence is identified using similar-term oppositions, search engines should return a combination of both confirming and disconfirming evidence in search results (contingent upon the availability of disconfirming evidence). This is one of the major findings of this manuscript and an important insight into how confirmation bias may unfold on the Web.

Decisions Made by Participants

Based on the results of our studies, we also posit that dissimilar-term opposition statements or hypotheses are more likely to induce biased decisions. The process by which this happens is as follows. Search queries used for a dissimilar-term opposition statement or hypothesis generate search results that include only confirming evidence about that statement or hypothesis, and individuals downloading from this result set usually have no choice but to agree with the validity of this statement or hypothesis. Therefore, even if we observe biased decisions, their underlying reason is different than other psychological reasons, such as selective search or biased interpretation commonly cited in extant work (see Park et al. 2013).

This implies that a similar-term opposition statement or hypothesis is more likely to generate unbiased decisions, since queries used for this type of statement or hypothesis create search results that include both confirming and disconfirming evidence about the validity of this statement or hypothesis. The findings of Study 1 provide support for this: among 38 participants, only 13 (or 34%) agreed with the validity of the statement in the existence of disconfirming evidence.

In an effort to shed light onto the reasons of biased decisions, we further examined the online activities of these 13 participants and found that five of them downloaded only the first page in their results, which happened to be confirming evidence, even though the results included disconfirming evidence. The other eight participants downloaded both confirming and disconfirming evidence but engaged in biased interpretation – ignored disconfirming evidence even after downloading and reading it. The reason for biased interpretation can be that the existence of confirming evidence alone may have increased participants' confidence in the validity of the statement provided to them. Nickerson (1996) provides support for this argument using Hempel's (1945) hypothesis, "all ravens are black," as an example:

> "*... observation of a single nonblack raven is enough to convince me that the statement that all ravens are black is false; however, the sighting of one or a few nonblack ravens will not necessarily shake my confidence in the truth of the statement that nearly all ravens are black, provided my sample is large enough to be considered representative of the general population and the percentage of the nonblack ravens in it is small.*" (Nickerson 1996, p.28).

However, a post-hoc analysis concerning decision confidence did not reveal any discernable differences in participants' level of confidence in their decisions in either study: decision confidence did not show any statistically significant differences between groups even after controlling for framing, initial familiarity, the type of documents read, or the number of documents read on confirming or disconfirming evidence. (See Appendix C for the results of this analysis.)

Limitations

This paper is not without its limitations. First, we conducted the experiment online without the presence of an experimenter. Therefore, it is possible that participants may have been distracted while completing the experimental task. However, the logs of participants (obtained from both Amazon Mechanical Turk and the web server used for the studies) indicated that participants did not take more than 10 minutes to complete the task. So, the experimental task was not too long or difficult to allow for distractions.

Second, we neither manipulated task importance, nor offered any incentive for decision accuracy. Although this raises concerns over participants' engagement in the experimental task, there is evidence that most participants took the experiment seriously: if we use the number of downloaded pages from search results as a proxy to the level of engagement of a participant, we can state that more than 70% of participants were engaged in the task since they downloaded at least two documents from their search results.

Third, the data collection instrument failed to save participants' demographic information such as age, gender, or proficiency with search engines. Therefore, we were not able to control for these factors in our data analyses.

Fourth, we did not define the terms hypertension or low blood pressure to participants. This may raise concerns over some of the findings, especially if participants do not know that hypertension and low blood pressure are the opposite ends of the same spectrum.

Finally, one can criticize the prevalence of dissimilar-term opposition statements in real-world by suggesting that there may be very few real-world scenarios that may fall under this category, or that it is unlikely in real-world that disconfirming evidence does not include crucial keywords (as in Study 2, where abstracts that support the link between coffee consumption and low blood pressure do not include the keyword "hypertension" anywhere in them). Even though these criticisms have validity, our experience is otherwise. First, there are many scenarios that fall under the dissimilar-term opposition category. Consider the link between attention and personality: a group of researchers argue that higher levels of attention are linked to introversion, while another group of researchers argue that they are associated with extraversion (see Koelega 1992). Similarly, consider the issue of aggressiveness among children under the age of six: certain studies posit that aggressiveness is more prevalent among boys, while others suggests that it is more common among girls (Maccoby and Jacklin 1976; Tieger 1980). Second, it is possible that searches conducted in titles or abstracts of these studies may not necessarily include keywords associated with counter arguments. In fact, the abstracts used in this study were created based on the real abstracts of peer-reviewed studies, which did not include the keywords observed in the counter argument (see Appendix A for details).

