Diane Judith Rodden. What Presentation of Search Engine Results Do Health Information Searchers Prefer? A Master's paper for the M.S. in I.S. degree. January, 2019. 258 pages. Advisor: Dr. Robert Losee

A study of a sample of online health information searchers was conducted to see what their preferences are with respect to four different display styles for search engine results on health topics. Screen shots of search result display screens were presented to the participants via a Qualtrics (www.qualtrics.com) online survey. The other display types were Display 1: Google standard display, Display 2: Google enhanced with faceted browsable categories, Display 3: Google enhanced with a word cloud for each search result, and Display 4: Google enhanced with an overview word cloud for collection of search results. For each search task, participants were asked to rate the search engine results displays for quality indicators, using Likert-type item rating scales. At the end, in three concluding questions, the participants were asked to choose the display(s) that were best at meeting three specific criteria, based on overall impressions. The evaluations by the participants suggest that the standard Google search results display and the Google screen enhanced with faceted browsable categories were favored over the other two display types.

Headings:

Health information searching

Tag cloud

Search engine results

Query refinement

Information retrieval by facet

Visualization of search results

WHAT PRESENTATION OF SEARCH ENGINE RESULTS DO HEALTH INFORMATION SEARCHERS PREFER?

by
Diane Judith Rodden

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

January, 2019

Approved by:

____

Advisor

Table of Contents

Acknowledgements

Thank you to my master's paper advisor, Dr. Robert Losee, for his guidance and encouragement.

This paper is dedicated to my parents, John and Janet Rodden, and to my brother, Paul Rodden, for the constant understanding, love, and support that they have provided throughout my life.

What Presentation of Search Engine Results Do Health Information Searchers Prefer?

"The true test of the performance of search engines is their ability to satisfy their users." (Agrawal, Gollapudi, Halverson, & Ieong, 2009, p. 13).

Adults of all demographic groups use the Internet to search for health information. For instance, 65% of U.S. male Internet users and 79% of U.S. female Internet users have used the Internet to look for health topics (Fox & Duggan, 2013). Surprisingly, among U.S. Internet users ages 18-29, 76% search for health information online; among U.S. Internet users age 65 and over, only 58% search for health information online (Fox & Duggan, 2013).

Search engines are the most popular way to search for health information. 13% of Americans begin their search with a dedicated health information site, such as WebMD, but 77% launch their search from a general-purpose search engine like Microsoft Bing (formerly Live Search), Yahoo!, or Google (Fox & Duggan, 2013). And, in the age group 18-29, which is the group most likely to search for health information, 82% start searching with a general-purpose search engine (Fox & Duggan, 2013).

How the search results are presented is important to how the users are able to efficiently and accurately select relevant documents from the search results (Gwizdka, 2009). In major search engines, such as Bing and Google, the traditional search engine results page includes a ranked list of the results; with each result represented as a document summary, the page title, and the URL (Al-Maqbali, Scholer, Thom, & Wu, 2010; Capra, Arguello, & Scholer, 2013; Gwizdka, 2009; Kelly & Azzopardi, 2015).

Although this traditional presentation is common, it might not be the best display for user satisfaction (Fuenzalida, & Martinez-Ramirez, 2017; González-Ibáñez, Proaño-Ríos; Kelly & Azzopardi, 2015). In fact, health information consumers who use general-purpose search engines instead of a government health site, a commercial site like WebMD or a university medical site, report higher levels of effort and frustration during their health searches (LaValley, Kiviniemi, & Gage-Bouchard, 2017). Yet, Heo and Hirtle warn of the difficulties in visualizing search engine results, such as how to identify how much and what types of data to include in the results to achieve the goal of providing more data without providing too much distracting data (2001).

Given the prevalence of health information searching that is done through the general-purpose browsers, studying how to present the search results in a way that optimizes the search experience could produce useful findings. For instance, if an improved display for health search results is identified, a browser extension or personalization of the search results could be implemented to assist health information searchers (Aula, Khan, Guan, Fontes, & Hong, 2010).

Research Question

When searching for health information on the Internet using a general-purpose search engine, how do users prefer to see the results presented out of the following search result display screens:

- Standard Google (this is the control);
- Google enhanced with faceted browsable categories;
- Google enhanced with a word cloud for each search result; or
- Google enhanced with an overview word cloud for the collection of search results?

Google was picked as the basis for the search engine results page designs because it is familiar and trusted. When Jansen, Zhang, and Schultz conducted a study of various search engines by brand, users scrutinized the search results less and expressed more confidence in the search results from Google and Yahoo! (2009).

Glossary

General-purpose search engines are types of information retrieval systems that index documents throughout the Internet and retrieve relevant documents based on the query terms input by the user (Atsaros, Spinellis, & Louridas, 2008).

