Belief Dynamics and Biases in Web Search

RYEN W. WHITE and ERIC HORVITZ, Microsoft Research

We investigate how beliefs about the efficacy of medical interventions are influenced by searchers' exposure to information on retrieved Web pages. We present a methodology for measuring participants' beliefs and confidence about the efficacy of treatment before, during, and after search episodes. We consider interventions studied in the Cochrane collection of meta-analyses. We extract related queries from search engine logs and consider the Cochrane assessments as ground truth. We analyze the dynamics of belief over time and show the influence of prior beliefs and confidence at the end of sessions. We present evidence for confirmation bias and for anchoring-and-adjustment during search and retrieval. Then, we build predictive models to estimate post-search beliefs using sets of features about behavior and content. The findings provide insights about the influence of Web content on the beliefs of people and have implications for the design of search systems.

Categories and Subject Descriptors: **H.3.3 [Information Storage and Retrieval]**: Information Search and Retrieval – *Search process*, *Selection process*.

Additional Key Words and Phrases: Belief dynamics; Search Interaction, Cognitive biases.

1. INTRODUCTION

People use search and retrieval systems to learn about the world and to inform decisions. From a Bayesian perspective, people searching for information can be viewed as starting a search session with personal probabilities about the truth of assertions of facts and outcomes of interest. As content is retrieved and reviewed, such beliefs may be revised. The influences of retrieved content on the beliefs of searchers can be monitored by tracking how their beliefs change as they examine search engine result pages (SERPs) and landing pages [Joachims et al. 2007; Buscher et al. 2008]. Belief updating with exposure to new information can be influenced by assessments of the credibility of information and to multiple heuristics and biases of judgment [Tversky and Kahneman 1974] that describe nuances of human belief updating. Beyond pursuing an understanding of how people update their beliefs in light of the review of specific information, we must also consider *how* people select, examine, and digest content during search and retrieval sessions. Searchers' interests, anxieties, and biases may draw people toward information pertaining to particular outcomes [White and Horvitz 2009; White and Horvitz 2010; White 2013]. The information selected and reviewed by searchers may also be influenced by structure of the presentation of content (e.g., the ordering of concepts in a presentation [White and Horvitz 2010]). Biases of judgment demonstrated in cognitive psychology include such findings as people having preferences for information that supports their prior position over information refuting it, regardless of the factual correctness [Tversky and Kahneman 1974; Baron 2007; Ariely

Authors' addresses: R. White, Microsoft Research, One Microsoft Way, Redmond WA 98052. E. Horvitz, Microsoft Research, One Microsoft Way, Redmond WA 98052; Email: {ryenw,horvitz}@microsoft.com.

DOI:http://dx.doi.org/10.1145/0000000.0000000

2008]. In addition, prior beliefs may contribute to overconfidence, which can make it difficult for people to update their beliefs [Griffin and Tversky 1992].

The dynamics of information needs have been well studied in the information science community [Bates 1989; O'Day and Jeffries 1994; Belkin et al. 1995]. However, beliefs and belief revision has not been deeply explored in search and retrieval settings. A better understanding of how beliefs change during search can assist in the development of enhanced ranking methods [Losada and Barreiro 1999] or adaptive search algorithms [Lau et al. 2004] (although these referenced methods focus on beliefs about *relevance* rather than facts or task outcomes). Modeling search beliefs enables richer models for recommendation or personalization, which have typically been modeled solely at the level of topical interest, e.g., [Teevan et al. 2005]. Additionally, information about the beliefs of searchers can be used to build systems capable of persuading searchers to adopt a specific view or perspective [Fogg 2002; Berkovsky et al. 2012], or to mitigate the effects of misinformed beliefs on task outcomes—especially for scenarios involving decision making under uncertainty.

We investigate belief revision during search. We focus on the efficacy of medical interventions and define beliefs in terms of an assessed likelihood of intervention efficacy. We conducted a study using remote participants and an in situ belief measurement methodology where we assigned search tasks via a crowdsourcing platform, and asked participants to examine search results. We assessed participant beliefs and associated confidence estimates before, during, and after each search task. Assessing confidence in addition to assessing likelihoods is important as levels of confidence may directly influence action [Heath and Tversky 1991; Griffin and Tversky 1992]. Confidence is also related to other aspects of search engine measurement, such as satisfaction [Kuhlthau 1991]. We examine several aspects of belief dynamics, such as the relationship between beliefs and confidence in search, as well as belief updating given exposure to Web content.

We shall present a methodology for capturing searchers' beliefs in situ during search-based question-answering tasks, deployed via a crowdsourcing platform. We show how we can employ the methodology to analyze belief updating in search scenarios for a range of tasks in the health domain, for which we have truth as asserted by a trusted consensus study. We find that there are strong biases in searchers' observed interaction behavior toward reported prior beliefs (i.e., strong evidence of confirmation biases [Tversky and Kahneman 1974]). We modify the study to control for content biases associated with the quantity and ranking of results related to certain outcomes (positive skew). We show that evidence of confirmation bias persists even after the removal of these additional effects. We also study the relationship between confidence and beliefs. We find that searchers exhibit significant overconfidence, similar to that observed in non-search settings [Griffin and Tversky 1992], which influences the degree of belief updating during searching. Finally, we develop predictive models capable of estimating post-search belief ratings using information available before the search and/or features of search interaction behavior. We show that we can attain strong performance at this task using implicit behavioral and content signals only (i.e., without belief or confidence ratings captured from searchers).

The remainder of this article is structured as follows. In Section 2, we describe related research in the areas of information need formation and dynamics, beliefs and biases in search interaction behavior, belief revision in search systems, confidence, crowdsourced user studies, and content quality and searcher trust. Section 3 describes the studies that we perform as part of our investigation into belief dynamics, including predictive models of belief updating. In Section 4 we report the findings of our studies, including the performance of the predictive models. We discuss our findings and their implications in Section 5, and conclude in Section 6.

2. RELATED WORK

Related work can be found in several different realms of study. Relevant prior research includes work on: (i) the dynamics of information needs, (ii) beliefs and biases in search and retrieval, (iii) beliefs in retrieval systems, (iv) the relationship between confidence and beliefs, (v) crowdsourcing as a methodology for user studies, and (vi) content quality and searcher trust.

2.1 Information Need Dynamics

Questions arise in the context of advancing one's state of knowledge or to inform a forthcoming decision. Searches for information can be directed or undirected, depending on whether there is a clear objective associated with the information obtained. Information scientists have analyzed the cognitive mechanisms behind the search for information, including the development of models for how information needs emerge [Belkin et al. 1982; Taylor 1968] and how they evolve during search [Kuhlthau 1991; Marchionini 1995]. Search is thought to be motivated by an incompleteness [Mackay 1960; Taylor 1968] or a "problematic situation" [Belkin et al. 1982] in the mind of the searcher that develops into a desire for information. There are a number of ways in which this may be characterized, including a gap [Dervin 1983], a visceral need [Taylor 1968], an anomaly in a searcher's knowledge state [Belkin et al. 1982], or as an unstable collection of noumena [Marchionini 1995]. Although researchers have studied changes in information needs during the course of search episodes [Bates 1989; O'Day and Jeffries 1994; Belkin et al. 1995], little research has been on how beliefs in facts or outcomes of actions evolve during a search episode.

2.2 Beliefs and Biases

As searchers review information delivered by a search system in response to their expressed needs, their beliefs about the topics they are searching on may be updated. The normative standard for modeling belief updating is to employ the rules of probability, where prior beliefs are updated with Bayes rule into posterior beliefs about hypotheses in light of additional evidence. Bayesian models of the updating of beliefs include Jeffrey conditioning, which introduces a particular approach to handling uncertainty about the validity of evidence presented to people [Jeffrey 1990]. Jeffrey conditioning has been used to guide the provision of implicit relevance feedback in search [White et al. 2005].

Many have discussed and studied the irrationality of people from the perspective of normative belief updating and decision making [e.g., Elster 1979; Gigerenzer and Todd 2000; Simon 1955, Tversky and Kahneman 1974]. Biases in peoples' beliefs can influence their search behavior, significantly affecting their judgment and decision making. Psychologists have examined biases in beliefs [Gigerenzer and Todd 2000; Klayman and Ha 1987; Tversky and Kahneman 1974]. Belief dynamics has also been researched extensively in the same community [Anderson 1981; Hogarth and Einhorn 1992; Tversky and Kahneman 1974]. Research on cognitive dissonance and selective exposure to attitude-supporting information [Festinger 1957; Fischer et al. 2011; Frey 1986; Hart et al. 2009] suggests that information seekers favor information that supports their beliefs, driven by the pursuit of both accurate and confirmatory information, dual objectives that may be in tension, especially for strongly-held beliefs.

Prior research has considered biases of different kinds in the context of information seeking. These include presentation biases, associated with the sequence order in which content is presented to searchers by search systems [Joachims et al. 2007]. Other biases include those associated with the content of captions, whereby the presence of particular terms [Clarke et al. 2007] has been shown to result in clickthrough *inversions* where searchers favor lower ranked results. Other features of captions, such as the presence of potentially-alarming content [White and Horvitz 2013] (e.g., the phrase “heart attack” appearing in captions for the query [chest pain]) or even bolded terms in result titles [Yue et al. 2010], can influence the likelihood of accessing the associated content. Ieong et al. [2012] demonstrated that domain preferences could also influence search-examination behavior, drawing searchers to choose results with favored Web domains irrespective of relevance.

There is growing interest in biases in search and the effect of search results on searchers' cognitive states [White and Horvitz 2009; White 2013] and emotional states [Lauckner and Hsieh 2013]. Recent research in the medical informatics community has shown that health information seekers may be affected by cognitive biases, in particular confirmation biases [Lau and Coiera 2007; 2009]. Other work has provided evidence that searchers demonstrate a preference for content that answers posed questions affirmatively, favoring content with “yes” answers for a balanced set of *yes-no* health questions, and that beliefs may be insensitive to manipulations favoring particular answers [White 2013; 2014]. Liao and Fu [2013] performed a user study of belief updating in information access, using controversial topics and side-by-side presentation of opposing views. They found that participants preferentially selected content that reinforced existing attitudes, but also that their attitudes were moderated by access to opposing content. They also found that those with strong beliefs were least susceptible to attitude moderation.

People's beliefs guide decisions and information gathering. Thus, biases in beliefs and belief updating can have detrimental impact on actions in the world. People have also been shown to trust the output of search engines [Joachims et al. 2007]. As an example of the influence of search engines on beliefs, people performing self-diagnosis via symptom searches have been found to often associate the ranking of search results as an ordering over the likelihoods of different medical conditions [White and Horvitz 2009]. Furthermore, the ranking of results can be skewed in searchers on common, typically benign symptoms, where less likely but more concerning medical disorders may be presented at the top of result lists [White and Horvitz 2009]. The content of search engine result lists is also likely to be skewed toward the effectiveness of medical treatment options, and certain query terms, such as “can” or “help”, contribute significantly to result skew White and Hassan [2014].

Aggregated clickthrough behavior is often leveraged by search engines to improve their performance [Joachims 2002; Agichtein et al. 2006]. Ranking algorithms may learn to select and order search results that are biased toward particular perspectives if many searchers select information that is aligned with popular beliefs, including myths and common misconceptions, driven by preferential attachment [Cho and Roy 2004; Goldman 2006]. On the potential value of providing diverse viewpoints in search results, recent studies have found that people may benefit from exposure to a range of opinions [Mankoff et al. 2011; Munson and Resnick 2010].

2.3 Belief Revision in Retrieval Systems

Researchers in information retrieval (IR) have studied the use of belief revision in search systems, especially systems that adapt to user interests. Conceptually, given a retrieval context, the IR system needs to specify a focus (i.e., a searcher's specific interest) over a context. Models of human information processing have been invoked to explain how stimuli might trigger a spreading activation process [Card et al. 1983]. Formal models of the belief revision process have been used to quantify query-document similarity values employed in search result ranking [Losada and Barreiro 1999].

Logan et al. [1994] studied belief revision in the context of the cognition of a librarian agent, which revised its beliefs (albeit inefficiently) based on natural language feedback. Lau et al. [2004] proposed a mechanism for modeling beliefs in IR systems using belief revision and information flow, such that it outperforms more traditional means of information filtering. The primary focus in these models is beliefs concerning relevance, rather than beliefs about target facts or outcomes, as we focus on in our study. The contribution in many of the prior studies about beliefs and retrieval comes in the context of research on models of cognition and retrieval. We are not familiar with prior studies that seek to measure the dynamics of belief during search and retrieval.

2.4 Beliefs and Confidence

An important aspect of understanding beliefs and their impact relates to confidence. Confidence can govern human behavior [Heath and Tversky 1991; Griffin and Tversky 1992] but has been largely unexplored in search and retrieval settings. Kuhlthau [1991] showed that optimism and confidence can drive the search process, and may be heightened toward the termination of longitudinal search processes. Cognitive psychologists have studied *overconfidence*. The overconfidence effect is a bias in which subjective confidence in the accuracy of an answer exceeds the objective accuracy of that answer [Pallier et al. 2002]. Psychologists have found that people may resist updating their beliefs in light of new evidence [Heath and Tversky 1991; Griffin and Tversky 1992]. Overconfidence effects have been observed in many disciplines, including medicine [Lusted 1977], clinical psychology [Oskamp 1965], and negotiation [Neale and Bazerman 1990]. Explanations for this effect center on people's tendency to focus on the salience of the evidence over its credibility. As part of our study, we explore the relationship between searcher beliefs and confidence. We also examine the impact of pre-search confidence on belief dynamics. In doing so, we address the need for a better understanding of confidence and the connection between confidence and beliefs in information-seeking contexts.

2.5 User Studies via Crowdsourcing Platforms

User studies of search behavior have traditionally been conducted in a laboratory setting (see [Kelly 2009] for an excellent summary). The recent emergence of crowdsourcing platforms such as Mechanical Turk and CrowdFlower has enabled low-cost, carefully-controlled studies of human behavior [Kittur et al. 2008; Paolacci et al. 2010]. This approach is being used for a range of experiments, including those associated with satisfaction modeling [Ageev et al. 2011], estimating attention [Lagun and Agichtein 2011], and evaluating relevance [Alonso et al. 2008]. White [2014] also employed a crowdsourced methodology to collect belief ratings from searchers before and after engaging with a search engine (but importantly not during the search process as we do in this study). We outline the differences between the current study and the White [2014] study at the end of this section.

2.6 Content Quality and Searcher Trust

As mentioned above, searchers may often blindly trust the rank ordering of the search results (so-called "trust bias") and select highly-ranked results irrespective of relevance [Joachims et al. 2007]. Trust is an important aspect of belief revision. Richardson et al. [2003] modeled trust in information shared by others in online settings and its relationship with personal beliefs. Trust is also related to confidence: if people trust the content that they view, they are more likely to be confident in the beliefs that they form. The high degree of trust that searchers place in search engine result rankings can lead to heightened concerns, erroneous beliefs, and negative emotional outcomes [White and Horvitz 2009; Lauckner and Hsieh 2013]. Searchers have been shown to associate the result ranking for medical symptom queries with a ranking of health conditions by likelihood of occurrence, even though search engines ranking algorithms do not consider veracity or normative data [White and Horvitz 2009]. Lauckner and Hsieh [2013] studied the effect of health content on the emotional state of Web searchers posing queries for medical symptoms. They showed that presenting serious illnesses in snippets at higher ranked positions led to negative emotional outcomes such as heightened searcher anxiety.

The level of trust that people place in online health information is uncertain and affected by factors such as age and gender (e.g., younger people are more likely to trust online health information) [Hesse et al. 2005]. Trust bias is potentially problematic in the health domain since online health information is often of low quality [Eysenbach and Kohler, 2002; Bengeri and Pluye 2003] and health seekers have been shown to ignore key quality indicators such as source validity or source creation date when examining health content [Fox 2006]. Cline and Haynes [2001] suggest that public health professionals should be concerned about the prevalence of online health seeking; in their study they consider the potential benefits of this activity, synthesize quality concerns, and identify criteria that could be used to evaluate online health information. Although search engines are an important part of obtaining health information, 70% of U.S. adults still turn to physicians or other health care professionals for information, care, or support regarding serious health concerns [Fox and Duggan 2013].

More generally, it has been argued that the lack of regulation over online health content raises important ethical and legal challenges [Boyer 2013]. To address quality concerns, services have emerged that offer external verification on the reliability of health-related web content (e.g., Health on the Net (hon.ch) and URAC (urac.org)). These sites assign quality scores to Web pages based on human review of their content; although importantly, they do not verify the correctness of any claims made on those sites. These labels, and other reliability signals, have been used for ranking within specialized websites [Gaudinat et al. 2006] or to predict escalations in concerns following the review of Web content [White and Horvitz 2010].

Measures of Web page quality have also been shown to be effective in supporting result selection decisions [Schwarz and Morris 2011] and can impact searcher trust [Sillence et al. 2004]. Schwarz and Morris [2011] identified page features associated with the credibility of online content, and presented methods to augment search-result presentation with credibility features to help people find more trustworthy information and make more reliable decisions. Sillence et al. [2004] studied the influence of design and content on the trust and mistrust of health Websites via an observational study with a small number of participants engaged in structured and unstructured search sessions. They found that aspects of the design could engender mistrust in the content, whereas the credibility of information and personalization of content served to engender trust. The *HealthTrust* system [Fernandez-Luque et al. 2012] leverages social network analysis to find trustworthy social media in online health communities.

2.7 Contributions over Previous Work

The research presented in this article extends prior work in a number of ways. We explore dynamics of belief during the *process* of search and retrieval. Traditionally, the focus in research on the information seeking process is on changes in information needs that occur during searching. We focus on the changes in subjective probability distributions about the truth of medical facts and their association with observed search behaviors. In doing so, we examine changes in beliefs *during* the search process as well as before and afterwards. Since confidence affects belief revision, we study participants' assessments of their own confidence in likelihoods as part of characterizing beliefs. Considering confidence can provide insights on the influence of strongly-held versus weakly-held beliefs on belief updating with the review of content. We construct predictive models of belief revision based on both explicit self-reporting data and implicit signals gathered based on content accessed and aspects of search activity.