Practical Implications

The Web is an incredibly large and rich content repository that is only as useful as the search strategies we employ to find information: if we fall prey to our cognitive limitations or other extraneous traps, it is not difficult to steer ourselves toward one-sided, partial, or biased arguments, and therefore, make biased decisions. To say that one-sided or biased arguments always lead to suboptimal or incorrect decisions would be too simplistic an argument, because there can be many contexts in which biased arguments can be preferred. For example, they can help us make faster and easier decisions through heuristics (Gigerenzer and Todd 1999). However, there can be certain contexts in which biased arguments and our susceptibility to confirmation bias can cause more problems than we realize. For example, patients can experience confusion, anxiety, or uncertainty (Broom 2005; Meisel and Pines 2012; Wagner et al. 2001), physicians can diagnose incorrectly (Meisel and Pines 2012), investors can make risky decisions (Constable 2014); superstitious individuals can engage in harmful rituals (see Wagstaff 2014); and scientists can pursue unsubstantiated theories (see Gould 1981). This study concerns those contexts in which confirmation bias does more harm than good, and provides an initial step toward identifying the role of search contexts in understanding how confirmation bias unfolds on the Web.

Since search contexts, along with search engines, can be considered an extraneous factor, over which users have little to no control, one of the most important strategies for reducing their biasing effects can be training Internet users, perhaps starting at an early age, about conducting online searches for

hypothesis-testing tasks with special emphasis on formulating problems, devising search queries, and ultimately seeking both confirming and disconfirming evidence (Anderson 1982; Ferreira et al. 2006; Lau and Coiera 2009; Tweney et al. 1980).

Alternatively, we can fine-tune search engines to handle queries written toward dissimilar-term opposition scenarios (see Kayhan 2013). For example, search engines can use dictionaries, thesauri, or sematic networks (such as those used in Unified Medical Language System – UMLS) to have more contextual awareness and identify oppositions. Then, users can be provided with query recommendations, or more comprehensive search results that incorporate counter arguments so that they can make more informed decisions - interested readers can refer to the work of Kayhan (2013) to find more information about these strategies.

Conclusion

In this manuscript, our goal was to understand how search contexts can contribute to individuals' tendency to engage in confirmation bias on the Web. Using two studies, we demonstrated that scenarios, for which disconfirming evidence can be identified using a different term or phrase, can lead search engines to generate a result set consisting mostly – and sometimes only – of confirming evidence. This, in turn, leads to biased decisions. Our findings deepen our understanding of how confirmation bias unfolds on the Web when individuals seek information about the validity of a statement.

References

Anderson, C. A. 1982. "Inoculation and Counterexplanation: Debiasing Techniques in the Perseverance of Social Theories," *Social Cognition* (1), pp. 126-139.

Ariely, D. 2008. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York: Harper Collins

Broom, A. 2005. "Medical Specialists' Accounts of the Impact of the Internet on the Doctor/Patient Relationship," *Health* (9:3), pp. 319-338.

Burke, V., Beilin, L., German, R., Grosskopf, S., Ritchie, J., Puddey, I., and Rogers, P. 1992. "Association of Lifestyle and Personality Characteristics with Blood Pressure and Hypertension: A Cross-Sectional Study in the Elderly," *Journal of Clinical Epidemiology* (45:10), pp. 1061-1070.

Constable, S. 2014. "Twisting the Data." Retrieved November 6, 2014, from http://online.wsj.com/articles/confirmation-bias-what-it-means-1415048451

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. 2008. "An Experimental Comparison of Click Position-Bias Models," in: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. Palo Alto, California, USA: ACM, pp. 87-94.

Eysenbach, G., and Köhler, C. 2002. "How Do Consumers Search for and Appraise Health Information on the World Wide Web? Qualitative Study Using Focus Groups, Usability Tests, and in-Depth Interviews," *British Medical Journal* (324:7337), pp. 573-577.

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., and Sherman, J. W. 2006. "Automatic and Controlled Components of Judgment and Decision Making," *Journal of Personality and Social Psychology* (91:5), p. 797.