Ranked list is a presentation of the search engine results as a display of the documents that are found by the search engine to be relevant to the search terms entered, where the results are listed in an order determined by a mathematical algorithm. Google, the most common search engine worldwide (Humenberger, 2011), displays results as a ranked list.

Faceted browsable results refers to display of the search results as categories with "…meaningful labels organized in such a way as to reflect the concepts relevant to a domain" (Hearst, 2006, p.60).

According to the Oxford Dictionary of Journalism, a word cloud is "a graphic representation of the frequency with which certain words are used in a speech, statement, document, judgement, or similar. The higher the frequency, the bigger the depiction." (Harcup, 2014). Word clouds can also add different fonts or different font colors to enhance the graphic representation.

Review of the Literature

Theoretical Framework

It might be useful to first briefly consider a theoretical framework which can inform our understanding of information seeking. A central concept of information retrieval is that an information need is the motivation for an information search (for example, see Broder, 2002 or Shneiderman, Byrd, & Croft, 1997). An information need may start as a vague affective state of discontent that may not even be expressible as a question or problem (Taylor, 1968). Because an information need is a difference between what is already known and what will be acquired, it is unique to each individual (Perry, 1963).

For electronic information retrieval to occur, the information need will need to be transformed from an ambiguous thought to a concrete question, and ultimately, modified to a query suitable for a search engine (Perry, 1963; Taylor, 1968). Information foraging theory positions information searching as a task where the searcher will weigh the anticipated information gain of using a given information source against the expected time, effort, and energy that will be expended to decide the value of undertaking or continuing the search, much like an organism weighs the value of pursuing a food source (Pirolli & Card, 1995). Electronic information retrieval is often a multi-step, iterative process (M.J. Bates, 1989; Marchionini, 1995). For instance, some of the possible steps described by Marchionini include the user reviewing the search results to decide if the items meet his or her expectations, examining a document, looking at other results in the list, and modifying the query (1995). Bates describes a berrypicking model in which information is collected, the original problem or task is reconceptualized, and then, more pieces of information are gathered (1989). Eventually an outcome is reached in which the searcher finds information to address the information need, partially or fully; or the user gives up and stops (Marchionini, 1995). Feild, Allen, and Jones define search satisfaction “as the fulfillment of a user’s information need” (2010, p.35).

Internet Use, Searchers, and Search Behavior

Next, it may be helpful to review some demographic information of Internet users, along with some background literature on searching. Men and women are using the Internet at the same rates, with 81% of all American men and 81% of all American women reporting that they are Internet users (Fox & Duggan, 2013). In the United Kingdom, the digital divide between men and women has also disappeared, with equal numbers of each sex utilizing the Internet (Dutton & Blank, 2013). In the 18-29 age group, 95% of Americans are Internet users (Fox & Duggan, 2013). By race and ethnicity, there are still some gaps in Internet use as white Americans have an 83% rate of Internet use, while African Americans and Hispanics have 74% and 73% rates of use, respectively (Fox & Duggan, 2013). Fox and Duggan found a large divide in Internet use by education, with only 47% of adults without a high school diploma reporting Internet access, but 96% of college graduates stating that they are Internet users (2013).

From 2013 to 2015, the percentage of Americans with household broadband has declined, as some households switched to smartphones for their Internet access instead, due to cost of service or cost of computers (Horrigan & Duggan, 2015). According to Horrigan and Duggan, some of the issues faced by those having Internet access only through a smartphone are inability to use the Internet adequately after the data-cap is exceeded for the month and disconnection of service when financial difficulties arise for the household (2015). For instance, 62% of African Americans had broadband service in their homes in 2013, but in 2015, only 54% of African Americans have broadband Internet home service, with a 9-percentage point increase in African Americans that are now smartphone users for all Internet access (Horrigan & Duggan, 2015).

On average, as of 2016, around 40% of senior citizens are smartphone users, 51% have broadband Internet in the homes, and 34% are social media users (Anderson & Perrin, 2017). However, the senior citizen demographic spans multiple decades of ages and technology use varies widely by age group, ranging from almost 60% of adults ages 65-69 using smart phones to only around 15% of those over 80 using smartphones, according to the survey work done by Anderson and Perrin (2017).

Despite the widespread use of the Internet, evidence about user success in searches is conflicting. One problem seems to be short queries. In a study of web searching patterns of students in seven classes at a U.S. college, vague and short queries were often used by the students (Nowicki, 2002). U.S. web searchers use one-word queries in 20-29% of their queries, while Europeans use one-word queries 25-35% of the time (Jansen & Spink, 2006). Consistent with the findings of Jansen and Spink, in a study of German university students who were asked to recall and write down the last query for non-academic purposes that they input into a general-purpose search engine, the queries were found to be one-word queries 32.5% of the time and two-word queries 30% of the time (Lewandowski, 2008).