In related work by White [2014] on beliefs about health questions with *yes* and *no* answers, a simpler analysis of changes of qualitative assessment following exposure to content was undertaken, using estimates of two physicians as ground truth (i.e., questions where both physicians agreed that the answer is either *yes* or *no*). In contrast, we investigate queries about intervention efficacy, employ an internationally-recognized source of medical evidence as ground truth, capture beliefs from participants during the search, as well as before and after, consider the role of confidence, and explore the challenge of predicting post-search beliefs based on implicit and explicit signals collected during the search process. We also employ a more sophisticated sampling methodology for controlled experiments.

3. STUDY

We now describe the study aimed at understanding belief dynamics during search and retrieval. We focused on health search given its importance (e.g., 80% of U.S. adults report searching for health information online, mainly for consequential tasks such as medical self-diagnosis and treatment [Fox and Duggan 2013]), and our familiarity and expertise in health information search and retrieval. This section describes our research questions, data, experimental instruments, and the methodology followed.

3.1 Research Questions

The following seven research questions guided our investigation:

—**RQ1:** How do beliefs change during the search process?

—**RQ2:** What is the influence of pre-search beliefs on different aspects of the search process, including (i) the search results selected, (ii) the time spent reviewing results, and (iii) the eventual outcome of the search process (post-search beliefs)?

—**RQ3:** What is the effect of the content of pages viewed, including the position of strong evidence with the pages examined, on belief revision?

—**RQ4:** How do some of the salient aspects of RQ1-RQ3 vary when we control for the availability of content related to each outcome (*helps* vs. *does not help*) on the SERP?

—**RQ5:** What is the relationship between beliefs and confidence? How accurate are searchers even if they are confident that they have achieved the correct answer? That is, is there evidence of overconfidence and how does that impact beliefs?

—**RQ6:** Is there evidence of synthesis of content encountered during the search, meaning that beliefs may not change immediately to reflect examined content? This has implications for real-time belief adaptation by search systems. We operationalize that in terms of whether belief ratings change following the last page view (i.e., in the time between providing the final post-page rating and the post-search rating)?

—**RQ7:** Can we accurately estimate the nature of belief revision based on features of searcher behavior and content? How does the predictive accuracy shift with leveraging information about searchers (from multiple observations over time) and with gaining self-assessments of prior beliefs about the topic at hand?

Title: *Melatonin for the prevention and treatment of jet lag*

Background: *Jet lag commonly affects air travelers who cross several time zones. It results from the body's internal rhythms being out of step with the day-night cycle at the destination. Melatonin is a pineal hormone that plays a central part in regulating bodily rhythms and has been used as a drug to re-align them with the outside world.*

Summary: *Melatonin is remarkably effective in preventing or reducing jet lag, and occasional short-term use appears to be safe. It should be recommended to adult travelers flying across five or more time zones, particularly in an easterly direction, and especially if they have experienced jet lag on previous journeys. Travelers crossing 2-4 time zones can also use it if need be.*

Figure 1. Title, background, and plain language summary from sample Cochrane review on the use of melatonin for jet lag (label=*helps*).

Answers to these questions are important in understanding belief dynamics during search and retrieval and the ultimate influence of search on beliefs. Answers would help guide the designs of systems that provide searchers with the most valuable content, considering the strengths of searchers' beliefs as well as their topical interests.

3.2 Data

Performing our study required data from a variety of sources, including ground truth, search engine results, and labels for the content retrieved by the search engine.

3.2.1 Ground Truth

We focus on medical findings about the efficacy of treatments or other actions to improve health or alter the course of a condition. We sought a source of information that could serve as ground truth to balance the outcomes for questions used in our study. We did this to help ensure that there was no bias in the answer distribution for the underlying tasks. Accuracy is especially important in this context since the results retrieved for health searches can inform decisions regarding self-treatment and the pursuit of professional medical care [White and Horvitz 2009; Fox and Duggan 2013].

We use information contained in topics covered by Cochrane reviews as a proxy for ground truth about the efficacy of medical interventions. The Cochrane corpus contains systematic reviews of the efficacy of interventions authored by panels of experts. Cochrane reviews are recognized as the highest standard in evidence-based health care [Higgins 2008]. They are used by physicians and healthcare practitioners throughout the world in making evidence-based treatment decisions [Sackett et al. 1996]. Cochrane reports have been found to be more recent and rigorous than systematic reviews and meta-analyses published in paper-based journals [Jadad et al. 1998] or industry reviews, such as those involving pharmaceuticals [Jørgensen et al. 2006]. Direct analysis of the review quality in comparison to other systematic reviews has shown that Cochrane reviews are of superior quality to other sources [Petticrew et al. 2002].

Cochrane reviews investigate the influence of interventions for prevention, treatment, and rehabilitation. They also assess the accuracy of diagnostic tests for a given condition in specific patient groups and settings. Each review addresses a clearly formulated question, e.g., "Can melatonin prevent or treat jet lag?" During creation of the review, a corpus of existing primary research on a topic that meets certain criteria is collated, and then it is assessed by a panel of medical experts using a set of guidelines to establish whether or not there is conclusive evidence about a specific treatment. The reviews are updated regularly to ensure that treatment decisions are based on up-to-date and reliable evidence. Abstracts of the reviews are available on the Cochrane library website (cochrane.org/cochrane-reviews). These comprise a number of sections, including title, background, objectives, methods, results, conclusions, and a plain language summary. Figure 1 presents fields from an example review abstract.

As part of a previous study [White and Hassan 2014], we obtained 4906 abstracts for Cochrane reviews for research purposes. The reviews discuss a range of treatment options, with titles including *Exercise for depression*, *Topical treatments for fungal infections of the skin and nails of the foot*, and *Cranberries for treating urinary tract infections*. We joined the content of these reviews against the queries appearing in search logs using a multi-step matching methodology comprising: (i) computing overlap with Cochrane review titles, (ii) controlling for the term order to focus on queries a similar intent to the review (i.e., query and review mentioned concepts in the same order), and (iii) verifying via human annotators that the selected queries have the same intent as the Cochrane reviews (see Section 3.2.2.2 for more details). Using this method, the above three reviews found matches with logged queries. However, many of the reviews were highly specific, focusing on detailed treatment options (e.g., one has the title "Hypertonic saline solution administered via nebulizer for acute bronchiolitis in infants"), and sufficient matches with logged queries could not be found, even after replacing some complex terminology with simpler variants. Non-matching reviews were ignored in our analysis. For queries corresponding to matching reviews, we obtained their results (both SERP captions and the content of each search result) from the Microsoft Bing search engine, and search interactions from the logs of the same search engine. As mentioned earlier, the Cochrane reviews provided us with ground truth on which to base the selection of a balanced set of search tasks employed in our user study.

3.2.2 Question Queries

We now describe the question queries that we used for our analysis. The specific queries selected were used as the basis for the search results assigned to the Cochrane reviews used in this analysis. Queries were mined from the aggregated search logs of consenting users of the popular Microsoft Bing Web search engine. The queries were selected as part of earlier work [White and Hassan 2014], but we include a description of the methods below for completeness in this article.

3.2.2.1 Search Engine Query Logs

We automatically extracted question queries from a random sample of the logs of queries issued by over 10 million consenting users of the Microsoft Bing search engine during a three-month period from July to September 2013. The data includes user identifiers, timestamps, queries, result clicks, and the captions (titles, snippets, and URLs) of each of the top 10 results. To remove variability from cultural and linguistic variations in search behavior, we only include log entries from searchers in the English-speaking United States locale. Given these logs, we sought to extract queries where the intent appeared to suggest that the searcher was seeking information about the efficacy of a medical intervention. To be more certain that queries were associated with such an intent, we targeted cases where we observed searchers constructing queries as questions. Questions started with words such as "can", "should", "does" and had significant overlap with the Cochrane reviews. To help ensure data quality, we performed the following additional filtering: (i) selected queries issued by at least five users, (ii) selected SERPs with same 10 results/captions and same result ordering across all instances of the query in the three-month period, and; (iii) focused on query instances that were either the only query in the session or the terminal query in the

Table I. Sample queries in question form identified via the filtering process, per truth label.

Helps	Inconclusive	Does not help
does echinacea help colds	can ear drops remove wax	can probiotics help colitis
can caffeine help asthma	does yoga help epilepsy	does ginkgo biloba help tinnitus
can acupuncture help migraines	does methadone help pain	do antibiotics help with colds
does melatonin work for jet lag	do orthodontics help tmj	do steroids help neuropathy
does zinc help colds	can haldol be used for vomiting	can magnesium stop cramps

session with no preceding queries with term overlap. Queries were normalized with transformation to lower case with surplus whitespace and punctuation removed.

3.2.2.2 Mapping Question Queries to Reviews

We mapped the question queries from the search logs to the matching Cochrane reviews to obtain the answer considered as ground truth for each question query. These data are used to select tasks for our study (balanced in terms of outcome) and also in terms of accuracy calculations for our analysis of overconfidence. We did the following:

—**Overlap with titles:** The titles of Cochrane reviews are observed to follow the template: <intervention> for <condition>. To match searcher' questions with a particular review, we required that both the intervention and the condition appear in the candidate query. To improve coverage, we used synonyms for both the intervention and the condition sourced from the Unified Medical Language System (UMLS) [Lindberg et al. 1993]. The UMLS is a well-known medical repository comprising over 60 families of biomedical vocabularies, and maintained by the United States National Library of Medicine. It integrates over two million names for 900,000 health-related concepts. For each of the matching concepts appearing in search queries, we generated a variant of that query for each of the synonyms in the UMLS.

—**Sequence of terms:** To avoid cases where terms overlap, but the order of query terms implies a different intent (e.g., [does the common cold increase zinc levels] matching against the review title "Zinc for the common cold") we imposed an order constraint on the terms in the queries. Specifically, we required that the intervention preceded the condition in the candidate query. Applying this filter meant that we could miss queries that did not satisfy the sequence ordering constraint. However, the query logs were voluminous and we could still find a sufficient number of matches for our study using this precision-oriented approach.

—**Verification with human judges:** The previous two steps were automated to handle large volumes of queries. The workflow generated a filtered set of 2495 distinct queries that was small enough to verify manually. To ensure that the queries we selected were high quality, we created a human judgment task. Crowdworkers from Clickworker.com (provided under contract to Microsoft Corporation) were used to verify that the candidate query matched the intent expressed in the Cochrane review. Clickworker provides rapid access to a large pool of human judges for crowdsourcing tasks. Three judges were provided with the query and the title and background of the review (see Figure 1 from earlier for an example of these fields). Judges were asked to indicate on a three-point scale—*yes*, *somewhat*, and *no*—whether the query had the same intent as the Cochrane review. Each query was reviewed by at least two judges and up to three to obtain a simple majority. We only retained queries where the majority opinion was *yes*.

In total, 268 Cochrane reviews were matched with this methodology, with 1342 distinct matching queries, and tens of thousands of query instances in our log data. The sequence of steps pruned our data significantly, but ensured that high quality queries

Table II. Ratings assigned to results at various rank positions. Ratings range from 0-100 inclusive. Ratings were provided by third party judges. Rank describes the rank position in the result list at which the average result ratings were computed. All ratings at the rank position and above are used. For example, Rank=3 means that the ratings for the top three results (ranks 1,2, and 3) were averaged in computing the result ratings reported for Rank = 3 in the table.

Ground truth	*Result rating*	*N*	*Rank*
All	63.082	585	1
All	63.611	1755	3
All	63.661	4680	8

Table III. Result ratings grouped by the ground truth label assigned to the original question. Ratings range from 0-100 inclusive. Ratings were provided by third party judges. Rank is defined as in Table II.

Ground truth	*Result rating*	*N*	*Rank*
Does not help	58.970	194	1
Inconclusive	61.259	196	
Helps	68.314	195	
Does not help	57.021	582	3
Inconclusive	65.458	587	
Helps	67.924	586	
Does not help	57.077	1556	8
Inconclusive	65.511	1564	
Helps	67.934	1560	

were chosen, e.g., [do probiotics help colitis]. Table I lists a random sample of five queries generated via this approach for each of the three types of ground truth labels described in the next section: *helps*, *inconclusive*, *does not help*. We can see that the queries were similar in each group, even though the answers to the questions were different. Before analyzing how search engines handle these queries, we needed a clear label for the recommendation given by the Cochrane reviews.

3.2.2.3 Labeling Review Recommendations

Labeling review recommendations involved reading the Cochrane summary and assigning a label. This would have been a challenging task to crowdsource since it would require careful reading of the task description and consistent labeling across all 268 reviews. To address this concern, the authors of a previous study [White and Hassan 2014] performed this task.[1] Each of the two authors of that publication reviewed the titles and plain language summary portion of each of the reviews independently and discussed disagreements, amending a small number of judgments in light of these discussions. Answers were provided on a three-point scale: *helps*, *inconclusive*, and *does not help*. The exact agreement between the judges was high (97.4%, free-marginal κ=0.959).

Overall, 45.5% of the matching reviews were labeled *helps* (around half had *might help*), 26.9% *does not help* (around half had *might not help*), and 25.0% *inconclusive*. We used only reviews with agreement and ignored those remaining (2.6%). This label

[1] The recommendations of the Cochrane review panel were clearly stated in the summaries used for this labeling task. Given the high agreement between the two judges (97.4%, κ=0.959), and that we only used review recommendations where both judges agreed on the recommendation, it is unlikely that author labeling resulted any bias that significantly affected the reviews chosen or the outcomes of the current study.

distribution provides our base rates, but to afford us more control over the study, we downsampled reviews and queries to provide an equal distribution of the outcomes.

3.2.2.4 Downsampling Reviews and Queries

To be sure that the conclusions reached from our analysis were easily interpretable and reliable, e.g., not affected by skew in the task outcomes, we sought to create a balanced set of answers for the three outcomes. Given that inconclusive set was the minority class (with 67 instances), we randomly downsampled the *helps* and *does not help* classes such that there were 67 reviews for each outcome; 201 reviews in total and 33.3% of each. We use this set of reviews in the remainder of our analysis. To prevent query-related bias toward particular outcomes, we also randomly sampled the queries within each of the categories so that there was an equal number of queries and a similar distribution of query terms (including question prefixes) for each of three answers. During this process, one question query was randomly selected per each Cochrane review. The downsampling also addressed concerns about selection bias connected to intervention intent (e.g., people may be more likely to search about helpful interventions), review authorship, or the query-review join.

3.2.3 Page Labels

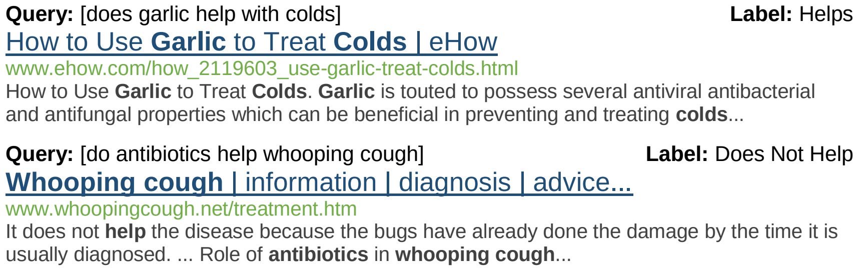
Given that we had access to the results returned by the Bing search system for queries, we sought to understand the nature of the answers contained within those pages. Although methods exist for extracting answers from documents automatically, e.g., [Abney et al. 2000; Dumais et al. 2002], we were concerned about the reliability of such methods given page complexity. We created a human-intelligence task with multiple human judges per page to address noise in the judgments. Crowdsourced judges were provided to Microsoft Corporation under contract by Clickworker.com. Participants were based in the United States and were required to be fluent in English. They were compensated financially for each judgment that they provided. To avoid skewing the page labels toward any one judge, we imposed a limit of 100 labels per judge.

The page labeling task was recognition oriented and presented judges with a query about the efficacy of a medical intervention related to a medical condition (e.g., [does echinacea cure colds]), a Web page, and the opportunity to provide an answer rating. Specifically, judges were instructed to review the full page and do the following:

"*Use the content of the page below to rate the likelihood (from 0 to 100% chance) that the treatment will effectively address the condition*".

Judges were provided with two additional options: (i) **No answer**: the page shared terms with the question query but did not offer an answer, and (ii) **Error**: the judge encountered trouble in loading the page. We solicited judgments from crowdworkers for all pages in the top ten results. Although our judges may be affected by biases in a similar way to our study participants, they were assigned results and could not select particular results as was the case with our study participants; this reduced the impact of selective exposure to attitude-supporting information. Pooling ratings across multiple judges also helped to ameliorate some of the effects of individual judge biases on the page labels obtained from this procedure.

For each query-URL pair we obtained labels from three judges and averaged their ratings to obtain a single rating per pair. Page access errors were encountered for around 3% of pages. These pages were ignored in our analysis. To ensure that judges make an effort to complete the task, we requested that they also indicate the content from the page that provided the strongest evidence for their assessment. Table I shows



Query: [does garlic help with colds] **Label:** Helps

How to Use **Garlic** to Treat **Colds** | eHow

www.ehow.com/how_2119603_use-garlic-treat-colds.html

How to Use **Garlic** to Treat **Colds**. **Garlic** is touted to possess several antiviral antibacterial and antifungal properties which can be beneficial in preventing and treating **colds**...

Query: [do antibiotics help whooping cough] **Label:** Does Not Help

Whooping cough | information | diagnosis | advice...

www.whoopingcough.net/treatment.htm

It does not **help** the disease because the bugs have already done the damage by the time it is usually diagnosed. ... Role of **antibiotics** in **whooping cough**...