Festinger, L. 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Feufel, M. A., and Stahl, S. F. 2012. "What Do Web-Use Skill Differences Imply for Online Health Information Searches?," *Journal of Medical Internet Research* (14:3), p. e87.

Gigerenzer, G., and Todd, P. M. 1999. *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.

Gilovich, T. 1991. *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. New York: Free Press.

Gould, S. J. 1981. *The Mismeasure of Man*. New York: W.W. Norton.

Hempel, C. G. 1945. "Studies in the Logic of Confirmation (I.)," *Mind* (54:213), pp. 1-26.

Hodkinson, C., and Kiel, G. 2003. "Understanding Web Information Search Behavior: An Exploratory Model," *Journal of End User Computing* (15:4), pp. 27-48.

Horn, L. 1989. *A Natural History of Negation*. Chicago IL: University of Chicago Press.

Hostler, R. E., Yoon, V. Y., and Guimaraes, T. 2005. "Assessing the Impact of Internet Agent on End Users' Performance," *Decision Support Systems* (41:1), pp. 313-323.

Huang, H.-H., Hsu, J. S.-C., and Ku, C.-Y. 2012. "Understanding the Role of Computer-Mediated Counter-Argument in Countering Confirmation Bias," *Decision Support Systems* (53:3), pp. 438-447.

Ieong, S., Mishra, N., Sadikov, E., and Zhang, L. 2012. "Domain Bias in Web Search," in: *Proceedings of the fifth ACM international conference on Web search and data mining*. Seattle, Washington, USA: ACM, pp. 413-422.

Jee, S. H., He, J., Whelton, P. K., Suh, I., and Klag, M. J. 1999. "The Effect of Chronic Coffee Drinking on Blood Pressure: A Meta-Analysis of Controlled Clinical Trials," *Hypertension* (33:2), pp. 647-652.

Kakol, M., Jankowski-Lorek, M., Abramczuk, K., Wierzbicki, A., and Catasta, M. 2013. "On the Subjectivity and Bias of Web Content Credibility Evaluations," in: *Proceedings of the 22nd International Conference on World Wide Web Companion*. Rio de Janeiro, Brazil: International World Wide Web Conferences Steering Committee, pp. 1131-1136.

Kale, M., Bishop, T., Federman, A., and Keyhani, S. 2011. ""Top 5" Lists Top $5 Billion," *Archives of Internal Medicine* (171:20), pp. 1858-1859.

Kayhan, V. O. 2013. "Seeking Health Information on the Web: Positive Hypothesis Testing," *International Journal of Medical Informatics* (82:4), pp. 268-275.

Keselman, A., Browne, A. C., and Kaufman, D. R. 2008. "Consumer Health Information Seeking as Hypothesis Testing," *Journal of the American Medical Informatics Association* (15:4), pp. 484-495.

Klayman, J. 1995. "Varieties of Confirmation Bias," *Psychology of Learning and Motivation* (32), pp. 385-418.

Koelega, H. S. 1992. "Extraversion and Vigilance Performance: 30 Years of Inconsistencies," *Psychological Bulletin* (112:2), pp. 239-258.

Kulviwat, S., Guo, C., and Engchanil, N. 2004. "Determinants of Online Information Search: A Critical Review and Assessment," *Internet Research* (14:3), pp. 245-253.

Lau, A., and Coiera, E. W. 2007a. "Do People Experience Cognitive Biases While Searching for Information?," *Journal of the American Medical Informatics Association* (14:5), pp. 599-608.

Lau, A., and Coiera, E. W. 2009. "Can Cognitive Biases During Consumer Health Information Searches Be Reduced to Improve Decision Making?," *Journal of the American Medical Informatics Association* (16:1), pp. 54-65.

Lau, A. Y. S., and Coiera, E. W. 2007c. "How Do Clinicians Search for and Access Biomedical Literature to Answer Clinical Questions?," *MEDINFO 2007*, Brisbane, Australia.

Lueg, J. E., Moore, R. S., and Warkentin, M. 2003. "Patient Health Infor Mation Search: An Exploratory Model of Web-Based Search Behavior," *Journal of End User Computing* (15:4), pp. 49-61.

Maccoby, E. E., and Jacklin, C. N. 1976. *The Psychology of Sex Differences*. Stanford, CA: Stanford University Press.