Frustration is another problem. In one study involving university students, searchers reported frustration in 50% of all queries, including 33% of the queries where they successfully completed the search task (Feild, Allan, & Jones, 2010). The students listed the following as causes of frustration during the search tasks performed in the study: "(1) off-topic results, (2) more effort than expected, (3) results that were too general, (4) un-corroborated answers, and (5) seemingly non-existent answers." (Feild, Allan, & Jones, 2010, p.37). One participant in a study on querying and relevance of search results recounted his or her frustrating experience during performing a study task where results were centered around hardware repair under warranty, when the desired context was upgrading existing hardware (Patil, Alpert, Karat, & Wolf, 2005). White and Dumais report that, based on search engine logs, the population of users who switch search engines either during or between a search engine may exceed 70% of users (2009). Based on survey responses from search engine switchers, 10% had switched due to frustration, 24% had switched due to dissatisfaction, and 23% were attempting to achieve better results (White & Dumais, 2009). A combined total of 57% of switches were due to perceived lack of success with the search (White & Dumais, 2009). One study looked at the effect of search engine branding (Google, Yahoo!, and Microsoft Live/Bing) on the ratings that users assign to the quality of the search results and did not find any effect of branding on quality ratings (Bailey et al., 2010).

Additionally, search success can be hampered by gaps in searching skills, lack of topic knowledge, technical difficulties, and other factors (Savolainen & Kari, 2006). Wirth, Sommer, von Pape, and Karnowski found that search expertise was predictive of successful searches in tasks involving finding sources of relevant information on a topic (2016). Search success in college students was found to be highly correlated with frequent use of search engines and the Internet (Nowicki, 2002). In a study of South Africans, Blignaut and McDonald found that increased web experience did not improve the likelihood that participants with lower socioeconomic status would find the correct information when conducting Internet searches, although the researchers theorize that lower socioeconomic users might need additional web experience above and beyond the web experience that was considered in this study (2012).

Yet, Yahoo! and Google perform quite well in experimental studies, with one study finding that in the top 20 results, both search engines returned relevant documents 48% of the time and when looking only at the top ranked result, it was relevant around 74% of the time (Lewandowski, 2008). User satisfaction (based on user assessments of relevance, accuracy, coverage, and ordering of the results) appears to be correlated with Google effectiveness, as calculated by precision and cumulative gain (Al-Maskari, Sanderson, & Clough, 2007). In a survey of residents of the United Kingdom, 72% of retired adults stated that they believe that they have the necessary skills to evaluate quality and credibility of information on websites, compared with 82% of residents still in the workforce (Dutton & Blank, 2013). And, as users get more topic knowledge, their searching techniques change to narrower query terms (Wildemuth, 2004).

Health Information Searchers

About 30% of Americans also state that using medical information from the Internet has substantially helped them personally or someone that they know (Fox & Rainie, 2002). Among American Internet users, 80% of American Internet users have used the Internet at least once (but not necessarily within the last year) to search for health information online (Fox, 2011), with 72% having looked for health information online within the last year (Fox & Duggan, 2013). In another study, 78% of Internet users reported having used the Internet to search for health information (LaValley, Kiviniemi, & Gage-Bouchard, 2017). In 2012, 31% of U.S. cell phone users reported searching for health information by cell phone (Fox & Duggan, 2013).

In adults who are Internet users, 73% of white Americans, 69% of African Americans and 66% of Hispanic Americans look for health information on the Internet (Fox & Duggan, 2013). African Americans, Hispanics, and white Americans are all equally likely to have used the Internet to search for health or medical information (Fox & Rainie, 2002).

Income does have an impact on health information searching. In U.S. Internet-using households, 80% of households with income of $75,000 and over use the Internet for health information, but only 65% of households with income of $30,000 and under go online for health information (Fox & Duggan, 2013). And, when Americans cancel their broadband due to financial constraints (Fox & Duggan, 2013), it can impact their ability to access health information. In 2015, 66% of Americans surveyed reported that not having a broadband Internet connection at home would interfere (either a lot or a little) with obtaining health information (Horrigan & Duggan, 2015). Moreover, 38% of Americans who do not have a broadband Internet connection at home state that it does interfere a lot with health information searching (Horrigan & Duggan, 2015).

The United States is not the only country utilizing the Internet for health information. For instance, AlGhamdi & Moussa studied health information searching in Saudi Arabia, finding that among Internet users, almost 60% used the Internet to search for health information, with more frequent health information searchers being high-income or university-educated (2012). In Australia and New Zealand, a telephone survey found around 17% of respondents use the Internet to search for health information (Gauld & Williams, 2009). In a 2015 survey in Vietnam, the most frequent sources of health information among adults ages 18-60 were health television shows (used by around 50%) and the Internet (used by around 32%) (Nguyen, Nakamura, Seino, & Vo, 2017). In a survey of residents of the United Kingdom, 69% of respondents used the Internet to search for health information (Dutton & Blank, 2013).