Figure 2. Sample captions that were assigned the label *helps* and *does not help* per the definitions introduced earlier.

the distribution of page ratings both overall and at different rank positions. The findings show an overall skew in the results toward *helps* (page ratings >> 50) at all ranks (all $t(584\,|\,1754\,|\,4679) \geq 3.20$, $p < 0.001$), as well as a general increase in the result ratings as the ground truth transitions from *does not help*, through *inconclusive*, to *helps* ($r = 0.7841$, $p < 0.001$). We retained the skewed result distribution for most of the analysis in this article since it accurately reflects the circumstances under which searchers must pursue these and similar objectives on search engines. Also note that over all of the SERPs used in this study, on average there were almost four pages (3.78) that received an answer rating of below 30 or above 70 (with general (73.2%) agreement between the search-result labels for the 3-4 results, for each query). The volume of strongly-labeled data suggests that there was sufficient rational grounds on each SERP for participants to revise their overall beliefs. For completeness, we also experimented with a controlled setting where we balanced the quantity and ranking of answer pages in the results. We describe that setting in Section 3.4.2.

3.2.4 Caption Labels

To more fully understand participant's examination of result lists provided by the search engine, we need to consider the content of the captions presented. Captions can significantly impact result examination behavior [Clarke et al. 2007; Yue et al. 2010; White and Horvitz 2013]. To this end, we performed labeling of the captions using the same definitions as were available for the pages (minus the Error option since all captions were present). We also recruited judges via crowdsourcing to provide answer labels for captions. Judges were sourced from the same judge pool as was used to obtain the page labels, but different judges were employed than those used in the generation of the page labels. Given a question query, judges were asked to label the content of the caption using the same rating scale as used for the pages. Figure 2 provides sample captions labeled as *helps* and *does not help* by judges. The statistics for the caption judging were similar to that of the page judging, and followed a similar trend with variations in result accuracy, so we shall not report on them explicitly in this article.

3.2.5 Summary

We have described the process by which queries to generate results were chosen, and how labels were generated for answers, captions, and results used in our analysis. The dataset allows for a rich analysis of different aspects of the search process, including the nature of the results that are returned to searchers for review, and the types of captions that are selected by searchers as a function of pre-search beliefs. Before proceeding, we describe the design of the assessment interface used by study participants.

3.3 Assessment Interface

We now describe the belief assessment interface through which we presented question queries and associated results, and also captured participant beliefs. It is worth noting that our participants were not self-motivated to perform the medical searches, in the same way as patients or concerned searchers might be. We wanted participation to be straightforward, while also ensuring that we could clearly delineate the aspects of the search process that were of most interest given our research questions. As such, we designed an interface that comprised three phases: *pre-search, during*, and *post-search*. Each phase is described in more detail below. Appendices A1-A3 presents the interface shown at each of the three phases of the judgment process.

—**Phase 1. Pre-search:** The participant is presented with the title of the Cochrane review (e.g., "Acupuncture for insomnia") and was asked to provide a probability that the described treatment is effective (on a scale ranging from 0-100). They were asked to base their judgment only on their personal knowledge, beliefs, or experience and were explicitly requested not to search the Web. We refer to this as their *pre-search belief* rating. In addition to this rating, they also provided their level of confidence in the assessed probability (on an 11-point scale, from 0-10 inclusive). Appendix A1 shows a screenshot of the interface used for this phase.

—**Phase 2. During:** The participant is then presented with search results from the search engine for the query that maps to the Cochrane review per the process described earlier. Participants can select as many results as they would like from the ranked list to determine the efficacy of the treatment (similar to a real information-seeking task). For each result selected, a new tab containing the content appears, occluding the result list. When the searcher has completed their examination of the result content, they return to the original tab to continue examining the search results page. When they return to the results page, they are presented with an assessment popup that seeks changes in the participant's likelihood and confidence in the efficacy of treatment following examination of the search result. Appendix A2 presents (i) the search-result examination interface, and (ii) the popup shown to participants once they select and review a result, and return to the SERP. In the example displayed in Appendix A2b, the participant updates their belief (from 50 to 70), and their confidence (from 6 to 7) in light of their interpretation of the information presented in the result examined (and possibly other factors such as a synthesis of the information encountered in the search episode this far (discussed more in Section 4.6)).

—**Phase 3. Post-search:** The participant is asked to provide information on their posterior belief and the associated confidence level following their examination of the result list and any results that they have selected during the process (Appendix A3).

The interface clears the ratings textboxes prior to each phase. As a result, participants need to re-enter their belief rating at each phase. This was done intentionally so that participants had to reconsider their belief and confidence ratings periodically as they moved through the search task. Participants did not have to complete all search tasks in a single pass. Tasks were assigned from the pool of available tasks on a first come, first served basis. Participants could therefore stop and return later, as long there were still tasks available for them to complete. Note that once a participant started on a particular task, they needed to complete all three of the phases outlined above for their judgments to be recorded. They could not terminate a given task midway and expect to resume that task at a later date.

3.3.1 Logging

All of the ratings provided by participants are recorded by the system in addition to a timestamp for each action that was performed. We also recorded several aspects of participant interaction, including the clicks on results and the dwell time on those results. The logged times capture the amount of time that participants spend with the result content in focus, on the SERP, and the total time spent per task.

3.4 Measuring Belief Dynamics

We employed two experimental settings in the measurement of belief dynamics: (i) a *natural* setting that presented the results retrieved by the search engine in their original rank order, and (ii) a *controlled* setting, where we controlled the availability and rank ordering of the search results shown. To avoid learning effects, each setting employed different sets of experimental participants.

3.4.1 Natural Setting

We used the questions generated per the description earlier. There were a total of 201 questions, 67 of each of the three answer types described above. These were randomly assigned to participants, with one query per review similar to those in Table I selected to generate the SERP shown to searchers. The belief dynamics for each review were captured from 10 judges. We could not compel the remote participants to complete all questions assigned. Instead, participants completed as many tasks as they desired. We imposed an upper limit of 100 tasks per participant, to ensure that we did not receive too many judgments from a single participant (which could skew our results toward particular participants' ratings). For the data analyzed in this study, there were 85 participants, who each completed on average 30.2 search tasks.

For all of the analyses, we removed tasks that judges took less than ten seconds to complete. Given the complexity of the judgment task, we believed that it was not possible to complete the task in good faith given such a short time. This resulted in an exclusion of 5.5% of our tasks which had these short completion times, and the complete exclusion of four judges from the analysis, who may have been spammers. The analysis of the natural setting data are from judgments and behaviors for the remaining 94.5% of tasks (*n*=1890).

3.4.2 Controlled Setting

As mentioned earlier, one of the advantages of using a crowdsourcing platform for the judges and participants in our experiments is that additional experiments with task variations can be run with ease. The experiment described in the previous section examined search results in a natural setting. In that setting, results returned by the search engine exhibited a slight bias towards *helps*, as evidenced by the result ratings reported in Table II. White and Hassan [2014] identified the reasons behind this bias, which include terms in the query (e.g., including the query term “help” leads to more results suggesting intervention effectiveness), the use of aggregated searcher behavior, and the availability of content in the search engine index (i.e., more pages with *helps* content are present in the index and available to be ranked by the search engine). More broadly, White and Horvitz [2009] examined distributions of online medical content independent of search queries, indexed materials, and interventions. They showed that such distributions of online content, and links between that content and medical symptoms, often diverge from distributions that are representative of prior and posterior probabilities of medical disorders (as drawn from the medical literature in their study). Understanding search behaviors in light of biases in online content and search results is important in learning about search engine usage in natural settings. However, the biased result list could skew results pertaining to engagement with SERPs. As such, we also pursued a strategy of controlling for the rank and availability of content related to the main outcomes. To do this, we created a judgment task whereby we first collected judgments from crowdsourced third-party judges (also drawn from a pool provided by

Clickworker.com) for pages that explicitly discussed the effectiveness of the treatment options. We then created ranked lists using these pages comprising an equal number of pages with *helps* and *does not help* labels, arranged in a random order.[2] This allowed us to study belief updating in a setting unaffected by biases in the quantity and rank ordering of the search results returned to searchers.

In this part of the analysis, we created a separate crowdsourced judgment task. Judges were asked to use the Web search engine of their choosing to find the URLs of five pages that only discussed the effectiveness of the intervention and in a separate task judges were asked to find the URLs of five pages that only discussed the ineffectiveness of the intervention. Judges also provided an indication of which of the five URLs that they found had the strongest evidence. For this analysis, we used a set of 40 tasks: 20 where the Cochrane ground truth label was *helps* and 20 where the label was *does not help*. Ten judges were assigned per task, resulting in a maximum of 100 URLs per topic (50 *helps* and 50 *does not help*). However, there was some overlap in the URLs found, on average 28.2% of the URLs located for the task were found by multiple judges. The five most popular URLs for each label type were used to generate a SERP for each intervention, randomly sorted. Ties were broken via secondary ordering using the number of judges who rated a URL as providing the strongest evidence. Across all SERPs, the average ratings for the *helps* pages was 82.64 (standard deviation, SD=12.33) and for *does not help* pages the average rating was 15.21 (SD=10.55). This suggests that, as with the natural setting, there is sufficient information available in the result lists to provide rational grounds for participants to revise their beliefs.

Titles and captions for these URLs were obtained from the Microsoft Bing search engine. To create the query-based captions we used the logged query that was most similar to the title of the associated Cochrane Review (based on the token-based edit distance between the query and the title, following removal of stop words) that also retrieved the URLs in the top-10. We were then able to extract the corresponding caption for the query-URL pair from our search log data. Pre- and post-search belief and confidence ratings using this artificially-generated list of search results were captured in the same way as described in the previous sections. As part of this experiment, additional judgment tasks (using the approach described in Section 3.2.4) were run to collect ratings for the snippets and search results identified using this method.

3.5 Predicting Belief Dynamics

The analysis so far has focused on characterizing the nature of any changes in belief over the course of the assigned search tasks. It is unlikely that a search system would have direct access to the assessed beliefs and confidences of searchers. Requiring that searchers report their belief as they perform tasks would likely be intolerably costly and could also introduce additional biases or other affects [Nielsen and Levy 1994; Czerwinski et al. 2001]. If we are to harness information about human belief updating to enhance the value of search to users, we may need to make inferences about belief updating. In one step in this direction, we now focus on the feasibility of predicting

[2] We focused on *helps* and *does not help* categories only (excluding *inconclusive*) because these had the strongest influences on beliefs, and we wanted to create balanced SERPs comprising an equal number of results with each rating, ordered randomly.

post-search beliefs and belief revision given attributes that are easily observable. Specifically, we consider whether observable aspects of participants' search activity could be useful in predicting belief updates associated with result examination.

This model could make predictions about the nature of the updates in a searcher's beliefs that could be expected to occur over the course of a session. Such a model could be used to update a long-term model of searcher beliefs that could be used to better personalize the search experience, in retrospective search log analysis to estimate beliefs and belief revision, and in generating training data for ranking algorithms that can optimize for their likely impact on searcher beliefs in addition to relevance.

We focus on predicting post-search belief given signals from search behavior and the history of ratings from the searcher, including the pre-search rating in some of our experiments. Feature development was performed using the dataset from the natural setting described in the previous section. As a training set, we collected an additional dataset using the natural setting from a separate pool of judges (the total number of tasks, following filtering of noisy search tasks (those with a duration of less than ten seconds) was 1879). In this set, participants only provided pre-search and post-search ratings (i.e., no ratings were collected during the search). Collecting ratings during the search would bias search behavior and we wanted to use signals derived from this behavior as features in our predictive models. As a *test set* to evaluate our predictive models we re-ran our experiment to generate a fresh set of data of equivalent size (*n*=1912) with a separate pool of judges.

Table IV. Features used for classification and regression tasks. Some of the features rely on access to external data sources: *Third-party judges, **Participant self-reports, ***Medical reviews.

Feature	**Feature description**
Search Activity	
NumClicks	Number of pages viewed
Num<Label>Pages	Number of *helps* / *inconclusive* / *does not help* pages viewed*
NumUniqueDomains	Number of unique domain(s) viewed
NumUniqueTLDs	Number of unique top-level domains (e.g., .edu, .gov) viewed
ClickPosition	Average click position(s) in ranked list of results
TimeOnTask	Total time on task
TimeOnSERP	Total time on SERPs
TimeOnPages	{Total | Average} dwell time on pages
ClickedResultRating	{Minimum | Average | Maximum} rating of clicked result(s)*
SeenExplanationOnPage	Was page rating explanation seen (using reading time)?*
Result and Captions	
ExplanationPagePosition	Position of page-rating explanation (% of length)*
AvgResultRating	Average page rating of the results shown on the SERP*
AvgCaptionRating	Average caption rating of the results shown on the SERP*
AvgContentLength	Average content length
User History	
UserNumTasks	Number of tasks completed
UserAvgChange	Average historic belief change (if available)**
UserAvgPreSearch	Average historic pre-search belief (if available)**
UserAvgPreConfidence	Average historic pre-search confidence (if available)**
UserAvgPostSearch	Average historic post-search belief (if available)**
UserAvgPostConfidence	Average historic post-search confidence (if available)**
UserAvgAccuracy	Average historic accuracy (if available) (rating + Cochrane)**
UserActivityContent	Average historic values for search activity and the content accessed prior to the current task (if available)*
Question	
QueryStartsWith<Term>	Query starts with one of {can, do, does, should, will}
GroundTruth	Ground truth label on intervention efficacy (Cochrane)***
Task	
PreSearchRating	Pre-search belief rating**
PreSearchConfidence	Pre-search confidence rating**

3.5.1 Features

We experimented with several features for predicting belief updates. The features can be grouped into five classes: (i) search activity—features of the search behavior for the current user for their current search task; (ii) content of the SERP and the pages selected; (iii) question—features of the query and the ground truth (answer from Cochrane); (iv) user history—features of the user history for their previous search tasks, focused on their self-reported ratings of confidence and belief, and the third-party ratings of the content accessed, and; (v) task—the self-reported pre-search belief and the pre-search confidence ratings. Table IV presents a description of the features used in our experiments. The features were chosen based on the data that we could access via our remote experimental methodology and ratings from self-reports. Many of these features are explained by descriptions in the table. We provide additional discussion on others. *ExplanationPagePosition* describes the relative position in the page of the strongest evidence that contributed to the page rating assigned by third party judges (averaged across all judges, per page-topic pair). Based on the position of that explanation on the page, *SeenExplanationOnPage* uses an estimate of human reading speed, approximately five words per second [Ziefle 1998], to estimate whether the searcher read the explanation (as a binary feature). This feature captures some of the interaction between explanation position and dwell time, which may make it a useful estimator of whether content was actually examined. Features that capture the starting word of the query (e.g., *QueryStartsWithCan*) are based on previous work, which has shown that these terms can be important in determining the nature of the search results [White and Hassan 2014], e.g., the query term “can” denotes possibility and is likely to retrieve results which favor intervention efficacy. For each instance, the user history class comprises the feature values averaged over all tasks for that participant observed up until that point (if there are any preceding tasks). This is used as a proxy for longitudinal information about the searcher from their search history (that can be used for applications such as personalization [Teevan et al. 2005; Bennett et al. 2012]), which may be unavailable if it is the first query or search task observed from that searcher.

Some of the features rely on having access to data from third-party judges or self-reporting by participants. In some cases, the human labels serve as a proxy for other automated methods for detecting answers in content (e.g., the presence of answers on pages (Abney et al. 2000)) and in assigning the ground truth label. In others, the labels reflect participants own belief and confidence at a particular point. While it is unlikely that searchers would provide explicit ratings about their pre-search beliefs in the course of performing search for information, search systems may be able to infer these beliefs based on the nature of the content that searchers visit and their behaviors as they interact with search results (e.g., the duration of dwells on pages expressing clear viewpoints on topics of interest).

3.5.2 Prediction Tasks

We focused on the following two prediction tasks:

—**Predict post-search rating bucket:** Multi-class classification task, i.e., one of the following three outcomes: helps, inconclusive, and does not help.

—**Predict exact post-search rating:** Regression task, where the goal is to predict the exact post-search rating (in the range [0,100]). This is more challenging than assigning the rating to a particular bucket, but this also allows for more accurate estimation of belief ratings—meaning that subtle belief changes can be detected.

For all of these tasks, we employed ten-fold cross validation and report performance averaged across ten random experimental runs. In running our experiments, we stratified the folds in the cross validation by participant, allowing us to determine the performance of our methods in predicting post-search belief ratings for new (unseen) users. We experiment with three baselines: (i) *random*: generate a random class (classification) or a random rating (regression), (ii) *marginal*: always predicts the dominant class (*helps*) for classification or the most popular specific rating (i.e., 100) for regression, and (iii) *confidence*: always updates the initial belief by the typical belief revision associated with the initial confidence. To address issues with data sparseness in the *confidence* baseline model, we bucket the initial confidence values in our training set into 11 groups (i.e., 0,10,20,…,90,100). For each of these groups, we then compute the average observed belief update and use that as the update in our experiment. When the model is applied as a baseline in our study, the confidence value of the searcher is first bucketed and the assigned value for that bucket is used for the belief update in the model. Note that the *confidence* baseline is only applied in comparison against the full model, which has access to all features—including self-report features on pre- and post-search confidence and beliefs.

3.5.3 Models and Metrics

For each of our prediction tasks, we used Multiple Additive Regression Trees (MART) [Friedman et al. 2000] to train regression and classification models. MART employs gradient tree boosting methods for regression and classification purposes. In MART, classification is performed by thresholding the output of the regressor. Advantages of MART include model interpretability (e.g., a ranked list of important features is generated to assist in better understanding the model), facility for rapid training and testing, and robustness against noisy labels and missing values.

We measure the performance of our predictive models using different metrics: accuracy and area under the receiver operator characteristic curve (AUC) for the classification task, and mean absolute error (MAE) and normalized root mean square error (NRMSE) (defined as root mean squared error (RMSE) divided by the range of observed values, i.e.,100) for the regression task. To compute the AUC for the multi-class classification we use the method from Hand and Till [2001], which computes an unweighted average over a set of binary comparisons between the different label options. These metrics allow us to quantify the effectiveness of our predictive models in estimating the post-search belief at coarse and granular levels. For the regression task, we focus on MAE and RSME (lower is better), and for the classification task we focus on accuracy and AUC (higher is better).