Marchionini, G., and White, R. 2007. "Find What You Need, Understand What You Find," *International Journal of Human Computer Interaction* (23:3), pp. 205-237.

Markoff, J. 2008. "Microsoft Examines Causes of 'Cyberchondria' " *NYTimes.com*, Retrieved August 7, 2012, from http://www.nytimes.com/2008/11/25/technology/internet/25symptoms.html?_r=1

Meisel, Z. F., and Pines, J. M. 2012. "'July Effect' Revisited: Why Experienced Docs May Not Deliver the Best Care." *Time*, Retrieved August 7, 2012, from http://healthland.time.com/2012/07/17/the-july-effect-revisited-why-the-most-experienced-doctors-dont-always-deliver-the-best-medical-care/#ixzz22tGX5n4e

Nickerson, R. S. 1996. "Hempel's Paradox and Wason's Selection Task: Logical and Psychological Puzzles of Confirmation," *Thinking and Reasoning* (2:1), pp. 1-31.

Nickerson, R. S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* (2:2), pp. 175-220.

Nurminen, M. L., Niittynen, L., Korpela, R., and Vapaatalo, H. 1999. "Coffee, Caffeine and Blood Pressure: A Critical Review," *European Journal of Clinical Nutrition* (53:11), p. 831.

Palatini, P., Dorigatti, F., Santonastaso, M., Cozzio, S., Biasion, T., Garavelli, G., Pessina, A. C., and Mos, L. 2007. "Association between Coffee Consumption and Risk of Hypertension," *Annals of Medicine* (39:7), pp. 545-553.

Park, J., Konana, P., Gu, B., Kumar, A., and Raghunathan, R. 2013. "Information Valuation and Confirmation Bias in Virtual Communities: Evidence from Stock Message Boards," *Information Systems Research* (24:4), pp. 1050-1067.

Periti, M., Salvaggio, A., Quaglia, G., and Di Marzio, L. 1987. "Coffee Consumption and Blood Pressure: An Italian Study," *Clinical Science* (72:4), p. 443.
Rabin, R. C. 2012. "Doctor Panels Recommend Fewer Tests for Patients." *NYTimes.com*, Retrieved June 18, 2013, from http://www.nytimes.com/2012/04/04/health/doctor-panels-urge-fewer-routine-tests.html?_r=0
Schulz-Hardt, S., Frey, D., Lüthgens, C., and Moscovici, S. 2000. "Biased Information Search in Group Decision Making," *Journal of Personality and Social Psychology* (78:4), pp. 655-669.
Steelman, Z. R., Hammer, B. I., and Limayem, M. 2014. "Data Collection in the Digital Age: Innovative Alternatives to Student Samples," *MIS Quarterly* (38:2), pp. 355-378.
Stensvold, I., Tverdal, A., and Per Foss, O. 1989. "The Effect of Coffee on Blood Lipids and Blood Pressure. Results from a Norwegian Cross-Sectional Study, Men and Women, 40-42 Years," *Journal of Clinical Epidemiology* (42:9), pp. 877-884.
Swann, W. B., Griffin, J. J., Predmore, S. C., and Gaines, B. 1987. "The Cognitive–Affective Crossfire: When Self-Consistency Confronts Self-Enhancement," *Journal of Personality and Social Psychology* (52:5), pp. 881-889.
Taylor, S. E., and Brown, J. D. 1988. "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin* (103:2), pp. 193-210.
Tieger, T. 1980. "On the Biological Basis of Sex Differences in Aggression," *Child Development* (51:4), pp. 943-963.
Tversky, A., and Kahneman, D. 1974. "Judgment under Uncertainty: Heuristics and Biases," *science* (185:4157), pp. 1124-1131.
Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., and Arkkelin, D. L. 1980. "Strategies of Rule Discovery in an Inference Task," *Quarterly Journal of Experimental Psychology* (32:1), pp. 109-123.
Uiterwaal, C. S. P. M., Verschuren, W. M. M., Bueno-de-Mesquita, H. B., Ocké, M., Geleijnse, J. M., Boshuizen, H. C., Peeters, P. H. M., Feskens, E. J. M., and Grobbee, D. E. 2007. "Coffee Intake and Incidence of Hypertension," *The American Journal of Clinical Nutrition* (85:3), p. 718.
Wagner, T., Hibbard, J., Greenlick, M., and L., K. 2001. "Does Providing Consumer Health Information Affect Self-Reported Medical Utilization," *Medical Care* (39), pp. 836-847.
Wagstaff, K. 2014. "'Slender Man' Cited in Stabbing Is a Ghoul for the Internet Age." Retrieved November 6, 2014, from http://www.nbcnews.com/storyline/slender-man-stabbing/slender-man-stabbing-report-finds-suspect-competent-stand-trial-n231461
Wakabayashi, K., Kono, S., Shinchi, K., Honjo, S., Todoroki, I., Sakurai, Y., Umeda, T., Imanishi, K., and Yoshizawa, N. 1998. "Habitual Coffee Consumption and Blood Pressure: A Study of Self-Defense Officials in Japan," *European Journal of Epidemiology* (14:7), pp. 669-673.
White, R. 2013. "Beliefs and Biases in Web Search," *Proceedings of the 36th international ACM SIGIR: Conference on Research and Development in Information Retrieval*, Dublin, Ireland: ACM, pp. 3-12.
Yue, Y., Patel, R., and Roehrig, H. 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*. Raleigh, North Carolina, USA: ACM, pp. 1011-1018.