Health Information Search Behavior

Fox found that health topics of interest to e-health information users include specific medical conditions, specific medical treatments, health insurance, dementia, nursing homes, and medication safety (2011). About 48% of the health information searchers are looking for information that pertains to the medical condition(s) of a friend or family member, not themselves (Fox, 2011). Mental health information, information to prepare for an appointment with a medical professional, and alternative medicine were also of interest to close to the majority (Fox & Rainie, 2002).

Almost three-quarters of online health information consumers use general-purpose search engines (eg. Google, Microsoft Bing/Live, or Yahoo!) when searching for health information on the Internet (Fox & Duggan, 2013; Stvilia, Mon, & Yi, 2009). Most health web searchers select a site from the first 3-5 search results and rarely check the next page of search results (Feufel & Stahl, 2012; van Deursen, 2012). Fox and Rainie reported similar results indicating that health information consumers visit around 2-5 sites after using a general search engine or portal to launch their search (2002). Health information searchers who were older, with less searching expertise, or with less education used more one-word queries when compared with a group of searchers who were younger, more experienced, and better educated (Feufel & Stahl, 2012). Similarly, van Deursen found that less educated e-health users were more likely to use overly general or one-word queries (2012). Eysenbach and Köhler also reported that 65% of participants in their study used one-word queries instead of better, more-specific queries (2002). In one study, 10% of participants did not separate their query terms with spaces when using a general-purpose search engine to find answers to health questions (van Deursen, 2012). Women are more likely than men to have a health information site that they prefer to use and to be users of government health sites (LaValley, Kiviniemi, & Gage-Bouchard, 2017).

Certifications and Other Attempts to Guide Health Consumers

Health On the Net Foundation (HON) is a non-profit international organization that promotes Internet health information standards and offers certification to sites that meet HON standards (Health On the Net Foundation (HON), 2017). DISCERN is a national consumer health information quality standard developed by a panel of UK health professionals, consumer health advocates, and health journalists (Charnock, Shepperd, Needham, & Gann, 1998; Charnock, Shepperd, Needham, & Gann, 1999). DISCERN uses features such as citation of references, presenting information without bias, and currency to help health information consumers evaluate quality in health information (Charnock, Shepperd, Needham, & Gann, 1998; Charnock, Shepperd, Needham, & Gann, 1999). Similarly, the European Union created quality standards for health websites, with guidelines about principles such as authority, privacy, and transparency (Commission of the European Communities, Brussels, 2002). An example of a government health search site is PubMed. PubMed is an index of over 28 million citations of biomedical and health information (National Center for Biotechnology Information [U.S.], 2018). PubMed is managed by the National Center for Biotechnology Information at the U.S. National Library of Medicine.

Quality and Credibility of Health Websites

Studies of health and medical websites have found a range of quality and credibility. Credibility can be defined as believability (Fogg et al., 2001). Researchers who looked at 200 websites returned by a Google search on the search term "antioxidants" found that although commercial sites were frequently found in the 200 search results, Google displayed more government and health portal sites in the top ranked sites; these government and health portal sites had high quality scores when evaluated by using the Journal of the American Medical Association (JAMA) health criteria (Aslam, Gibbons, & Ghezzi, 2017). Additionally, many of the sites in the top Google results from Aslam, Gibbons, and Ghezzi had Health-on-the-Net (HONCode) certification (2017). Bedell, Agrawal, and Petersen looked at the content, usability, and reliability of websites providing diabetes information, finding that only 17% of the sites met established reliability criteria, while 60% of the commercial sites featured advertising for sales of a specific product (2004). Additionally, Bedell, Agrawal, and Petersen found that usability was hampered on several sites due to excessive advertising content (2004). In looking at 165 published studies in which quality of health information websites were assessed by experts, common criteria used in many of the 165 studies to evaluate quality were accuracy, completeness, authoritativeness, trustworthiness, inclusion of references, listing of authors, and having up-to-date information (Y. Zhang, Sun, & Xie, 2015). Y. Zhang, Sun and, Xie determined, based on their meta-study of 165 studies, that over 55%

of the authors identified quality problems with online health information, while 37% of the studies found that quality of ehealth information is of variable quality (2015).