4. FINDINGS

We now present the findings of our study, grouped by research question.

4.1 Belief Dynamics (RQ1)

A central goal of this work is to understand how people's beliefs change during the course of reviewing Web content. Recall that we balanced the ground truth such that there was an equal distribution of each answer type. We asked participants to provide their estimate of the probability that the treatment was effective, between 0 and 100 inclusive (higher means more effective). Despite this balancing, participants' initial (pre-search) beliefs were found to be skewed positively towards *helps*, i.e., average pre-search rating across all of the tasks = 60.70, median = 62). This bias was evident regardless of ground truth (i.e., all per-truth-label median belief ratings ≥ 56). As mentioned earlier, it is unlikely that participants had a strong attachment to the topics discussed in the Cochrane reviews given the range of topics and the fact that they were not self-motivated to pursue the searches. However, participants may have been open to the possibility that the interventions could be effective, resulting in an initial skew toward *helps*. The observed bias toward *helps* may also be founded in part in the topic selection process used by Cochrane, per the organization's work to address common beliefs and questions. More investigations are needed to understand the basis for the biases that we identified about the efficacy of interventions.

Turning our attention to subsequent changes in beliefs, we observe that most beliefs changed (in 66.7% of the search tasks there was some change in the belief, as illustrated in Figure 3) but most changes in belief were slight (in 66.9% of the cases where a change in belief was noted, the absolute rating change was ≤ 20). This may be

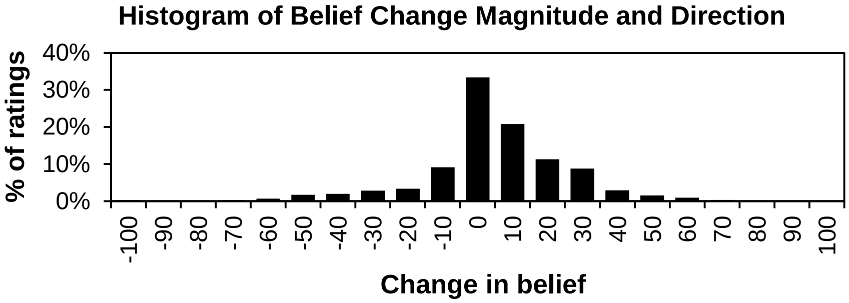
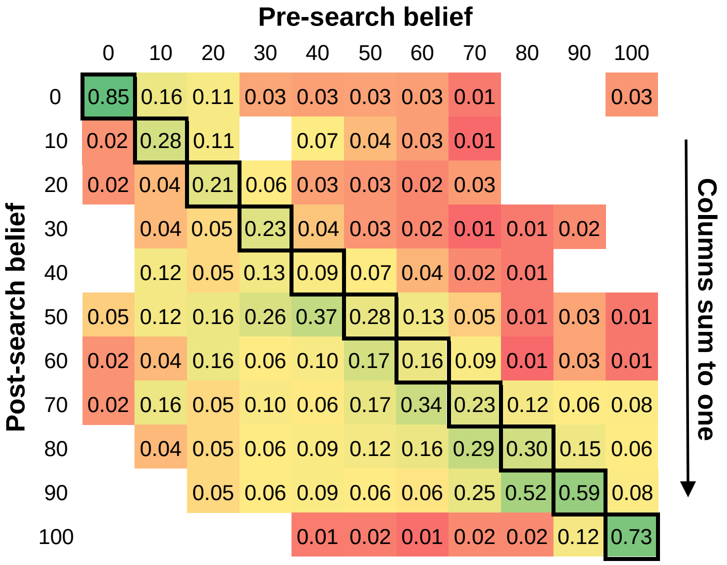


Figure 3. Histogram of nature and direction of belief change during the search task (start vs. end) (*N*=1890).



	0	10	20	30	40	50	60	70	80	90	100
0	0.85	0.16	0.11	0.03	0.03	0.03	0.03	0.01			0.03
10	0.02	0.28	0.11		0.07	0.04	0.03	0.01			
20	0.02	0.04	0.21	0.06	0.03	0.03	0.02	0.03			
30		0.04	0.05	0.23	0.04	0.03	0.02	0.01	0.01	0.02	
40		0.12	0.05	0.13	0.09	0.07	0.04	0.02	0.01		
50	0.05	0.12	0.16	0.26	0.37	0.28	0.13	0.05	0.01	0.03	0.01
60	0.02	0.04	0.16	0.06	0.10	0.17	0.16	0.09	0.01	0.03	0.01
70	0.02	0.16	0.05	0.10	0.06	0.17	0.34	0.23	0.12	0.06	0.08
80		0.04	0.05	0.06	0.09	0.12	0.16	0.29	0.30	0.15	0.06
90			0.05	0.06	0.09	0.06	0.06	0.25	0.52	0.59	0.08
100					0.01	0.02	0.01	0.02	0.02	0.12	0.73

Figure 4. Matrix representing *P*(*post-search belief* | *pre-search belief*). Columns sum to one. Cells on diagonal (dark borders) indicate unchanging beliefs (i.e., pre-search belief = post-search belief) (*N*=1890).

due in part to previously-identified heuristics and biases of judgment, including anchoring and adjustment [Tversky and Kahneman 1974], where people have been found to rarely deviate significantly from an initial judgment.

To simplify the analysis, we bucket the beliefs by deciles and present the distribution of belief changes as shifts among buckets in Figure 3. When belief updates occurred, the movement was largely in a positive direction (53.1% toward *helps* vs. 16.8% toward *does not help*) following review of both SERPs and individual search results. Figure 4 illustrates the transitions in more detail, focused on the conditional probability *P*(*post-search belief* |*pre-search belief*). Cells on the diagonal depict cases where beliefs remain unchanged as a result of searching. From the figure, we see that frequently, beliefs either remain strongly held (upper left or lower right) or move toward *helps* following searching.

We made clicking on search results optional since we wanted to improve the realism of the search tasks assigned to remote participants. Clicks were observed for approximately 40% of search tasks, evenly distributed across participants such that no single searcher or group of searchers was an outlier. On average, there were 1.81 clicks per task and 45% of these tasks had a single click. The 40% clickthrough statistic is similar to clickthrough statistics for informational queries appearing in Web search engine query logs [Teevan et al. 2008]. This is one indication that participants' behavior was similar to that observed in real settings.

We were concerned that the lack of clicks could be caused by spammers, which is a great concern in crowdsourced studies [Raykar and Yu 2011]. Time on task has been used as a way to remove erroneous judgments in crowdsourced experiments [Kazai 2011]. We filtered out those tasks which took less than 10 seconds to complete, which likely identified some portion of crowdworkers who are not attending to tasks. Analysis of the remaining tasks revealed no correlation between whether any result was selected and the total time spent on the search task (the Pearson point-biserial correlation was 0.09, p = 0.79). The lack of correlation increased our confidence that the tasks without clicks were likely performed by non-spammers

Interestingly, search tasks where participants clicked on results were 34% more likely to show changes in beliefs than those tasks where no results were selected (i.e., when participants *only* reviewed SERP content) (***p*** < 0.01). This suggests that the content of the pages has an additive impact on beliefs beyond the effect of the SERP. It may also reveal something about the nature of the beliefs held by the participants who elected to click (e.g., perhaps these beliefs were more malleable) or could be linked to levels of engagement by participants in the task. We discuss this in more detail later in the article. As mentioned earlier, for this (natural) setting, we use the results returned from the engine in the order that they were originally presented, so as to accurately model current the practice in Web search. As we show later (Section 4.4), even when we artificially balance the rank and quantity of evidence, participants were still more likely to seek content that: (i) supported the pre-search beliefs, and (ii) tended to justify the efficacy of interventions.

4.2 Effect of Pre-Search Belief on Search Behavior (RQ2)

In addition to considering the nature of pre-search beliefs and how they change as a result of searching, we considered the influence of the pre-search beliefs on the examination of search results. Information about the examination behavior of searchers can be harnessed to guide the optimization of retrieval performance of search engines [Joachims et al. 2002; Agichtein et al. 2006]. Examination behaviors include selection of results and the review of landing page content measured in terms of dwell time on the page. Dwell times provide a sense for how deeply participants explored the results that they chose to examine. To simplify much of our analysis for the remainder of this article, we group participants' pre-search ratings into three buckets with the following ranges: (i) *does not help*: [0,30); (ii) *inconclusive*: [30,70], and (iii) *helps*: (70,100]. We experimented with other thresholds and discovered that, within reason, the findings were largely insensitive to the specific thresholds selected.

4.2.1 Clicks and Outcomes

Table V presents the clicks and the ratings assigned to the snippets of text drawn from pages that participants selected. As expected, the ratings (assessed during the phase of labeling by crowdworkers) of clicked snippets lies around 60 given that the result ratings in the top-ranked results have a similar value. There is a strong correlation between the pre-search rating and the post-search rating (r = 0.769, p < 0.001). From the table, we can see that beliefs become less extreme as a result of examining search results. One explanation is that searchers become more uncertain as a result of reviewing information and change their position to align with the content that appears in the retrieved Web pages (i.e., a belief rating of around 60). However, our findings on

Table V. Average rating of clicked captions and average post-search belief, given pre-search belief.

Pre-search Bucket	*Average Pre-search Belief*	*Average Clicked Snippet Rating*	*Average Post-search Belief*	*N*	*Percentage of clicks*	
					Helps	*Does not help*
Does not help	13.13	60.66	48.63	274	39.31%	60.69%
Inconclusive	52.33	64.39	65.78	282	61.44%	38.56%
Helps	86.70	68.48	81.07	260	75.27%	24.73%

Table VI. Average dwell time given pre-search belief and result rating.

Pre-search Bucket	*Result rating*	*Dwell time (seconds)*	*N*
Does not help	Does not help	43.02	122
	Inconclusive	38.15	50
	Helps	30.36	79
Inconclusive	Does not help	29.89	80
	Inconclusive	30.05	52
	Helps	36.45	127
Helps	Does not help	28.18	47
	Inconclusive	30.63	48
	Helps	41.83	144

Table VII. Average number of clicks, total dwell time, total SERP time, and total time on task.

Pre-search Bucket	*Num. clicks*	*Total time (seconds)*			*N*
		Dwelling on result	*Dwelling on SERPs*	*On task*	
Does not help	1.80	76.76	26.86	103.62	274
Inconclusive	2.05	65.05	19.20	84.25	282
Helps	1.56	60.51	13.90	74.41	260

confidence presented later (Section 4.5) suggest that confidence generally increases as a result of searching, across the full range of belief ratings. The findings suggest that those who clicked made a concerted effort to revise beliefs to align with available content. In addition, almost all of the search result lists contained snippets and/or results with a third-party answer rating that contradicted the participants' pre-search beliefs. The exposure to alternative perspectives may lead to attitude moderation, similar to that noted in related studies [Liao and Fu 2013].

When we consider the nature of the captions that searchers selected, we observe a strong correlation between pre-search beliefs and the rating of the clicked caption ($r = 0.8112$, $p < 0.001$). This clearly demonstrates the connection between prior beliefs and search behavior and may be explained in part by biases of confirmation and tendencies to anchor on initial beliefs [Tversky and Kahneman 1974]. Studies of selective exposure [Frey 1986] provide evidence that people favor opinion-reinforcing information. Related work on cognitive dissonance [Festinger 1957] has shown that people experience positive feelings when viewing confirmatory information. The use of a human labeling methodology allows for the detection of this bias and its extent to be quantified in a way that was not possible in previous work. Considering the percentage of clicks

Table VIII. Top 3 patterns in belief changes over the course of the task (for tasks with ≥ 3 clicks, *N*=206).

Pre-search Bucket	***Pattern***	***P(Pattern∣Pre)***	***P(Open∣Pre)***	***N***
Does not help	*Dec,Dec,Dec*	0.357	0.164	69
	Inc,Inc,Inc	0.225		
	Same,Same,Same	0.128		
Inconclusive	*Same,Same,Same*	0.279	0.209	85
	Inc,Inc,Inc	0.252		
	Dec,Dec,Dec	0.191		
Helps	*Inc,Inc,Inc*	0.482	0.089	52
	Dec,Dec,Dec	0.123		
	Same,Same,Same	0.080		

on each of the two main outcomes (*helps* and *does not help*, far right columns in Table V), there are clear variations depending on pre-search rating ($\chi^2(2) = 13.93$, $p < 0.001$).

4.2.2 Result Dwell Time

In addition to considering the results that people selected, we also studied the amount of time that they spent on search results as a function of the pre-search rating and the rating assigned to the clicked result. Dwell time is an important signal in a number of applications, including satisfaction estimation [Fox et al. 2005; Kim et al. 2014] or relevance estimation [Yan et al. 2014], so being able to measure the impact (if any) of pre-search rating on page dwell time, has practical significance. The descriptive statistics depicting the dwell time are shown in Table VI. We performed a two-way analysis of variance (ANOVA) with *pre-search rating* and *result rating* as the two factors of interest. The findings reveal some interesting differences: (i) that when the pre-search rating is *does not help* or *helps*, participants spend significantly more time examining results that agree with their pre-search rating (around 40s for agree versus around 30s for disagree) ($\boldsymbol{F}(2,248 \mid 236) \geq 3.97$, $\boldsymbol{p} = 0.02$), and (ii) when the pre-search rating was *inconclusive*, the trend was less clear ($\boldsymbol{F}(2,256) = 1.48$, $\boldsymbol{p} = 0.23$). Searchers spend less time examining content that contradicts their current position; an observation that has been made in other situations beyond information search (e.g., in political science [Garrett 2009; Knobloch-Westerwick and Meng 2009]). The fact that agreement and disagreement led to large changes in dwell time has broad implications for behavioral analysis in retrieval settings. For example, it means that there is value in considering alignment with prior beliefs, in addition to page content [Liu et al. 2010; Kim et al. 2014], when interpreting page dwell times. To this end, later in this article we explore ways to automatically estimate beliefs in light of limited evidence about searchers' intentions via implicit signals mined from their search activity.

4.2.3 Total Time Spent

How people spend their time during searching may reveal their interests and even their beliefs in addition to their focus of attention. We measure temporal duration in a number of different ways in this study: (i) total dwell time on pages, (ii) total time on SERPs, and (iii) total time on task (a combination of (i) and (ii)). We also considered the number of search-result clicks as a function of the pre-search rating. To analyze these differences we employed a one-way multivariate analysis of variance (MANOVA) with the pre-search rating (rows in Table VII) as the independent variable and the number of clicks, total time on SERP, dwelling on pages, and on the task in total (i.e., the columns in Table VII) as the dependent variables. The MANOVA shows that there

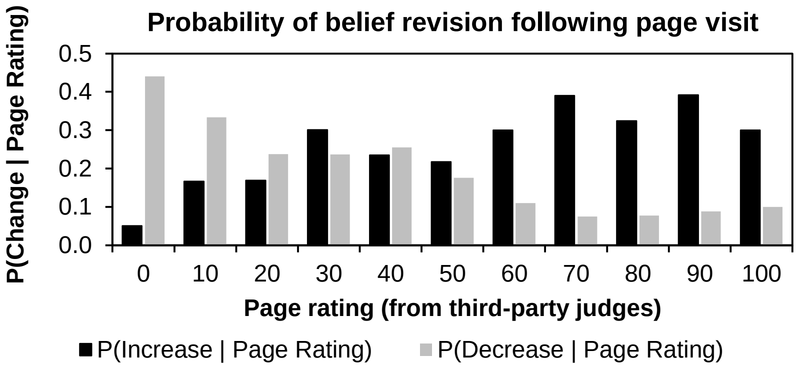


Figure 5. Changes in belief (Increase, Decrease) conditioned on the rating of the page reviewed immediately beforehand. P(Same | Page Rating) is not shown in the figure, but is approximately 50% across the full range of ratings.

Table IX. Relationship between the locations of the strongest supporting evidence in clicked results and changes in reported beliefs.

Location of evidence	**Belief update?**		***N***
	Yes	***No***	
Top half of page	78.57%	21.43%	330
Bottom half of page	56.49%	43.51%	124

Table X. Patterns in belief changes over the course of the task (for tasks with three or more clicks).

	Number of page views			
	0	**1**	**2**	**3+**
Absolute belief change	16.55	19.15	23.42	26.40
N	1194	230	280	206

are significant differences for all variables tested (all $\boldsymbol{F}(2,816) \geq 4.85$, $\boldsymbol{p} < 0.01$). From the statistics summarized in the table, we can draw two conclusions:

—There are more result clicks when the pre-search rating is *inconclusive*, perhaps because searchers need to explore more options before settling on an answer. There were the fewest number of result clicks associated with *helps*. Supporting information in those cases may be easier to find, and significant amounts of information is not required to validate a hypothesis that a particular intervention is effective.

—The total dwell time and total time on task, including SERP examination time, is higher for cases where the pre-rating is *does not help*. This may be related to the difficulty that searchers experience in locating negative information given the noted positive skewness in our Web search results, or the fact that it may be difficult to disconfirm hypotheses.

The findings clearly indicate behavioral differences associated with the pre-search rating. The significance of these differences suggests that there may be features of the pre-search behavior that are useful indicators of searchers' pre-search beliefs. This may be useful for predicting pre-search beliefs or to inform better models of post-search beliefs (since beliefs frequently do not change dramatically (per Figure 3)).

4.3 Changes in Belief during Searching (RQ3)

Our next research question concerns changes in beliefs over the course of the search session. To understand how beliefs change *during* search episodes, we also considered changes in belief related to the examination of results. We examined this in three ways: (i) patterns of belief change over the course of the task, (ii) changes in belief conditioned on the rating assigned to the page viewed (and the location of evidence therein), and (iii) the role of the quantity of information viewed in the updating of searcher beliefs.