APPENDIX A

The development of the abstracts

Using two databases (Google Scholar and PubMed), we searched for peer-reviewed studies that examined the relationship between coffee consumption and blood pressure. We identified four articles in favor of the relationship between coffee consumption and hypertension (see Burke et al. 1992; Jee et al. 1999; Palatini et al. 2007; Uiterwaal et al. 2007), and three articles in favor of the relationship between coffee consumption and low blood pressure (see Periti et al. 1987; Stensvold et al. 1989; Wakabayashi et al. 1998). Note that the abstracts of the articles in favor of hypertension did not include the phrase "low blood pressure" anywhere in them, and the abstracts of articles in favor of low blood pressure did not include the term "hypertension" anywhere in them.

The studies were published in a span of 20 years in different journals; used samples from various countries and age groups; reported other disorders; and used inconsistent control variables. Pilot studies revealed that these were a major source of confusion for participants, confounding the experimental task. Therefore, we changed these abstracts (by staying as faithful as possible to the original abstracts) and removed certain details about the study samples (such as participants' ages and nationalities), made sure each abstract reported the same control variables, removed findings about other health problems unrelated to hypertension or low blood pressure, and changed article details to make them look like they were published in the same year and in the same journal. We also created a fourth abstract in favor of the relationship between coffee consumption and low blood pressure to balance the number of studies.

For Study 1, we created another set of four abstracts – using the abstracts that favored the link between coffee consumption and low blood pressure – such that we replaced the phrase "low blood pressure" with "hypertension" and changed the tone of the corresponding sentences so that the new abstracts failed to support the link between coffee consumption and hypertension (to satisfy the "Coffee consumption & No hypertension" case). At last, we had three sets of abstracts based on the original seven studies: 1) four abstracts indicated that there was a link between coffee consumption and hypertension; 2) four abstracts indicated that there was no link between coffee consumption and hypertension; and 3) four abstracts indicated that there was a link between coffee consumption and low blood pressure. To see if the new abstracts were a good proxy of the original abstracts, we conducted a set of pretests – discussed next.

Pretest – Study 1

For Study 1, we focused on the two new sets of abstracts: those that indicated a link between coffee consumption and hypertension, and those that indicated that there was no such link. Our goal was to make sure that these abstracts did not lack credibility. Using student participants, we measured participants' level of confidence in the findings reported in these abstracts and compared them to those reported in the original abstracts. To this end, six participants were given all four original abstracts in favor of coffee consumption and hypertension, six participants were given all four new abstracts in favor of coffee consumption and hypertension, and five participants were given all four new abstracts that indicated that there was no link between coffee consumption and hypertension.

The abstracts were provided to each participant in random order to reduce order effects. Comparison of mean confidence scores of each participant revealed statistically significant differences ($F(2,14)=14.48$, $p<0.001$). Mean confidence score for the original abstracts were 4.42 (on a 7-point Likert scale) and statistically lower than the mean confidence scores for the new abstracts in favor of coffee consumption and hypertension (5.75, $p<0.001$), as well as the mean confidence scores for the new abstracts that indicated there was no link (5.60, $p<0.001$). There was no statistical difference between the two new sets of abstracts (5.75 versus 5.60, $p=0.60$).