In a study of breast cancer information on the web, Bernstam et al. found a much higher of accuracy than previous authors, with only 5.2% of 343 websites having inaccurate information (2008). Bernstam et al. also found that sites that contained alternative medicine information about breast cancer had an increased likelihood (15.6 times more likely) to provide inaccurate information and that using well-publicized guidelines to evaluate health websites did not screen out sites with inaccurate information (2008). In a sample of websites offering content on probiotics, websites were found to often exaggerate health benefits, overstate the benefits, and minimize the risks (Brinich, Mercer, & Sharp, 2013). These probiotic biases are more common in commercial sites (Brinich, Mercer, & Sharp, 2013). When 51 gastric cancer information sites returned as top results by general-purpose search engines were assessed for quality, the sites, in general, were found to be too commercial with incomplete information provided; websites in the top 10 search results were found to be of high-quality, with good accuracy, and complete (Killeen et al., 2011). Kitchens, Harle, and Li evaluated the quality of 5,249 webpages found by using Google to conduct a query for 2,069 health topics, finding that health website quality varies by category of health topic, with medical conditions and diseases having better quality information than categories such as alternative medicine, wellness and poisons/toxins (2014). One limitation of the work of Kitchens, Harle, and Li, which they acknowledged in their paper, was that they defined quality very narrowly, with the only way that a website could be judged as high quality would be to have a HON or similar certification (2014).

The problem of inaccurate health information online is not limited to the United States. In a study on epilepsy information available through top Chinese search engines using Chinese query terms, 18.8% of the information was inaccurate and 30.8% of the information was accurate but incomplete (J. M. Liu et al., 2015). When 20 Romanian websites on the topic of first aid for choking were assessed for quality, the average credibility score was 4.90 out of 14 using the eEurope 2002 standard credibility measures, with problems on the choking first aid sites such as not listing authors, not disclosing financial interest, and lack of references (Nadasan, Vancea, Georgescu, Tarcea, & Abram, 2011). Chang, Hou, Hsu, and Lai reviewed search engine results from the Taiwanese versions of Yahoo! and Google in response to health topic queries, finding that the top 50 quality award winning health websites in Taiwan rarely were displayed within the first 100 search engine results (2006). One study found that in querying for epilepsy information online using search terms in Chinese, 80.2% of the retrieved information had content that was relevant to the query terms (J. M. Liu et al., 2015). In looking at 75 Portuguese Brazilian language websites on dental caries, researchers found that when reviewed by experienced professionals, a large portion of the websites were of poor quality when examined in accordance with JAMA (Journal of American Medical Association) and DISCERN (Charnock, Shepperd, Needham, & Gann, 1998; Charnock, Shepperd, Needham, & Gann, 1999) established health website standards (Aguirre et al., 2017).

Dangers of Poor Quality Health Information Online

Erroneous or misleading health information can be dangerous in multiple ways. Health consumers diagnosed with cancer, due to their vulnerable emotional status, can fall victim to websites and social media sites that sell fraudulent and ineffective cures for cancer (U.S. Food and Drug Administration, 2018). When confronted with a difficult health decision with no clear evidence and/or under the cognitive influences of poor health, a health consumer can make decisions guided mostly by emotion (Peters, Lipkus, & Diefenbach, 2006). Almost 96% of online pharmacies do not meet U.S. federal laws, state laws, and/or U.S. National Association of Boards of Pharmacy professional pharmacy safety practices, with 74% of online pharmacies claiming to operate from Canada found neither to be located in Canada, nor to be legal in Canada (National Association of Boards of Pharmacy, 2017). The FDA warns consumers that fraudulent online pharmacies expose consumers to unregulated, ineffective dosages, or unknown dangerous substances (2018).

Also, e-health users might not expect health information to be misleading. For instance, 69% of U.S. health information consumers indicated that they had never encountered medical information on the web that was incorrect or untruthful (Fox & Rainie, 2002). Furthermore, in a survey, around 70% of U.S. health information consumers expressed the belief that most online information is correct and trustworthy (Fox & Rainie, 2002). But, 3% of Americans state that using health information from the Internet has been harmful to themselves or someone they know (Fox, 2011). Surprisingly, 30% of Americans have used the Internet to make a do-it-yourself diagnosis for their symptoms or those of a family member or friend, with over one-third stating that they did not follow-up with a medical professional (Fox & Duggan, 2013).

Internet Health Information Credibility Assessment by Users

There is a large body of literature on the capabilities and inclination of health information consumers to effectively judge Internet health information, but the results are mixed and also point to differences in the credibility criteria viewed as important by e-health consumers. The results of Cunningham and Johnson suggest that e-health users do have sufficient knowledge and inclination to evaluate health information online, as participants used cues such as reputation, authority, completeness, clarity, freedom from bias, terminology, word usage, and website design to judge the trustworthiness of sites (2016). Brady, Segar, and Sanders found that online health forum members listed articulateness and logic as markers for users whose posts would be viewed as credible (2016). The work of Robins, Holmes, and Stansbury identified good health website design as significantly associated with higher ratings of a health website's credibility by participants (2009). Also, site design features, such as listing the contact information for the organization operating the site, a menu, a privacy policy, and links to external sites, were identified by Rains and Karmikel as increasing the likelihood that users would rate the credibility of health websites positively (2009). Freeman and Spyridakis noted that participants showed a strong inclination to use author credentials, organizational authority, and reputation to evaluate online health information (2004) and in another study, participants most frequently cited author authority as most important (Stvilia, Mon, & Yi, 2009). Health information consumers with high health literacy and lower health literacy mentioned using established quality criteria to assess health websites, such as checking that the content authors are listed and verifying that a medical expert authored the information (Diviani, van den Putte, Meppelink, & van Weert, 2016).