4.3.1 Patterns of Belief Change

To estimate the degree of dynamism in information needs, we calculated changes in the belief over the current state following reviewing a search result, grouped based on the pre-rating in the same way as in the previous analysis. Changes were labeled *Inc* when there was an increase in belief rating about treatment efficacy (toward *helps*), *Dec* when there was a decrease (toward *does not help*), and *Same* when there was no change. Focusing on cases where there were at least three clicks (n=206), we can create patterns of change over the course of the task (e.g., *Inc,Inc,Inc* representing three consecutive increases in belief rating following the review of each of the three search results). Focusing on episodes with at least three clicks provided the opportunity to observe oscillation in the belief ratings of searchers, which would not be possible with one or two clicks.

Table VIII presents the dominant patterns within each bucket. The results show that three consecutive changes of the same type are most common. They also show that consecutive increases in belief are highly likely for those who believed *does not help* or *helps* before examining results (i.e., ***P****(Pattern | Pre)* = 0.357 and 0.482 respectively, where *Pre* denotes the pre-search belief (*helps*, *inconclusive*, *does not help*)). When we consider cases where participants were unsure at the outset of their searching, we find that the probabilities are more evenly distributed. Since these participants approached the task with a less strongly-held opinion, they may be more amenable to change.

To better understand changes within the search task, we also computed the probability that the participant exhibited signs of open-mindedness, ***P***(*Open* | *Pre*), to reflect the likelihood that beliefs would both increase and decrease within the same task as participants examined the results. Although this may simplify how people weigh and consider evidence, our methodology provided a unique opportunity to study this aspect of belief dynamics within the search process. To satisfy this definition, searchers had to both increase and decrease within three consecutive clicks (e.g., *Inc,Same,Dec* or *Inc,Dec,Inc*). The amount of information that people need to view for their beliefs to change may vary on an individual basis. Searchers may also consider additional factors such as source credibility and authority during belief revision [Wathen and Burkell 2002]. As a result, our analysis only focuses on a subset of the participants whose open-mindedness manifests in a particular way: oscillating between different outcomes. The results are summarized in the last column of Table VIII. They show that participants who were unsure at the outset of the task (i.e., *Pre* = *inconclusive*) were more open-minded per our definition than those who had formed an opinion before they searched. Interestingly, those who decided that a treatment was effective at the outset were almost half as likely to be open-minded (***P***(*Open* | *helps*) = 0.089), as those who believed beforehand that the intervention was ineffective (***P***(*Open* | *does not help*) = 0.164).

4.3.2 Reasons for Belief Change

To better understand the reasons for observed belief dynamics, we examined two factors: (i) the content of the pages that people viewed, and (ii) the collective volume of the information that people viewed, regardless of the answers contained therein.

4.3.2.1 Page Content

Given that we obtained ratings for each of the pages in our dataset, as well as labels from searchers with regard to how their ratings changed following reviewing the page, we can compute the effect of the page on the belief. In particular, we study whether that effect is positive (toward *helps*) or negative (toward *does not help*). At each value of page rating we can compute the fraction of changes that were positive and the fraction that were negative. Figure 5 shows that the nature of the page is associated with the type of change observed. Pages labeled as *does not help* often lead to changes in the negative direction (toward zero), although mainly for the extreme cases where the page rating ≤ 20. For most cases where the page ratings were 50 or more, people were much more likely to transition to *helps*. It appears that those who may have been unsure initially need only to see some supportive evidence to shift their belief positively.

4.3.2.1.1 Impact of Explanation Position

We also examined the relationship between *where* the content appeared on the page and whether there was a change in belief following the review of the content. White and Horvitz [2010] showed that the presence of serious illness mentions before benign explanations was more likely to lead to escalations in subsequent search queries. In this analysis, we were interested in whether there was a relationship between where on the page the strongest evidence for a page label was found (provided by third-party judges) and whether the searcher's belief was likely to change. To do this, we focused on pages that expressed either *helps* (rating > 70) or *does not help* (rating < 30) per the definitions introduced earlier. We divided the page into two parts: top 50% and bottom 50%, and computed how frequently beliefs changed *toward the overall page rating* when the strongest evidence appeared in either of those two locations.

The results in Table IX show that evidence appearing toward the top of the landing page is much more likely to result in belief updates when the result is viewed than that appearing toward the bottom ($\chi^2(1) = 11.031$, $p < 0.001$). Moving beyond the two levels, the correlation between the position of supporting evidence the result (expressed as a fraction of content following the removal of HTML markup) and the belief change was also strongly negative ($r_{pb} = -0.74$, $p < 0.001$), signaling a significant relationship between the position of information and belief updating. Explanations for this include differences in the amount of attention that the different parts of pages receive, as has been shown extensively in research on gaze tracking (e.g., [Buscher et al. 2008; Guo and Agichtein 2012]), the nature of the information that appears early in Web pages generally, and the interactions between these two factors.

Deliberate editorial decisions made by page authors could result in less compelling content being presented later in the page—an approach used often in the news media [Bell 1991]. To better understand the positioning of strong evidence on Web pages, we performed additional analysis. We converted the page ratings data from third-party judges into a binary judgment whereby 0 = weak or no evidence (i.e., *inconclusive* per the earlier definition) and 1 = strong evidence (i.e., *helps* or *does not help*). We then computed the correlation between the presence of strong evidence (0 or 1) and the position of the answer on the page (as a percentage of page content length, minus HTML and scripts as before). The findings suggest that stronger evidence was indeed more likely to be positioned near to the beginning of the page ($r_{pb} = -0.48$, $p < 0.001$).

Given the change in belief as a function of whether content is likely to have been examined and the strength of the evidence, care should be taken during page authoring to consider the nature and positioning of answer content so as to maximize the impact on reader beliefs. In addition, search engines may also want to consider the *position* of evidence and *strength* of evidence within documents should their objective be to influence or persuade searchers with the content that they surface, e.g., in scenarios where their goals include supporting searcher learning.

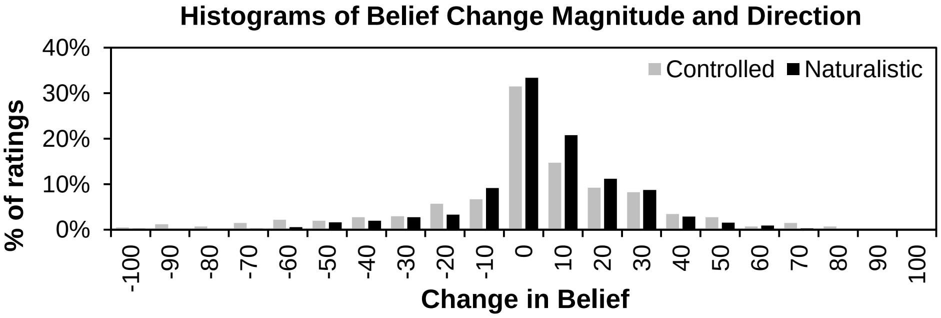


Figure 6. Distribution of rating changes for controlled SERP setting (gray bars) versus natural SERP setting (black bars) from Figure 3.

Table XI. Ratings of clicked results and average post-search belief, given pre-search belief for controlled setting.

Pre-search bucket	*Average Pre-search Belief*	*Average Clicked Snippet Rating*	*Average Post-search Belief*	*Percentage of clicks*	
				Helps	*Does not help*
Does not help	11.810	51.543	39.121	35.44%	64.56%
Diff $_{\text{Natural–Controlled}}$	+1.319	+9.107	+9.508	–3.87%	+3.87%
Inconclusive	53.125	55.529	60.434	58.69%	41.31%
Diff $_{\text{Natural–Controlled}}$	–0.794	+8.859	–5.342	–2.75%	+2.75%
Helps	87.044	60.877	80.542	70.40%	29.60%
Diff $_{\text{Natural–Controlled}}$	–0.352	+7.605	–3.474	–4.87%	+4.87%

4.3.2.2 Quantity of Information Viewed

Earlier in the article, we stated that we observed just under two clicks per task. Given the nature of this experiment, participants could choose varying numbers of pages, including no pages, whereby their assessment would be based entirely on the content of the SERPs that they examined (as happened for 60.1% of the tasks). In addition to examining the effect of the *answer type* of content viewed (as in Figure 5), we can also consider the *quantity* of content viewed. Our hypothesis is that the viewing of more information leads to more significant belief updates. Table X presents the average absolute change in rating as a function of the number of pages viewed per search task, irrespective of the label assigned to the pages viewed by the separate pool of judges. The findings show that, as the number of pages viewed increases, the absolute change also appears to increase. The correlation between the number of pages visited and the absolute belief change is both positive and significant ($r = 0.693$, $p < 0.001$).

4.4 Controlling for Content Bias (RQ4)

We ran an additional experiment with balanced SERPs where the pages were presented in random order. We referred to this as the “controlled” setting. In our analysis, we focus on two main aspects: (i) the changes in belief, and (ii) the nature of the results that were selected. We focus on SERP interactions in this analysis since these are most likely to be affected by changes in this controlled setting. Figure 6 shows a histogram of the belief changes for pre- and post-search. The results from the controlled study are denoted as “Controlled” in the figure. Those from other study described earlier (in Figure 3) are denoted as “Natural”, since they reflect the dynamics on the unaltered result set from the search engine. It is clear from the figure that the differences for Controlled are not as pronounced as those from Natural. However, they were similar and still significant (e.g., 43.3% move toward *helps* vs. 25.2% moved toward *does not help*, $p \leq$

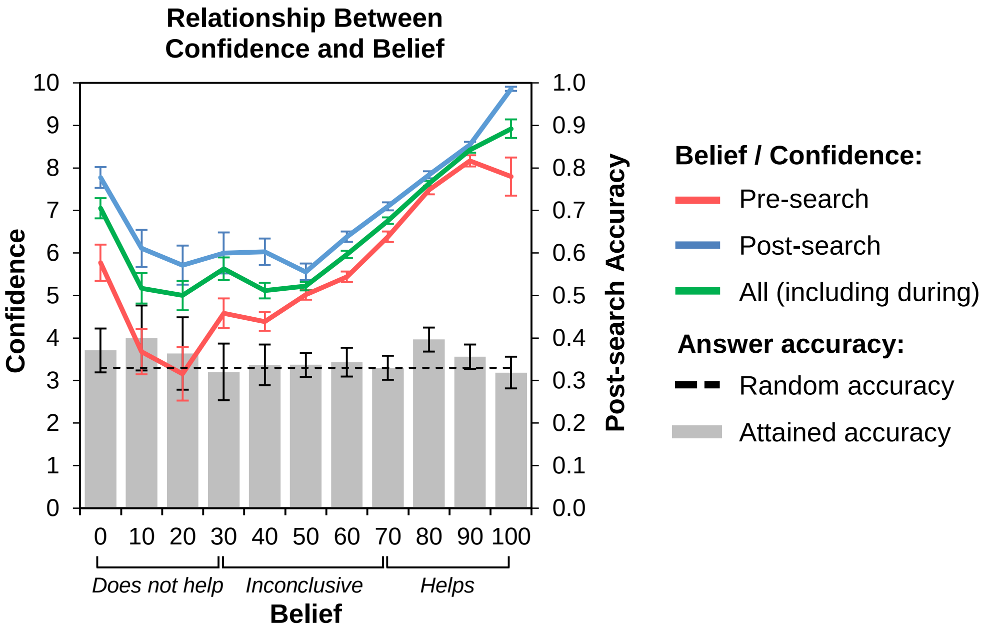


Figure 7. Confidence ratings versus belief for ratings captured pre-search, post-search, and across the full search episode including the per-page judgments obtained during the search (±SEM). Also shown below the x-axis are the belief rating ranges for the three task outcomes: *does not help*, *inconclusive*, *helps*.

0.03). Table XI shows the average pre- and post-search ratings, the average rating of the clicked snippets, and the fraction of the clicks on pages with *helps* and *does not help* labels—all grouped by the bucketed pre-search rating. The table also shows the differences between the values and those reported in the previous section. The results across these measures of search behavior and belief dynamics are very similar between the Controlled and Natural studies. Indeed, the ratings noted in the two studies are statistically indistinguishable (all $p > 0.21$). This shows that there are strong and predictable biases in the belief updates, that we observe with our methodology even when we control for the volume and rank of the content retrieved by the search engine.

4.5 Beliefs and Confidence (RQ5)

We are particularly interested in the relationship between beliefs about the truth of the assertion and the confidence with which those beliefs are held. Confidence, especially overconfidence, has been examined in studies of human judgment, e.g., [Griffin and Tversky 1992] but not in information seeking. Confidence is important in our setting because it may hinder belief updating given content retrieved by the search engine. We studied the relationship between beliefs and confidence and their joint influence on the dynamics of belief. We consider participants' beliefs and confidence before examining the search results (*pre*-search), after review of the search results (*post-search*), and overall, including per-page ratings captured from participants during the search process. Participants were asked to provide the confidence in their assessed probability on a scale from 0-10, higher is more confident. Figure 7 illustrates the distribution of

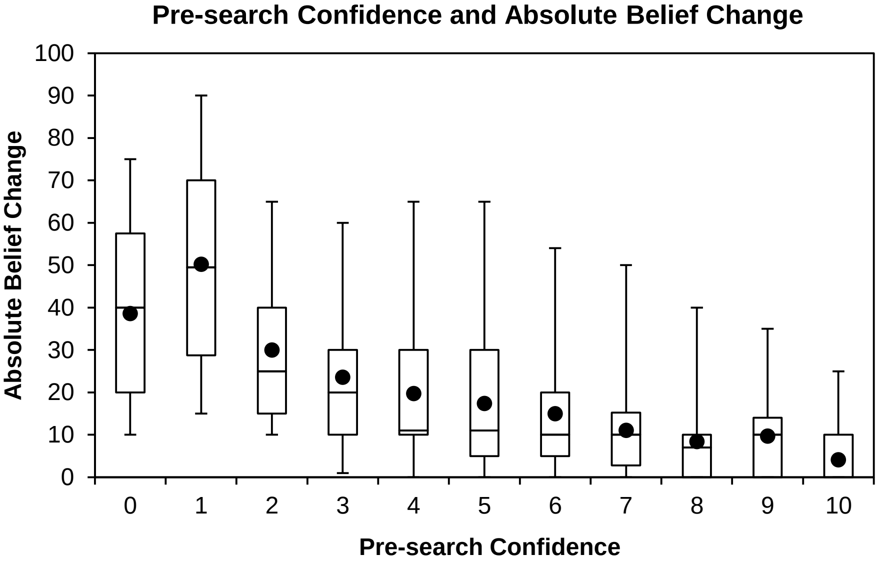


Figure 8. Box-and-whisker plot of the relationship between pre-search confidence and absolute belief change over the task. The mean belief change in each confidence bucket is also marked with a • symbol.

the ratings for three groups across the range of belief values.[3] In the figure, error bars denote the standard error of the mean.

Figure 7 shows that there that there are strong dependencies between assessed belief and confidence. When participants have stronger beliefs about the truth or falsity of assertions (closer to 0 or closer to 100), they appear more confident in the estimates that they provide. Confidence is minimized in the belief rating range [10,50], when beliefs are least certain. Across the full range of belief values, the confidence level before the searcher attempted the search task is consistently lower than that for when the search task was complete. As searchers gain more knowledge about the topic of the search during their examination of results, they become more confident in their beliefs about the answer they have selected. Previous research has also suggested that confidence may increase during the search process, albeit in longitudinal studies, and primarily in library settings [Kuhlthau 1991]. In this article we target changes in searcher confidence in a much shorter timeframe, i.e., within a single search episode.

Both belief and confidence may jointly influence information-gathering and world actions. Overconfidence in outcomes may suppress the collection of information and the updating of beliefs in light of the review of new information—and ultimately hinder decision making. Research in overconfidence in psychology suggests that people are frequently more confident in the accuracy of their answers than they should be per studies of accuracy of assessments [Griffin and Tversky 1992; Pallier et al. 2002].

To better understand the relationship between belief dynamics and the confidence values we analyzed the strength of participants' reported pre-search confidence in relation to the absolute magnitude of their belief update during the task. We plot the values as a box-and-whisker plot in Figure 8. The top and bottom of the box denotes the first and third quartiles, and the whiskers denote the maximum and minimum. We see that there is a clear negative relationship between the pre-search confidence

[3] In all cases, the belief ratings were collected at the same time as the confidence rating.

and the extent of the belief update. That is, as their pre-search confidence increases, participants tended to update their beliefs to a lesser extent ($r = –0.475$, $p < 0.001$). It appears that confidence is a significant moderator of belief dynamics of participants during our search tasks. In light of this, we were also interested in the impact of confidence on answer accuracy.

If we use the Cochrane reviews as ground truth, we can compute the accuracy of the ratings provided by participants. In our analysis we focus on post-search ratings since those are most closely connected with task outcomes. Answer accuracy is plotted in the histogram in Figure 6 using gray bars, with the secondary y-axis denoting the accuracy and the x-axis denoting the belief scores rounded up to the nearest bucket. We observe that the accuracy of the answers obtained is often no better than random (i.e., random accuracy is 0.33 given the equal distribution of three answer outcomes). This mirrors earlier results about near-random answer accuracy that were obtained on a different question set both in a natural setting (large-scale search log analysis) [White 2013] and in a remote user study with a methodology similar to that reported here [White 2014], and is significantly lower than the confidence values reported over the range of different belief ratings. Although the abstracts of the Cochrane reviews are accessible on online, they were not present in the sets of top-ranked search results returned by search engines for the topics used in our study. This demonstrates the challenge that searchers may encounter in obtaining a correct answer from search engines, even though the answer may be available somewhere on the Web.