Further, there were no within-group differences among the new abstracts: a one-way ANOVA showed that the mean confidence score for each new abstract was not significantly different from other means: $F(3,20)=1.23$; $p=0.33$ for abstracts in favor of coffee consumption and hypertension; and $F(3,16)=1.67$; $p=0.21$ for abstracts that indicate there is no link. Therefore, the pretest for Study 1 showed that the new sets of abstracts were a good proxy of the original abstracts.

Pretest – Study 2

For Study 2, we first focused on the believability of the findings reported in the new set of abstracts that favored the link between coffee consumption and low blood pressure. Therefore, we recruited another 13 student participants to measure their level of confidence in the results reported by the new abstracts. Six of these were given the all three original abstracts that favored the link between coffee consumption and low blood pressure, and seven were given all four new abstracts that favored the same link. We gave the abstracts to each participant in random order to reduce the order effects.

Participants had a higher level of confidence for the new abstracts. The mean confidence score, on a 7-point Likert scale, was 4.78 for original abstracts and 5.86 for new abstracts ($F(1,11)=16.16$, $p=0.002$). Further, there were no within-group differences among the new abstracts. A one-way ANOVA showed that the mean confidence score of the new eight abstracts were not significantly different from other means ($F(3,24)=0.65$; $p=0.59$).

Next, we focused on the similarity between the original abstracts and the new set of abstracts. To this end, we conducted Study 2 using only the original seven abstracts – four of which were in favor of coffee consumption and hypertension, and the remaining three were in favor of coffee consumption and low blood pressure. Thirteen other student participants took part in this pretest who did not know the number or nature of the pages hosted on the server. Six participants were assigned to the link between coffee consumption and hypertension, and the other seven were assigned to the link between coffee consumption and low blood pressure. As participants completed the experimental task, we recorded their queries and the keywords used in each query. Later, we sent these queries to the same experimental setup that indexed the eight new abstracts. Our goal was to compare the results obtained from the original abstracts with the results obtained from the new abstracts. If queries returned the same set of pages, then the new abstracts would be considered similar to the original abstracts.

For the relationship between coffee consumption and hypertension, all queries provided consistent results: if a specific query returned only confirming evidence from the original set of abstracts, it also returned only confirming evidence from the new set of abstracts. For the relationship between coffee consumption and low blood pressure, the new abstracts outperformed the original abstracts for certain queries. This was because one of the original abstracts reported an *inverse* relationship between coffee and blood pressure and was not being included in the search results when participants used the phrase "low blood pressure" in their queries. On the other hand, queries returned all four abstracts from the new set, providing better results. Overall, the pretest of Study 2 showed that the new abstracts was a good proxy of the original abstracts.

APPENDIX B

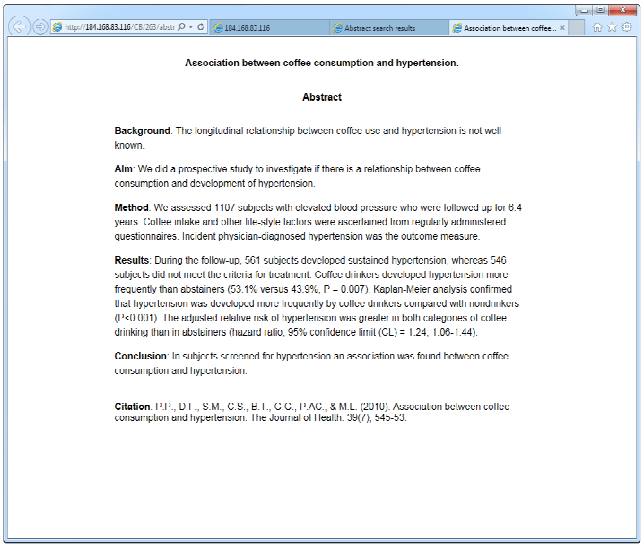

Figure B1. Sample Abstract

APPENDIX B (continued)