A 2005 study by B.R. Bates, Romina, Ahmed, and Hopson found that users asked to compare health information from a well-respected health source and unnamed sources did not rate the information from the unnamed sources as any less trustworthy or truthful (2006). Health information users may even interpret the Google ranking of the search results as an indicator of the trustworthiness of health websites (Diviani, van den Putte, Meppelink, & van Weert, 2016). Completeness was found to be an important marker of health website credibility by Dutta-Bergman, with users more likely to rate a test health website as credible when shown the version with “complete” information, as opposed to the versions with “jargon” information or “incomplete” version (2004). Another study found that more difficult terminology led to higher credibility assessments of online health information (Freeman & Spyridakis, 2004). When healthcare consumers were surveyed about what identifiers were most important in health sites, accuracy, reliability, credibility, trustworthiness, clarity, objectivity, utility, and verifiability were ranked the highest, out of 20 identifiers (Stvilia, Mon, & Yi, 2009). Website attributes that made e-health consumers more likely to assess the website as less credible include sources that are not identifiable, lack of a health certification indicator, lack of information on currency of the information, presence of a shopping cart, and inclusion of too much advertising or pop-up ads (Fox & Rainie, 2002).

Presence of advertising on online health information sites may have unusual effects on user assessment of the website’s quality. Advertising on sites with .org domains negatively impacts credibility, but advertising on sites with .com or .edu improves a website’s credibility (Walther, Wang, & Loh, 2004).

In a study by Crystal and Greenberg, when health information users were asked to highlight (with a computer mouse) which details were used by them to determine the relevance of search engine results (document summaries and documents themselves), the "key criteria identified included (in order of frequency of appearance) research, topic, scope, data, influence, affiliation, Web characteristics, and authority/person" as determined by a content analysis of user highlights (2006). "Scope" included highlights demonstrating the reading level/audience and geographic region of the document summary or document, while "Web characteristics" included highlights of the links to other documents and file type of the document (Crystal & Greenberg, 2006). The authors note that the study participants frequently highlight terms related to the attribute "research", which included the methods and population sample discussed in the medical websites, but the attributes "affiliation" and "authority/person" were among the least frequent criteria used to assess relevance, so the interplay of "…detailed investigation of research methods and limitations, on the one hand, and implicit acceptance of the credibility and authority of Web documents, on the other, deserves further attention." (Crystal & Greenberg, 2006, p.1379).

Although not specifically about health information sites, the work of Haas & Unkel (2017) on user assessment of credibility is relevant to this literature review. When participants were asked to assess the credibility of individual websites in the search engine results, it was noted that participants gave higher credibility scores to institutes and new sites, with lower credibility given to blogs and corporate websites (Haas & Unkel, 2017). The participants from the Haas and Unkel study assigned higher credibility to a search result if it mentioned a scientific study in the document summary (2017).

Over 1,400 Americans and Finish volunteers answered questions about various website attributes to see which ones they associated with website credibility in a study by Fogg et al. Website typographical errors and other quality control problems were frequently noted as signs of a low credibility site, with this effect especially pronounced in study participants less than 28 years old (Fogg et al., 2001). Mixing of commercial and advertising content with the informational content of a site is perceived by users as an indicator of low credibility in a website (Fogg et al., 2001). Attributes associated with high credibility include showing photos of employees, listing a physical address, usability, navigability, listing credentials, listing references, including links to sources, and personalization (Fogg et al., 2001).

Purported authority of the authors does not seem to have a strong effect on credibility assessments. When some participants were shown that the author of a health website was a medical doctor while some were shown that the author of the health website was a lay individual, there were no significant differences detected in how the participants rated the credibility or their willingness to implement the suggestions on the website (Hu & Sundar, 2010). In the same study, participants indicated that they would be more likely to act on information from a Web page than from a blog or a personal home page but did not indicate any differences in how they would view the credibility of a Web page, a blog, or a personal home page (Hu & Sundar, 2010). There were no significant differences in how participants rated the credibility of a nutrition website purported to be from the CDC and how participants rated the credibility of a nutrition website purported to be a personal blog (Jung, Walsh-Childers, & Kim, 2016).