Since many participants were examining results during the search tasks (40% of tasks had at least one click, average dwell time on pages was 68.3s), there are at least two explanations for the low accuracy: (i) participants are being misinformed by the content that they review in search results that are ranked based on content-match rather veracity [White 2013; White and Hassan 2014], and/or (ii) participants lack the domain knowledge before searching, meaning that they anchor on incorrect answers, and then do not deviate far from these erroneous pre-search beliefs during searching. Our data support both of these hypotheses: (i) the top-ranked search result in the natural setting was correct only 34.1% of the SERPs, and (ii) participants' pre-search beliefs were correct for only 35.6% of the search tasks. We also noted that participants modified correct pre-search beliefs to incorrect beliefs just as often as they modified incorrect pre-search beliefs to correct beliefs ($\chi^2(1) = 0.614$, $p = 0.43$). Preserving significantly more correct answers than incorrect answers would have suggested some domain knowledge among our participants and/or some support from the search engine in making this distinction. Although they are not captured on the same scale (confidence ranges from 0 to 10, accuracy ranges from 0 to 1), we assume that they can be compared for this analysis of overconfidence. Since the search engine returned more pages labeled *helps* and it is easier to confirm than refute an assertion, searchers with pre-search belief of *helps* may be more susceptible to overconfidence. Indeed, Figure 7 shows that overconfidence is maximized among those with a pre-search belief of *helps*.

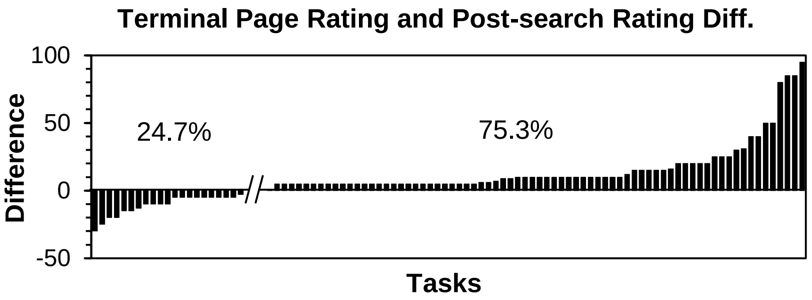


Figure 9. Distribution of differences between the terminal page belief and post-search belief, excluding 67.5% cases where ratings did not change (*N*=265).

4.6 Changes following Last Page View (RQ6)

Thus far, we have focused on the importance of the pre-search belief and the influence of examined content. We cannot expect belief updates to end with those associated with the terminal page view. Upon analyzing the data, we noted that quite frequently (on 32.5% of occasions), the belief provided for the last page viewed did not equal the post-search belief assessed from the participant at the end of their task. This suggests that other factors influence post-search ratings beyond the content of the pages (e.g., the SERP content), or that synthesis and reflection upon the information gathered during the task takes time and that people view their beliefs holistically in light of the sum of all information collected during the task. The distribution of the rating changes following the belief assessed for the last page is shown in Figure 9. 75.3% of the changes show an increase (i.e., toward *helps*). Looking more closely, we find that 70.2% of the changes represent a movement *further* from participants' pre-search beliefs, rather than a regression toward it, which we might have expected if the change in belief was temporary. The time between terminal page rating and the post-search rating was on average 31.45 seconds (median = 18 seconds), signifying that participants may have thought carefully and/or examined the SERP in quite some detail before concluding the task. During this time, searchers may be synthesizing the information that they have been exposed to during the full search episode. Indeed, participants who moved toward *helps* following the last page view were 4.30 times as likely to have viewed *helps* content than *does not help* content during the search task ($Z = 4.33$, $p < 0.001$). The average over all tasks and participants was 1.94 times as much *helps* content as *does not help* content. This is important since during belief modeling we may wish to make point estimates of belief status based on reviewed content such that tailored support can be offered in real time. However, searchers may not be ready for tailored content immediately following the page view. More work is needed to understand these updates and their temporal dynamics, as well as to better understand the longer-term influence of a sequence of belief updates, coming via exposure to a sequence of content.

4.7 Predicting Belief Dynamics (RQ7)

Given that we observe such clear patterns in searchers' belief updates, we were interested in whether we could build a predictive model capable of inferring revised beliefs following the review of content. To perform this prediction, we featurized the self-report data provided by participants as well as implicit signals from their search activity and the third-party judgments of the content that they accessed. We now present findings on this prediction task. We begin with all of the features, but we also: (i) assess the contribution of different feature classes, (ii) assess model performance for unde-

Table XII. Performance of predictive models versus the baselines. Results are reported in terms of accuracy, area under the receiver operator characteristic curve (AUC), mean absolute error (MAE), and normalized root mean squared error (NRMSE).

Feature class	***Classification***		***Regression***	
	Accuracy	**AUC**	**MAE**	**NRMSE**
Full model	0.8320	0.8412	9.530	14.351
Confidence baseline	0.6755	0.6880	15.391	23.227
Marginal baseline	0.6288	0.5000	33.513	45.973
Random baseline	0.3334	0.5000	37.806	49.208

cided searchers who do not possess a strong belief before searching, making the prediction task potentially more challenging given the significant influence of pre-search beliefs, and (iii) examine the performance of the model if we exclude information about beliefs and confidence explicitly provided by searchers—as this information is unlikely to be provided by searchers in practice, as well as the answer ratings for search results provided by third-party judges.

4.7.1 All Features

We employed the broad range of features described earlier in this section for both the classification and the regression task. Table XII shows that the performance of the model for both of these tasks is strong. The full model accurate assigns the correct rating bucket (helps, inconclusive, does not help) 83% of the time; well above the accuracy the three baseline models (which had accuracies ranging from 33% to 68%, depending on the features used). All differences the three baselines are significant at $p <$ 0.001 (using paired t-tests). The MAE for the regression task given the full model is 9.5 (out of 100). This means that on average the full model with access to all features could accurately predict the actual post-search belief rating within 15% of the actual belief rating assigned by study participants.

4.7.1.1 Feature Contributions

We also sought to understand the effect of individual features on the performance of our models. Table XIII lists the top ten features and their impact on classification performance, relative to the most important feature, namely *PreSearchRating*. Given the findings reported earlier in this article on how beliefs rarely changed dramatically, it seems reasonable to expect that the pre-search belief contributed most to the outcome of the search. The other most useful features relate to other aspects of the self-report data provided by participants, including their confidence levels and their historic beliefs. *UserAvgPreConfidence* and *PreSearchConfidence* were likely strong contributors because of the relationship between confidence and belief revision that we observed in our earlier analysis. That said, it is a concern that the most useful features in the model rely on searcher self-reporting, which may not happen in practice given the overhead for searchers. We found that *AvgResultRating* is also an important feature. As this statistic is a property of search result content, it could be considered during ranking. Doing so would require data on the nature of the result (e.g., the type of answer it contains). In practice, this could be estimated via methods such as sentiment analysis [Pang and Lee 2008] and answer extraction [Abney et al. 2000] rather than human judges as we employed in this study. The reliance on self-reporting and human-judgment-based features for our prediction experiments limits their generalizability. As such, in Section 4.7.4, we explore the performance of the models if we focus on features that are based only on signals derived available without self-reports or third-party page judgments.

Table XIII. Relative contributions of individual features to the overall performance of the predictive model (regression task). The top 10 features with highest evidential weight are listed in the table.

Feature	***Feature Class***	***Relative weight***
PreSearchRating	Task	1.0000
UserAvgPostSearch	User History	0.4723
UserAvgPreConfidence	User History	0.2897
AvgResultRating	Results and Captions	0.2147
UserAvgPostConfidence	User History	0.2078
AvgExplanationPosition	Results and Captions	0.2035
PreSearchConfidence	Task	0.1820
UserAvgChange	User History	0.1807
TimeSpentOnJudgment	Search Activity	0.1675
UserAvgPreSearch	User History	0.1624

Table XIV. Performance of different feature classes. Sorted by accuracy in descending order. Statistical differences in accuracy are noted w.r.t. the full model using paired *t*-tests at * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Similar differences hold for the other metrics for classification and regression.

Feature class	***Classification***		***Regression***	
	Accuracy	**AUC**	**MAE**	**NRMSE**
Task	0.7647*	0.7755	11.561	17.11
User History	0.7305**	0.7470	12.567	19.65
Search Activity	0.7300**	0.7321	12.655	19.81
Results and Captions	0.7187**	0.7222	13.772	21.50
Question	0.6689***	0.6801	16.554	25.27
All feature classes	**0.8320**	**0.8412**	**9.530**	**14.35**

4.7.2 Feature Classes

We were also interested in how each of the five feature classes contributed to the overall effectiveness of the prediction model. We assessed the impact of each class separately and reported the results in Table XIV. The results suggest that the features related to the self-report data (*Task* and *User History*) were most useful, but there were also promising aspects of the other behavioral and content features that may be useful in making predictions about belief updates in the absence of explicit self-reporting. We explore the effectiveness of models trained only on these features in Section 4.7.4.

4.7.3 Uncertain Searchers

The analysis described in this section suggests that an important factor in predicting belief revision is pre-search belief. When this belief is strong, the searcher is likely to stick with the prior beliefs regardless of the results returned by the search engine. However, some searchers are undecided at the outset of their search sessions and we wanted to understand prediction performance for this set of tasks. In particular, we wanted to verify that the model was not only working in cases when there were strong pre-search beliefs, i.e., that it could learn the correct directionality of the belief revision based only on search behaviors. In doing this, we built a predictive model over all searchers, but only evaluated based on the tasks where participants provided a pre-search rating of 50. While the pre-search rating would likely be unavailable to search engines in practice, we still believed that it was important to measure model performance for this potentially-challenging subset of tasks. The results show that the predictive performance for this particular set of tasks is still strong (i.e., classification accuracy = 0.7711, NRMSE = 17.59). The marginal performs less well (accuracy = 0.5996, NRMSE = 47.22) and our model significantly exceeds the performance of the marginal (e.g., 29% gain in accuracy, 63% reduction in NRMSE; both significant at p

Table XV. Relative contributions of individual features to the overall performance of the predictive model (Implicitly collected, non-self-report features only, classification task). The 10 features with highest evidential weight are shown in the table.

Feature	***Feature Class***	***Relative weight***
AvgResultRating	Results and Captions	1.0000
NumDoesNotHelpPagesClicked	Search Activity	0.5721
NumHelpsPagesClicked	Search Activity	0.5220
AvgCaptionRating	Results and Captions	0.4838
TimeOnSERP	Search Activity	0.4641
TimeOnTask	Search Activity	0.3600
AvgDwellTime	Search Activity	0.2440
MaxClickedResultRating	Search Activity	0.1788
AvgClickedResultRating	Search Activity	0.1745
GroundTruth	Question	0.1570

Table XVI. Relative contributions of individual features to the overall performance of the predictive model (Implicitly collected, non-self-report / non third-party features only, classification task). The 10 features with the highest evidential weight are shown in the table.

Feature	***Feature Class***	***Relative weight***
ClickPosition	Search Activity	1.0000
QueryStartsWithCan	Question	0.7681
TimeOnSERP	Search Activity	0.6555
NumUniqueDomains	Search Activity	0.4632
TimeOnTask	Search Activity	0.3877
QueryStartsWithShould	Question	0.3111
AvgDwellTime	Search Activity	0.2919
NumClicks	Search Activity	0.2375
TimeOnPages	Search Activity	0.2174
NumUniqueTLDs	Search Activity	0.1818

< 0.001). This also suggests that the *directionality* of belief change (i.e., *move positive*, *same*, *move negative*) can also be estimated effectively.

4.7.4 Implicit Features

Many of the features employed in the predictive model relied on having access to data from human judgments (self-reporting, third-party result labeling, or the nature of the ground truth). We wanted to understand how well we could predict the post-search rating without access to these features, since it is unlikely that a search engine would have access to them in practice. We removed two sets of features from the model: (i) remove self-reporting features, and (ii) remove self-reporting *and* third-party features:

—**Removing all features based on self-reporting:** Removing these features from the predictive model resulted in a decrease in overall performance (accuracy = 0.7377, NRMSE = 19.54), but performance remained well above that of the marginal models, which always predicts a revision toward *helps* (accuracy (from Table XII) = 0.6288). Aggregating the implicit features over preceding search tasks in a similar manner as with self-report features (*UserActivityContent*) leads to a significant increase in the predictive accuracy (from 0.7377 to 0.7869) and a reduction in NRMSE (from 19.54 to 16.30) (both significant at $p \leq 0.002$). In examining the implicit features that were most useful in predicting the revised belief (Table XV), we observe that a range of feature classes are represented. However, the most important features were related to the average page and caption ratings of the search results retrieved, the nature of the results that participants selected, and the time spent searching (both in total for the search task and on average dwell time per result viewed).

—**Removing all features based on self-reporting *and* third-party labels:** In an additional experiment, we removed all features that were based on self-reporting and judgment by third party judges. This reduced our feature set quite considerably (to only a subset of those in the *Search Activity* and *Question* classes), but we believe that it reflects quite accurately the features that search engines could model without any additional labeling. As expected, performance decreases with respect to the full model and the model without self-report features (accuracy = 0.6833, NRMSE = 22.04), but performance is still significantly greater than that of the marginal and random baselines (both $p < 0.01$). The ten most important features are listed in Table XVI. Interestingly, the most influential features relate to where in the results the searcher clicked, the nature of the query terms selected (specifically the presence of the word "can" and "should", which denote possibility and the pursuit of advice respectively), and the total amount of time spent examining the SERP. This suggests that how the search engine processes queries containing particular terms, ranks the search results, and presents the results via SERP captions are important factors in belief revision.

5. DISCUSSION AND IMPLICATIONS

We have studied the dynamics of the beliefs of people performing search using a methodology that allows us to crowdsource the study of changes in beliefs and confidence levels during search activities, and to assess the impact of the search engine and other factors on those updates. We have found that confidence increases during search tasks and identified dependencies among beliefs and confidence during information seeking episodes, specifically that high confidence limits belief updating. We also showed evidence for the presence of confirmation biases in the behavior surrounding result selections as well as in the dynamics of beliefs over the tasks assigned. Biases in judgment can be considered when analyzing search behavior retrospectively, but may also be harnessed to enhance search when considered in the generation of results in real time (e.g., by applying a searcher's belief profile in generating a personalized list of results). The automated methods that we present for inferring searcher beliefs from behaviors provide a means by which systems could model searchers' belief state without having to explicitly probe searchers to obtain that information. That said, more accurate predictions about beliefs and belief updating could be made if rating data could be collected from searchers in a lightweight way.

We summarize the primary insights from our study of belief dynamics as follows:

—Pre-search beliefs are important determinants of search interaction, and some beliefs are difficult to update, especially if they are strongly held at the outset. This finding suggests that it could be valuable for search providers to capture and represent searcher beliefs, especially pre-search beliefs, both during search sessions and during retrospective analyses of searcher behavior via search log data. This is important if search engines want to personalize the search experience to match a searcher's prior beliefs or transmit information that contradicts them (e.g., to correct erroneous beliefs).

—The relationship between page content and pre-search beliefs affects how people examine those pages (disagreement with prior beliefs results in longer dwell times), and the strength/location of evidence on the pages affects if and how beliefs are updated. One clear implication from this finding is that when interpreting dwell times for applications such as satisfaction estimation [Kim et al. 2014], search engines should consider the relationship between searcher beliefs, and the opinions and sentiment of the pages accessed during the search process.

—Searchers seek positive (*helps*-related) information, even when search engines control for the rank and the quantity of content in results (equal distribution of *helps* and *does not help* pages, random ordering of those pages). This bias in search behavior has been observed in previous studies in the search domain [White 2013], and needs to be considered as important background when interpreting clickthrough signals.

—Confidence plays an important role in belief revision during search. Participants who were more confident in their beliefs were less likely to revise them during their search. There was strong evidence of overconfidence, especially when the belief was *helps*. Searcher confidence has largely been ignored in retrieval settings, but its impact on belief revision during search is significant, and should be considered in belief modeling.

—Post-search beliefs can be inferred with reasonable accuracy using features of the user search activity and the search query. Performance is stronger given access to self-report data and labels from other sources about the nature of the results and the ground truth, but those could be difficult to attain in practice given the reporting overhead and the costs involved. Automated methods for labeling results and obtaining ground truth (e.g., from medical literature or normative base rates) may be feasible but further work is needed to determine their viability. Longitudinal analysis of belief revision yields more accurate models to provide estimates on pre-search beliefs for future sessions given one or more historic search sessions. This can be useful in interpreting and modeling beliefs and their dynamics during the current search episode.

Our findings demonstrate aspects of the nature of belief dynamics during the execution of search tasks. We focused on experiments where people could click to access pages as many times as they preferred and provided feedback on their beliefs after reviewing each page. The unconstrained scenario provided a relatively natural interaction experience. The assessments of belief after each page view enabled us to monitor belief dynamics over the course of the search session. However, we also believed that removing the need for per-page ratings would help ameliorate possible demand characteristics [Orne 1962] that may be evident from probing participants frequently about their belief state. We acknowledge that such artefacts of the experimental design could impact the reliability of our findings. To address this, we performed experiments that did not ask for per-page ratings and other variants that placed greater requirements on our participants (e.g., where they had to select at least n results for each task that they attempted)[4]. The results of all of these experiments were statistically indistinguishable from those reported earlier (all $p > 0.24$). We conclude that there is no significant impact from our particular experimental design choices regarding when and what ratings were elicited from participants during our crowdsourced studies.

We also found that participants were drawn toward results describing the effectiveness of treatments and results that supported their prior beliefs, if such content is shown to them. It could therefore be beneficial to develop presentation strategies that account for such a selection bias. Strategies might even include employing an “all or nothing” approach whereby *only* information aligned with a particular perspective is shown. Consideration of such support is needed as part of a broader discussion concerning the role of search engines in belief revision. Questions include: Should search engines model and shape beliefs? If so, should the goals of belief revision be some assessed notion of accuracy, or of a policy on communicating the distribution or diversity of possible answers? If searchers possess erroneous beliefs, should search systems try

[4] Although we required that people provide assessments for the pages that they viewed, this appeared to have had little impact on the number of clicks. Crowdworkers who performed multiple tasks exhibited little correlation between task number and the number of results clicked ($r = -0.02$, $p = 0.86$).

to correct them or facilitate access to supporting information? What is the impact of the search task on this decision, e.g., should the pursuit of potentially-harmful medical treatments be regarded differently than pursuing skewed political information? How should actions be communicated with searchers, e.g., with options to reverse system interventions? We believe that search engines should consider beliefs in interpreting search behavior and that they should consider the ramifications of inaccurate answers when making determinations about the display of results. For consequential topics such as health, accuracy should be paramount. For controversial topics such as political or moral issues, the presentation of balanced perspectives may be appropriate. More research and reflection are necessary for understanding how to best harness insights and machinery for inferring beliefs and belief updates in search and retrieval.