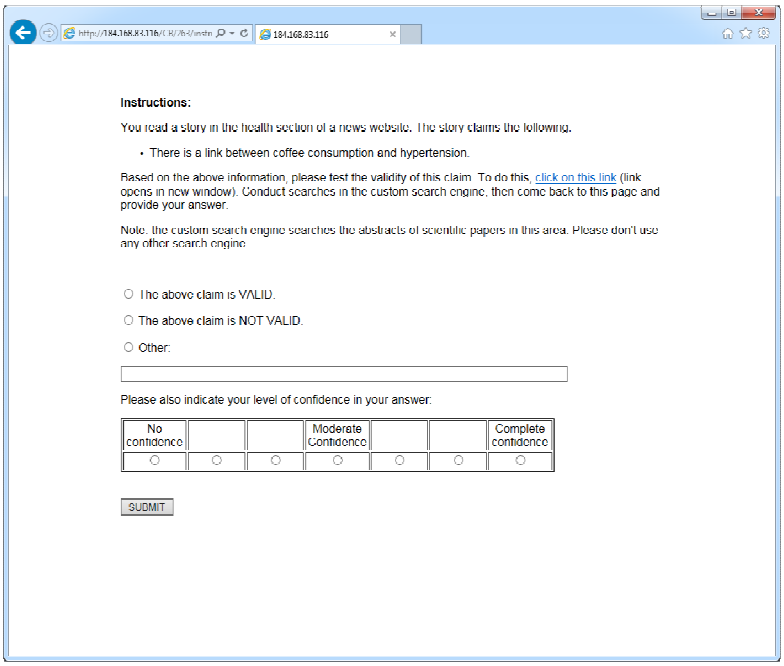

Figure B2. Instructions

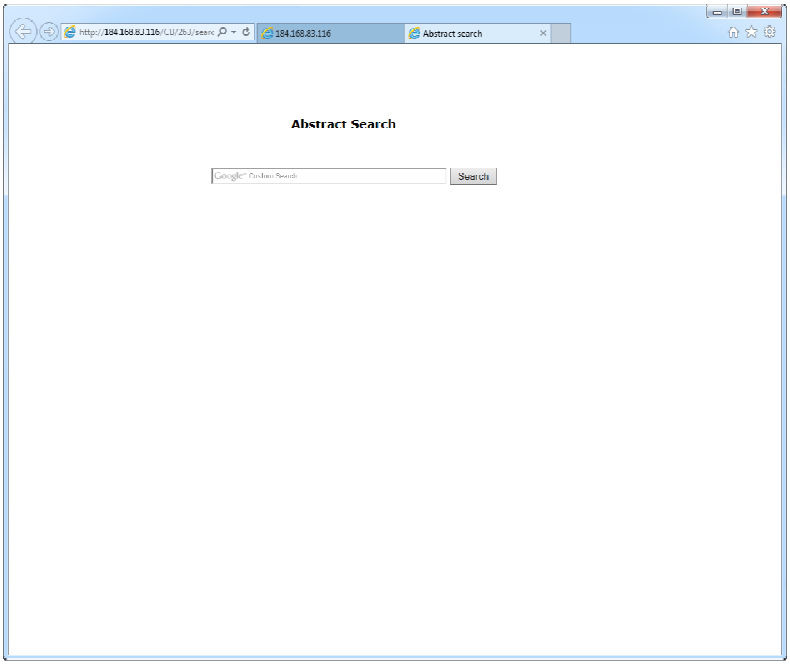

Figure B3. Search Page

Figure B4. Results Page

APPENDIX C

We examined participants' confidence in their answers using an ANCOVA, where the dependent variable was decision confidence, the independent variables were framing and participants' answers, and covariates were the number of downloaded pages on each argument, the type of downloaded pages (a binary variable), and participants' initial familiarity.

Study 1: none of the variables were significant. Participants who tested the positively framed statement were, on average, less confident in their answers than those that tested the negatively framed statement, but the difference was not significant (4.97 versus 5.42 respectively, p=0.32); and participants who indicated that the statement provided to them was valid were, on average, less confident in their answers than those who indicated that it was invalid (even after controlling for framing), but the difference was still not significant (5.0 versus 5.39, p=0.46).

Study 2: none of the variables were significant. Participants who tested the link between coffee consumption and hypertension were, on average, more confident in their answers than those who tested the link between coffee consumption and low blood pressure, but the difference was not significant (5.6 versus 5.0, p=0.54). Participants who indicated that the statement provided to them was valid were, on average, more confident in their answers than those who indicated that it was invalid (even after controlling for framing), but the difference was not significant (5.4 versus 5.2, p=0.79).