Eysenbach and Köhler noted that online health information consumers stated that when searching for health information online, they assessed the quality of the information as part of the search process by checking for indicators such as source authority, third party endorsements, and professional layout of the site, but Eysenbach and Köhler then discovered that, in practice, these same participants did not consistently check the source and authority when performing search tasks (2002). Participants in one study frequently failed to verify the quality and credibility of the websites they used to answer health questions (Feufel & Stahl, 2012), while 63% of participants in another study utilized irrelevant webpages and 36% used websites that were not credible (van Deursen, 2012). Fox and Rainie found that around 50% of health information consumers rarely review the quality and credibility of health website (2002). Around 25% of health information users are diligent about checking quality and credibility each time that they are using the Internet for health information, while the remaining 25% review health sites for quality and trustworthiness only sometimes (Fox & Rainie, 2002).

Effects of Age, Gender, and Socioeconomics on Credibility Evaluation

Some authors have examined the effects of age on assessing online health information credibility. Older users are more likely to disregard signals of low credibility (such as missing HON Health on the Net Network certifications or missing references to sources) but are less influenced by website reviews that promote low quality sites as being valuable (Liao & Fu, 2014). However, van Deursen found that senior citizens were less likely to use irrelevant and low-quality health websites to answer health related questions (2012). Age decreases a user's likelihood of reviewing the authority of a health website (Gauld & Williams, 2009).

Having a post-secondary education was associated with increased frequency of reviewing the credibility of an Internet health information site, according to a study of Australians and New Zealanders (Gauld & Williams, 2009). Benotsch, Kalichman, and Weinhardt conducted a study in which the study participants were HIV patients with varying levels of literacy, income, and education (2004). The HIV patients were asked to evaluate the credibility of a Journal of the American Medical Association (JAMA) HIV information page and a page of doubtful authority describing HIV treatment with components from goats. Benotsch, Kalichman, and Weinhardt observed that HIV patients with lower education, lower literacy, lower income, and lower HIV health knowledge rated the goat page as high in credibility, while participants with higher literacy and more HIV health knowledge rated the JAMA HIV information page as high quality (2004).

Diviani, et al. (2016) noted that health information consumers with lower health literacy used some additional non-standard evaluation criteria for health websites. Health information searchers may create their own criteria to help predict quality of health websites, such as whether they think the website is popular, whether the information has some jargon to demonstrate that experts have been involved, and whether the website abstains from promoting an ideology (Diviani, van den Putte, Meppelink, & van Weert, 2016).

Men and younger adults were found to be especially sensitive to a bad website design as a cue for poor credibility of a health website (Fox & Rainie, 2002). One study found that men placed less trust in Internet health information than females, but the difference was not significant (Gauld & Williams, 2009). Older users see markers that convey trustworthiness and expertise, such as listing credentials, listing references, and including links to sources, as signs of credibility in a website (Fogg et al, 2001). Men, overall, rate attributes of website credibility more critically than women (Fogg et al., 2001).

Tao, LeRouge, Deckard, and De Leo argue that the heuristics that a healthcare consumer uses to evaluate quality of a health website vary by the type of healthcare consumer (2012). For instance, Tao, LeRouge, Deckard, and De Leo found that healthcare consumers with an educational background in healthcare were more likely to consider level of detail, relevancy, and completeness as important indicators of health website quality when compared with healthcare consumers with an educational background in business (2012). Also, consumers with a business background felt that links to medical referrals should be provided by quality sites, whereas consumers with a healthcare background felt that external links should instead be given for more detailed information (Tao, LeRouge, Deckard, & De Leo, 2012). Personalization is especially important to higher income users as a marker of credibility (Fogg et al, 2001).

Emotional Factors Used in Internet Health Information Evaluation

Cunningham and Johnson found that participants also judged health information by whether they could relate or identify with it (2016) but the authors didn’t explore whether emotionally connecting to the narrative on the website was associated with actual accuracy and credibility of the website. Additional authors have also found that users of online health forums appraised health information as more trusted and credible if the user posting the information had similar health symptoms and life experiences as their own (Brady, Segar, & Sanders, 2016; Costello, 2015). Wang, Walther, Pingree, and Hawkins expanded on the concept of emotional connection and shared experience, reporting that online health information consumers prefer information from patients and other lay authors, rather than medical experts, in both online forums and on traditional websites (2008). Wang, Walther, Pingree, and Hawkins then proposed a model in which similarity of the author to the reader is what impacts the credibility assessment of the information and what causes the online health information consumer to act on the information (2008). An experiment by Flanagin, Hocevar, and Samahito observed that participants who felt more similarity with other users of an online (non-health-related) social media site felt the user-generated content was more credible and were more likely to indicate that they would follow the advice of the forum (2014).