Our analysis of directly predicting post-search belief ratings allows us to model the impact of content on searcher beliefs in the absence of specific historic data for searchers. The analysis of belief updates for searchers is performed retrospectively, the searcher need only engage in one search session for the system to make reliable predictions of their actual belief rating based on behavioral signals (at 73% accuracy), rising considerably (to 79% accuracy) given access to more behavioral evidence attainable from tracking a single user longitudinally across all of their judgment tasks, i.e., up to and including the last task observed for each searcher (the marginal has 63% accuracy). Predicting belief revision has a number of different applications, including understanding the impact of content on searchers' viewpoints and learning "persuasive" retrieval models that guide searchers toward accurate answers if they are available or present balanced perspectives to moderate searcher attitudes [Liao and Fu 2013]. Insights about beliefs and belief revision during search might also be used to detect the presence of procedures designed to persuade searchers. Making predictions retrospectively about beliefs and belief revision with information could also be useful for tasks such as better interpreting observed page dwell times from search logs, characterizing the impact of content on beliefs, and in generating training data for machine-learned ranking algorithms that consider belief revision in addition to relevance. Alternative challenges include predicting a searcher's pre-search belief such that the search experience could be tailored accordingly, predicting the impact that particular pages could have on searcher beliefs, or predicting the nature of the change in belief (e.g., increase, stay the same, decrease) given a search episode. We made some progress toward all of these goals in this article, but focused primarily on post-search beliefs.

Our findings lead to important design implications for search providers:

—**Consider beliefs in interpreting behaviors, including clicks and dwells:** Our findings suggest that pre-search beliefs and the relationship between these beliefs and the content of examined pages had a dramatic effect on information behavior. As such, these beliefs should be considered when interpreting actions such as clickthrough behavior (e.g., the extent to which a click is related to objective relevance and the extent to which it is related to the confirmation of an existing belief?) or dwell times (e.g., does the viewpoint on the page contradict the searcher's viewpoint? If so, dwell durations may be shorter than expected, irrespective of searcher satisfaction).

—**Label the nature of the content in search results:** In addition to labeling topicality of search results, which has been well explored (e.g., [Bennett et al. 2010]), search engines also need to label the answers or perspectives noted within search results. This could be performed manually for a particular set of queries as we did in this study. However, given the large number of documents and the scale of the Web, automated methods need to be developed. Methods already exist for automatic answer extraction, which could be used as the basis for this approach (e.g., [Abney et al. 2000;

Dumais et al. 2002]). Assessing the nature of content seems to be essential given the observed importance of the content labels in our predictive models of belief revision;

—**Consider the location of salient evidence in the content retrieved:** Search engines aimed at providing the most valuable support, considering biases and affordances of cognition (e.g., to support searcher learning), need to consider the influence of the position of the salient evidence within pages. Traditionally, information retrieval algorithms only consider the presence or absence of terms within documents. Our results suggest that pages with strong evidence placed near the beginning are most likely to result in belief revision following review. This shares similarities with our previous work [White and Horvitz, 2010], in which we found that the review of Web pages where serious illnesses were mentioned before benign explanations was more likely to lead to queries containing evidence of escalations in medical concerns;

—**Model beliefs during personalization:** Richer models of searcher intentions and interests can be developed by modeling searchers' beliefs as well as their topical interests. These beliefs can be communicated explicitly by searchers and updated based on observations about their interactions. These beliefs can also be captured implicitly based on search interaction behavior given evidence captured from one or more sessions. In doing so, there are challenges in determining which aspects of models of beliefs to update at a given time (i.e., the topic of the current search), and;

—**Reflect about information goals and human cognition:** Search providers should reflect about the goals of their services, and fold such reflections into designs. For example, designers interested in providing correct information to searchers may wish to consider how to best transmit that information to them, taking into consideration their likely prior beliefs, confidences, and likely belief updating. Designers may also wish to consider searchers' desire to have their belief validated, regardless of factual correctness. Favoring accurate answers may result in better decisions by searchers, while validating beliefs may lead to more engagement and higher levels of searcher satisfaction (more returning sessions over time, etc.). This is a decision that the search engine designers may wish to make on a query or on a searcher level, e.g., some queries may be consequential (e.g., health related) and the focus should perhaps be on veracity, whereas some searchers may have strongly-held, immovable beliefs or search on controversial topics for which validation and its role in engagement might be favored. Considering veracity in health-related searching more generally, medical software is already regulated by the United States Food and Drug Administration and the European Union. If search engines continue to provide medical information that plays a significant role in influencing people's health-related decision making, then the formal review and regulation of medical support in search engines may eventually be necessary. For search providers, the role of commercial outcomes (specifically advertising), and optimizing for such outcomes while also providing searchers with accurate results and/or evidence to validate their beliefs, requires further exploration and review.

We acknowledge several limitations of this study:

— We used crowdsourced judges given the number of ratings required. However, such judges do not have the same interests and goals as searchers pursuing health information. Thus, they will frequently lack the intrinsic motivation during search that people truly seeking answers would likely have. As such, the absolute numbers reported in this article should be interpreted with some care—however, the trends in our findings remain both clear and noteworthy. The results that we present may be sensitive to changes in the level of motivation searchers have with learning about the efficacy of medical interventions. There are reasonable concerns that crowdworkers were operating under extraneous time pressure that affected their task performance, and more generally, concerns about noise in crowdsourced data [Hsueh et al. 2009; Kazai 2011]. While it is difficult to determine levels of participant engagement, our participants did spend on average one-and-a-half minutes (88.7s) per task, and to improve quality we did exclude those who spent considerably less time on their tasks. Follow-up user studies are needed with participants who are fully motivated to learn about their health and well-being. Studies with real patients are certainly possible, e.g., the recent European Union Kreshmoi Project (kreshmoi.eu) has successfully studied patients searching for information related to their medical conditions (e.g., [Pletneva et al. 2013]).

— The use of search engine log data to find the seed queries to generate our result sets limits the reproducibility of our study. To run a similar experiment without access to logs, researchers could obtain query-like statements similar to those in our seed set via a crowdsourcing task. Given an assigned Cochrane review, crowdworkers could generate queries based on our requirements outlined in Section 3.2.2.2 (overlap with review title, sequence ordering of query terms, etc.). These queries could then be used as the basis for generating result sets similar to those used in our analysis.

— We focused on a single domain, on beliefs about the truth of assertions about the efficacy of medical treatments and interventions, relying solely on healthcare topics selected and reviewed by Cochrane, and taking the summary Cochrane assessments as ground truth. The Cochrane data may contain unforeseen, implicit biases in the types of problems that are considered and accepted by the Cochrane Collaboration for further review (e.g., reviews of conditions and treatments that are believable and popularly known or considered as true or false within medicine or across populations of laypeople). Although we controlled for the answer distribution truth as part of the experimental design, more research is needed into the nature of the reviews themselves and biases inherent in their construction. Our labeled data reflects the state of medical knowledge at the point that our analysis was performed. However, the nature of medicine is such that the findings of a single study may change the review recommendation, and as a result, alter our ground truth. In addition, quantifying the extent of any contradictory results within each of the selected topics and the potential impact of such uncertainty on the reliability of our conclusions is an important area for future work. Performing similar studies in other domains would be helpful for identifying generalizable cross-domain principles of the dynamics of belief with search and retrieval.

— The task-by-task nature of our experimental design means that the internal validity of our experiment could be threatened by factors such as maturation, regression toward the mean, and repeated testing (c.f. [Shadish et al. 2002]). Such factors may influence reliability of any conclusions drawn from the study. More work is needed on understanding the role, if any, that these factors play in our experimental outcomes.

—Beliefs and biases impact how people are affected by the content that they review (e.g., examining a landing page), as well as decisions about the information that they decide to review (e.g., in selecting a result on the SERP). The focus in biases in retrieval has largely been on document selection [Joachims et al., 2007; Yue et al., 2010; Ieong et al., 2012; White, 2013]. In contrast, studies of cognitive biases in psychology have examined the impact of fixed content on belief revision [Tversky and Kahneman, 1974]. Our analysis considers both factors, mixed in different ways throughout the article. Any *interactions* between biases, selection decisions, and content reviewed are not considered, but need to be studied to fully understand belief dynamics and biases.

— Finally, it is known from research on survey methodologies (c.f. [Dillman 2000]) and psychometrics (e.g., [Wallsten and Budescu 1983]) that eliciting subjective probability estimates from people can lead to unreliable data. Allowing our study participants to report their belief estimates on a 0-100 scale enabled a granular analysis of belief dynamics. However, this could lead to data collection biases (e.g., 59.7% of the ratings that we collected were 0 or multiple of 10). As part of our future work in this area, we need to explore the impact of alternative elicitation methods such as sliders or multipoint scales on the reliability, consistency, and distributions of our response data.

Directions of research also include further pursuit of understanding of the influence of specific aspects of examined content on searchers' short- and longer-term beliefs. We will also seek to explore ways to integrate belief revision into retrieval systems, as has been attempted in adaptive search systems [Lau et al. 2004]. The methodology that we adopted—specifically limiting engagement in the study to one task per topic per participant—is not conducive to studying searchers over longer timeframes. We only considered these predictions for the domain that we focused on and for a limited case library, containing at most 100 tasks per participant. We expect that data from search behavior over longer time frames could be leveraged to enhance the reliability of such inferences. Studies are needed with more tasks per searcher and for different domains to investigate the feasibility broader inferences about belief dynamics.

The page labels that we collected from a separate pool of judges provide insight into the nature of the pages, but richer analysis of the content and the structure of the content examined is also necessary. Advanced tracking mechanisms such as cursor/gaze tracking technology (e.g., [Buscher et al. 2008; Guo and Agichtein 2012]) could be useful in understanding which parts of Web pages searchers have examined and map that content directly to inferences regarding changes in belief. Previous work has shown that the structure of content on Web pages (e.g., location, relative positioning of discussion of benign and serious health concerns) about such information as the diagnostic implications of symptoms can affect behaviors and beliefs, with effects on escalations in follow-on queries [White and Horvitz 2010]. We showed some effects of position of evidence on pages on the likelihood of a belief update following the review (including considering the likelihood that content was reviewed, given the location of the evidence). We could also use other structural information to better understand the nature of the page (and hence better predict the update that will result), such as the presence of images or formatting of the content presented, on belief revision.

We also need to explore the impact of manipulations in the search results on the nature of the belief dynamics, e.g., by restricting the result set to only contain confirmatory information, or promoting answer pages of a particular type, e.g., ranking *helps* content above *does not help* content, as in previous work [White 2014]. These pages can help shape searcher beliefs and mitigate biases that may exist in their beliefs or the content returned by the search engine. Further work is also required regarding if and how to challenge or even shift the stance of searchers who possess strongly-held, but factually-incorrect beliefs. Emerging research at the intersection of personalization and persuasion suggests that there is value in systems that positively influence attitudes, intentions, and behaviors [Berkovsky et al. 2012]. In doing so, care needs to be taken in determining and framing the intended outcomes of any action taken by the system (e.g., for erroneous beliefs the emphasis should be on education, and the benefits of greater knowledge, rather than persuasion, which has negative connotations). Deeper inferences about searcher beliefs could also be used in more sensitive applications such as the optimal placement of display advertising for medical topics. In supporting the use of such inferences, search providers may need to restrict and review the advertisements shown or the advertisers themselves, so as to not erode searcher trust regarding how searchers' belief models are applied in practice.

6. CONCLUSIONS

We have presented studies of belief revision during and following online search episodes. Focusing on the domain of health search, we employed a crowdsourcing platform to gain access to participants in several studies and phases of study. In the experimental methodology we controlled the distribution of search results and captured beliefs at the outset, during the performance, and at the conclusion of search tasks. We collected probability estimates from crowd-sourced searchers reflecting their beliefs about the efficacy of medical interventions at these different points in time. The analyses revealed evidence of confirmation bias in result selection decisions and differences in dwell times on pages given agreement with prior beliefs. We found that participants spent less time on pages where the ratings disagreed with their belief at the outset of the search task. Searcher confidence was also found to be an important determinant of belief updating; highly-confident searchers were less likely to revise their beliefs. The findings highlight the importance of considering beliefs and biases when interpreting behavioral data such as click-through signals. We demonstrated the ability to make predictions about belief updating, even in the absence of self-report data. These findings suggest that search engines can model searcher beliefs by monitoring behaviors and content. We hope that the methods and results will motivate ongoing research and development of new models of search behavior that consider the nature and evolution of searchers' beliefs and the use of such inferences for personalizing search results and for recommending queries and content.

REFERENCES

Steven Abney, Michael Collins, and Amit Singhal. 2000. Answer extraction. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLC '00)*. Association for Computational Linguistics, 296–301.

Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2011. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM Press, New York, NY, 345–354.

Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving Web search ranking by incorporating user behavior information. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM Press, New York, NY, 19–26.

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *ACM SIGIR Forum*, 42, 2, 9–15.

Norman H. Anderson. 1981. *Foundations of Information Integration Theory*. New York: Academic Press.

Marcia J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13, 5, 407–424.

Mike Bengeri and Pierre Pluye. 2003. Shortcomings of health-related information on the internet. *Health Promotion International*, 18, 4, 381–387.

Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks. 1982. ASK for information retrieval: Part I - background and theory. *Journal of Documentation*, 38, 2, 61–71.

Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9, 3, 379–395.

Allan Bell. 1991. *The Language of News Media* (pp. 84-85). Oxford: Blackwell.

Paul N. Bennett, Krysta Svore, and Susan T. Dumais. 2010. Classification-enhanced ranking. In *Proceedings of the 19th International Conference on the World Wide Web (WWW '10)*. ACM Press, New York, NY, 111–120.

Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM Press, New York, NY, 185–194.

Shlomo Berkovsky, Jill Freyne, and Harri Oinas-Kukkonen. 2012. Influencing individually: Fusing personalization and persuasion. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 2, 2, Article 9.

Celia Boyer. 2013. The internet and health: international approaches to evaluating the quality of web-based health information. In *eHealth: Legal, Ethical and Governance Challenges*. Springer Berlin Heidelberg, 245–274.

Lyle A. Brenner, Derek J. Koehler, Varda Liberman, and Amos Tversky. 1996. Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, 65, 3, 212–219.

Georg Buscher, Andreas Dengel, and Ludger van Elst. 2008. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM Press, New York, NY, 387–394.

Stuart Card, Thomas Moran, and Alan Newell. 1983. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum, Hillsdale, NJ.

Junghoo Cho and Sourashis Roy. 2004. Impact of search engines on page popularity. In *Proceedings of the 13th International Conference on the World Wide Web (WWW '04). ACM Press, New York, NY,* 20–29.

Charles L. A. Clarke, Eugene Agichtein, Susan T. Dumais, and Ryen W. White. 2007. The influence of caption features on click through patterns in Web search. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. ACM Press, New York, NY, 135–142.

Mary Czerwinski, Eric Horvitz, and Edward Cutrell. 2001. Subjective duration assessment: An implicit probe for software usability. In *Proceedings of the IHM-HCI Conference (IHM-HCI '01)*, 167–170.

Brenda Dervin. 1983. An overview of sense-making research: Concepts, methods, and results to date. In *Proceedings of the International Communication Association*.

Don A. Dillman. 2000. *Mail and Internet Surveys: The Tailored Design Method* (Vol. 2). New York: Wiley.

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. 2002. Web question answering: Is more always better? In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. ACM Press, New York, NY, 291–298.

Jon Elster. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Gunther Eysenbach and Christian Köhler. 2002. How do consumers search for and appraise health information on the World Wide Web? Qualitative studies using focus groups, usability test, and in-depth interviews. *British Medical Journal*, 324, 7337, 573–577.

Leon Festinger. 1957. *A Theory of Cognitive Dissonance* (Vol. 2). Stanford: Stanford University Press.

Peter Fischer, Andreas Kastenmuller, Tobias Greitemeyer, Julia Fischer, Dieter Frey, and David Crelley. 2011. Threat and selective exposure: the moderating role of threat and decision context on confirmatory information search after decision. *Journal of Experimental Psychology: General*, 140, 1, 51–62.

Brian J. Fogg. 2002. Persuasive technology: Using computers to change what we think and do. *Ubiquity*, 5.

Susannah Fox. 2006. *Online Health Search 2006*. Pew Internet and American Life Project. http://www.pewinternet.org/2006/10/29/online-health-search-2006/. Accessed March 2015.

Susannah Fox and Maeve Duggan. (2013). *Health Online 2013*. Pew Internet American Life Project. http://www.pewinternet.org/2013/01/15/health-online-2013/. Accessed March 2015.

Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve Web search. *ACM Transactions on Information Systems (TOIS)*, 23, 2, 147–168.

Dieter Frey. 1986. Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, 19, 1, 41–80.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. *Additive Logistic Regression: A Statistical View of Boosting*. *Annals of Statistics*, 28, 2, 337–407.

R. Kelly Garrett. 2009. Echo chambers online? Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, 14, 2, 265–285.

Arnaud Gaudinat, Patrick Ruch, Michel Joubert, Philippe Uziel, Anne Strauss, Michèle Thonnet, Robert Baud, Stéphane Spahnif, Patrick Weberf, Juan Bonalg, Celia Boyera, Marius Fieschic, Antoine Geissbuhler. 2006. Health search engine with E-document analysis for reliable search results. *International Journal of Medical Informatics*, 75, 1, 73–85.

Gerd Gigerenzer and Peter M. Todd. 2000. *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.

Gerd Gigerenzer and Daniel G. Goldstein. 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103, 4, 650.

Eric Goldman. 2006. Search engine bias and the demise of search utopianism. *Yale Journal of Law and Technology*, 2005-2006.

Dale Griffin and Amos Tversky. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 3, 411–435.

David J. Hand and Robert J. Till. 2001. A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45, 2, 171–186.

William Hart, Dolores Albarracin, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135, 4, 555–588.

Chip Heath and Amos Tversky. 1991. Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4, 1, 5–28.

Bradford W. Hesse, David E. Nelson, Gary L. Kreps, Robert T. Croyle, Neeraj K. Arora, Barbara K. Rimer, and Kasisomayajula Viswanath. 2005. Trust and sources of health information: the impact of the internet and its implications for health care providers: findings from the first health information national trends survey. *Archives of Internal Medicine*, 165, 22, 2618–2624.

Julian P. Higgins. (Ed.). 2008. *Cochrane Handbook for Systematic Reviews of Interventions (Vol. 5)*. Chichester: Wiley-Blackwell.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing,* 27–35.

Robin M. Hogarth and Hillel J. Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1, 1–55.

Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. 2012. Domain bias in Web search. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM Press, New York, NY, 413–422.

Richard C. Jeffrey. 1990. *The Logic of Decision*. University of Chicago Press.

Thorsten Joachims. 2002. Optimizing search engines using click-through data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD '02)*. ACM Press, New York, NY, 132–142.

Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems (TOIS)*, 25, 2, Article: 7.

Alejandro R. Jadad, Deborah J. Cook, Alison Jones, Terry P. Klassen, Peter Tugwell, Michael Moher, and David Moher. 1998. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *Journal of the American Medical Association*, 280, 3, 278–280.

Anders W. Jørgensen, Jørgen Hilden, and Peter C. Gøtzsche. 2006. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: Systematic review. *British Medical Journal*, 333, 7572, 782.

Gabriella Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. In *Proceedings of the European Conference on Information Retrieval (ECIR '11).* Springer Berlin Heidelberg, 165–176.

Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3, 1-2, 1–224.

Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM Press, New York, NY, 193–202.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '08)*. ACM Press, New York, NY, 453–456.

Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the other way: Selective exposure to attitude-consistent and counter attitudinal political information. *Communications Research*, 36, 3, 426–448.

Carol C. Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42, 5, 361–371.

Dmitry Lagun and Eugene Agichtein. 2011. Viewser: Enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM Press, New York, NY, 365–374.

Annie Y. S. Lau and Enrico W. Coiera. 2007. Do people experience cognitive biases while searching for information? *Journal of the American Medical Informatics Association*, 14, 5, 599–608.

Annie Y. S. Lau and Enrico W. Coiera. 2009. Can cognitive biases during consumer health information searches be reduced to improve decision making? *Journal of the American Medical Informatics Association*, 16, 1, 54–65.

Raymond Y. K. Lau, Peter D. Bruza, and Dawei Song. 2004. Belief revision for adaptive information retrieval. In *Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. ACM Press, New York, NY, 130–137.

Carolyn Lauckner and Gary Hsieh. 2013. The presentation of health-related search results and its impact on negative emotional outcomes. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '13)*. ACM Press, New York, NY, 333–342.

Q. Vera Liao and Wai-Tat Fu. (2013). Beyond the filter bubble: Interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '13)*. ACM Press, New York, NY, 2359–2368.

Donald A. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. 1993. The unified medical language system. *Methods of Information in Medicine*, 32, 4, 281–291.

Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding Web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM Press, New York, NY, 379–386.

Brian Logan, Steven Reece, and Karen Spärck-Jones. 2004. Modelling information retrieval agents with belief revision. In *Proceedings of SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*. Springer London, 91–100.

David E. Losada and Alvaro Barreiro. 1999. Using a belief revision operator for document ranking in extended Boolean models. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM Press, New York, NY, 66–72.

Lee Browning Lusted. 1977. *A Study of the Efficacy of Diagnostic Radiologic Procedures: Final Report on Diagnostic Efficacy*. Chicago: Efficacy Study Committee of the American College of Radiology.

D. M. Mackay. 1960. What makes the question? *The Listener*, 62, 789–790.

Jennifer Mankoff, Kateryna Kuksenok, Sara Kiesler, Jennifer A. Rode, and Kelly Waldman. 2011. Competing online viewpoints and models of chronic illness. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '11)*. ACM Press, New York, NY, 589–597.

Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge: Cambridge University Press.

Sean Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '10)*. ACM Press, New York, NY, 1457–1466.

Margaret Ann Neale and Max H. Bazerman. 1990. *Cognition and Rationality in Negotiation*. New York: The Free Press.

Jakob Nielsen and Jonathan Levy. 1994. Measuring usability – preference vs. performance. *Communications of the ACM*, 37, 4, 66–75.

Vicki L. O'Day and Robin Jeffries. 1993. Orienteering in an information landscape: How information seekers get from here to there. *In Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 438–445.

Martin T. Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17, 11, 776–783.

Stuart Oskamp. 1965. Overconfidence in case-study judgments. *The Journal of Consulting Psychology*, 29, 3, 261–265.

Gerry Pallier, Rebecca Wilkinson, Vanessa Danthiir, Sabina Kleitman, Goran Knezevic, Lazar Stankov, and Richard D. Roberts. 2002. The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology*, 129, 3, 257–299.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1-2, 1–135.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5, 5, 411–419.

Mark Petticrew, Paul Wilson, Kath Wright, and Fujian Song. 2002. Quality of Cochrane reviews: Quality of Cochrane reviews is better than that of non-Cochrane reviews. *British Medical Journal*, 324, 7336, 545.

Natalia Pletneva, Uresova Zdenka, Jean-Jacques Altman, Vinay N. Postel, Patrice Degoulet, Jan Hajic, and Celia Boyer. 2013. Observations and lessons learnt from non-health professionals evaluating a health search engine. *Studies in Health Technology and Informatics*, 205, 940–944.

Vikas C. Raykar and Shipeng Yu. 2011. Ranking annotators for crowdsourced labeling tasks. In *Proceedings Advances in Neural Information Processing Systems (NIPS '11)*, 1809–1817.

Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. 2003. Trust management for the semantic web. In *Proceedings of International Semantic Web Conference (ISWC '03)*. Springer Berlin Heidelberg, 351–368.

David L. Sackett, William Rosenberg, J. A. Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. *British Medical Journal*, 312, 7023, 71–72.

Julia Schwarz and Meredith R. Morris. 2011. Augmenting web pages and search results to help people find trustworthy information online. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '11)*. ACM Press, New York, NY, 1245–1254.

William R. Shadish, Thomas D. Cook, and Donald Thomas Campbell. 2002. *Experimental and Quasi-experimental designs for Generalized Causal Inference*. Wadsworth Cengage learning.

Elizabeth Sillence, Pam Briggs, Lesley Fishwick, and Peter Harris. 2004. Trust and mistrust of online health sites. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '04)*. ACM Press, New York, NY, 663–670.

Herbert Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 99–118.

Herbert Simon. 1991. Bounded rationality and organizational learning. *Organization Science*, 2, 1, 125–134.

Robert S. Taylor. 1968. Question-negotiation and information seeking in libraries. *College and Research Libraries*, 29, 3, 178–194.

Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. 2008. To personalize or not to personalize: Modeling queries with variation in user intent. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM Press, New York, NY, 163–170.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5, 2, 207–232.

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

Thomas S. Wallsten and David V. Budescu. 1983. State of the art—Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 29, 2, 151–173.

Peter C. Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 3, 129–140.

C. Nadine Wathen and Jacquelyn Burkell. 2002. Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, 53, 2, 134–144.

Ryen W. White. 2013. Beliefs and biases in Web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM Press, New York, NY, 3–10.

Ryen W. White. 2014. Belief dynamics in Web search. *Journal of the Association for Information Science and Technology*, 65, 11, 2165–2178.

Ryen W. White and Ahmed Hassan. 2014. Content bias in online health search. *ACM Transactions on the Web*, 8, 4, Article: 25.

Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)*, 27, 4, Article: 23.

Ryen W. White and Eric Horvitz. 2010. Predicting escalations of medical queries based on web page structure and content. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM Press, New York, NY, 769–770.

Ryen W. White and Eric Horvitz. 2013. Captions and biases in diagnostic search. *ACM Transactions on the Web (TWEB)*, 7, 4, Article: 23.

Ryen W. White, Ian Ruthven, Joemon M. Jose, and C.J. van Rijsbergen. 2005. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems (TOIS)*, 23, 3, 325–361.

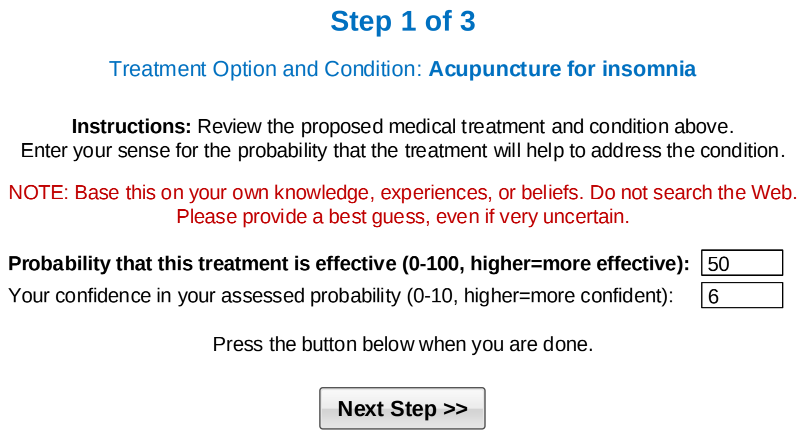
Janice C. Wright and Milton C. Weinstein. 1998. Gains in life expectancy from medical interventions: Standardizing data on outcomes. *New England Journal of Medicine*, 339, 6, 380–386.

Jinyun Yan, Wei Chu, and Ryen W. White. 2014. Cohort modeling for enhanced personalized search. *In Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*. ACM Press, New York, NY, 505–514.

Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. *In Proceedings of the 19th International Conference on the World Wide Web (WWW '10)*. ACM Press, New York, NY, 1011–1018.

Martina Ziefle. 1998. Effects of display resolution on visual performance. *Journal of the Human Factors and Ergonomics Society*, 40, 4, 555–568.

APPENDICES



Step 1 of 3

Treatment Option and Condition: **Acupuncture for insomnia**

Instructions: Review the proposed medical treatment and condition above.
Enter your sense for the probability that the treatment will help to address the condition.

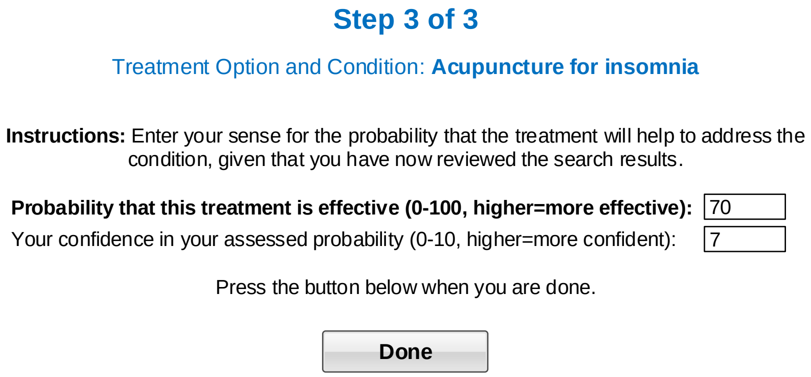
NOTE: Base this on your own knowledge, experiences, or beliefs. Do not search the Web.
Please provide a best guess, even if very uncertain.

Probability that this treatment is effective (0-100, higher=more effective): 50

Your confidence in your assessed probability (0-10, higher=more confident): 6

Press the button below when you are done.

Appendix A1. Interface for the elicitation of *pre-search* probabilities of treatment effectiveness (Phase 1).



Step 3 of 3

Treatment Option and Condition: **Acupuncture for insomnia**

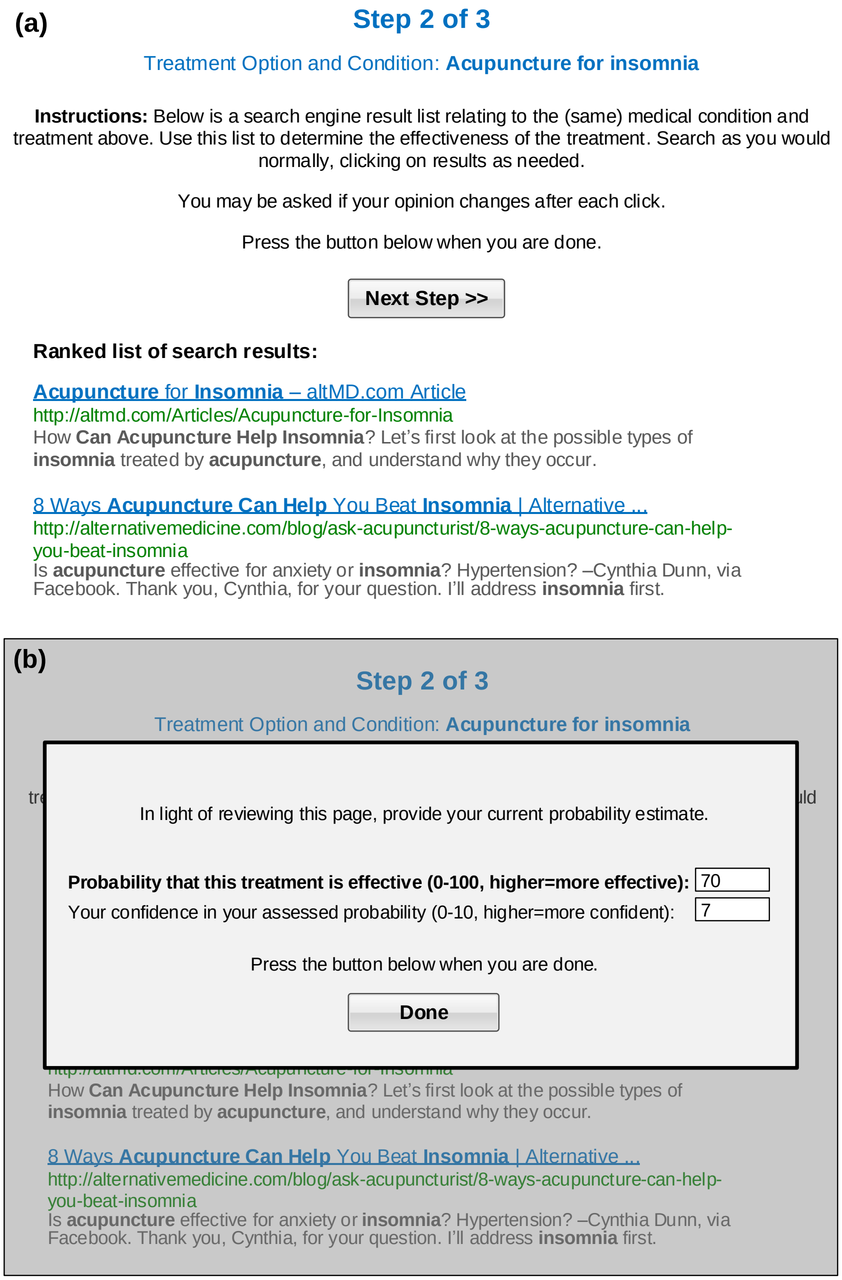
Instructions: Enter your sense for the probability that the treatment will help to address the condition, given that you have now reviewed the search results.

Probability that this treatment is effective (0-100, higher=more effective): 70

Your confidence in your assessed probability (0-10, higher=more confident): 7

Press the button below when you are done.

Appendix A3. Interface for elicitation of post-search probabilities of treatment effectiveness (Phase 3).



(a)

Step 2 of 3

Treatment Option and Condition: **Acupuncture for insomnia**

Instructions: Below is a search engine result list relating to the (same) medical condition and treatment above. Use this list to determine the effectiveness of the treatment. Search as you would normally, clicking on results as needed.

You may be asked if your opinion changes after each click.

Press the button below when you are done.

Ranked list of search results:

Acupuncture for **Insomnia** – altMD.com Article
http://altmd.com/Articles/Acupuncture-for-Insomnia
How **Can Acupuncture Help Insomnia**? Let's first look at the possible types of **insomnia** treated by **acupuncture**, and understand why they occur.

8 Ways **Acupuncture Can Help** You Beat **Insomnia** | Alternative ...
http://alternativemedicine.com/blog/ask-acupuncturist/8-ways-acupuncture-can-help-you-beat-insomnia
Is **acupuncture** effective for anxiety or **insomnia**? Hypertension? –Cynthia Dunn, via Facebook. Thank you, Cynthia, for your question. I'll address **insomnia** first.

(b)

Step 2 of 3

Treatment Option and Condition: **Acupuncture for insomnia**

In light of reviewing this page, provide your current probability estimate.

Probability that this treatment is effective (0-100, higher=more effective): 70

Your confidence in your assessed probability (0-10, higher=more confident): 7

Press the button below when you are done.

http://altmd.com/Articles/Acupuncture-for-Insomnia
How **Can Acupuncture Help Insomnia**? Let's first look at the possible types of **insomnia** treated by **acupuncture**, and understand why they occur.

8 Ways **Acupuncture Can Help** You Beat **Insomnia** | Alternative ...
http://alternativemedicine.com/blog/ask-acupuncturist/8-ways-acupuncture-can-help-you-beat-insomnia
Is **acupuncture** effective for anxiety or **insomnia**? Hypertension? –Cynthia Dunn, via Facebook. Thank you, Cynthia, for your question. I'll address **insomnia** first.

Appendix A2. (a) Search result list and (b) Solicitation popup shown following page review (Phase 2).