Emotional components other than similarity and commonality have been mentioned by participants in e-health information studies. For instance, in one study, participants mentioned that a quality Internet health website should have a photo of a likeable person or a photo of a likeable site owner (Eysenbach & Köhler, 2002). An empathetic tone to the website was mentioned as a marker of a quality health website by another user (Walther, Wang, & Loh, 2004). When asked to answer health questions that involved weighing two courses of action, participants in a cohort of older, less educated, and less experienced searchers were more likely than an experienced, younger, and more educated cohort to select search engine results that reflected their own personal opinions instead of looking for information written by experts on the topic (Feufel & Stahl, 2012). For a thorough discussion of the influences of affect on health information behavior, judgement, and motivation, see the paper by Peters, Lipkus, and Diefenbach (2006).

Subjective feelings, such as the way that a site makes the user feel, and consistency with the user's opinions were listed by Italian-speaking health information consumers as examples of evaluation criteria that they have utilized to evaluate Internet health websites (Diviani, van den Putte, Meppelink, & van Weert, 2016). Fox and Rainie also noted that health consumers seem to like reading the same information about their medical conditions that they already know (2002). While it often confirms the pre-existing experiences of the health consumers, they also learn something new in many cases (Fox and Rainie, 2002).

Freeman and Spyridakis found that greater interest in the topic of the health information led to higher ratings of credibility (2004). Rains & Karmikel also studied interest in the health information topic on a website, finding that references, listing the name of the article's author, providing numeric statistics or values, and including testimonials or stories were all associated with more positive attitudes toward the topic of the health information site (2009).

Effects of Bias on Health Information Credibility

Biased medical information can change the medical attitudes and knowledge of health consumers (Allam, Schulz, & Nakamoto, 2014). In the study by Allam, Schulz, and Nakamoto, Google's searching algorithm was manipulated to either return results from authoritative and objective vaccination information from reputable sites or to return results from sites with anti-vaccination communication intentions (2014). After viewing the anti-vaccination search results, participants had lower knowledge levels of vaccination literacy and more concern about vaccination detrimental effects than expressed on their pre-search vaccination literacy assessment (Allam, Schulz, &

Nakamoto, 2014). Participants who viewed the anti-vaccination search results were likely to recommend these sites and did not detect that the sites were low-quality (Allam, Schulz, & Nakamoto, 2014).

Fox and Rainie noted that over 85% of health website consumers felt that sites that had information agreeing with other sites were more credible (2002). However, agreement between sites may lead to a false sense of security if the sites are all repeating the same information from a common source (Fox & Rainie, 2002). When a search engine presents multiple sites with similar content, some health information searchers consider this overlap to be an indicator of the credibility of the information (Diviani, van den Putte, Meppelink, & van Weert, 2016).

In their 2014 paper, White and Hassan present findings from their study which suggest that search engine algorithms might have the tendency to return results biased toward positive outcomes in health searches through a few mechanisms. For instance, if health consumers incorporate the terms “can” or “help” in their query, such as “can X help with Y condition”, then pages frequently containing either of those words (and indicating successful treatment) will likely appear near the top of the search results, even though a large sample of health documents indicates that the distribution of the health corpus is 33% positive, 33% inconclusive and 33% negative (White & Hassan, 2014). Another factor that contributes to the bias is that users often click on the top results, so this makes documents that matched terms like “can” and “help” continue to be displayed in top ranked positions as their popularity grows with more clicks (White & Hassan, 2014). Results indicating treatments that do not work and inconclusive results are often displayed in lower ranked positions in the search engine results which do not get viewed by the searchers even if they are the accurate results (White & Hassan, 2014).

Health information and Social Media

Web 2.0, the next generation of the Internet, includes user-generated content and social media, along with expanded types of media, such as video (Christensson, 2008) and can include health content. Users with higher than average interest in using social media to connect with users who have similar medical conditions include people who have developed a new medical condition, people who have chronic conditions, people with weight management concerns, and people trying to quit smoking (Fox, 2011).

Dalmer argues that more studies are needed to alleviate a gap in knowledge of how social media health websites are being evaluated by online health consumers (2017). Y. Zhang, Sun, and Xie note that a large meta-study of 165 previous studies on ehealth revealed that very few studies have been done where experts evaluated content of user-generated health content for quality and credibility, reflecting another gap in the body of knowledge (2015).

One large study of social media and health was done by Moorhead et al., which examined 98 previous studies on social media health websites. Moorhead et al., found both benefits and concerns with the use of social media for obtaining health information, including broader abilities of patients to collaborate for emotional support, knowledge sharing, and advocacy, with caveats of potential issues of quality, credibility, and privacy (2013).

Adams conducted a literature review to find the status of online health information in the era of web 2.0, finding that there is still continuing difficulty in defining and rating credibility on the web, with social media providing opportunity for author reputations to be incorporated to help users validate quality (2010). Adams also noted that web 2.0 has not eliminated the online health information issues of advertising combined with health content and non-disclosure of conflicts of interest, such as commercial relationships (2010).

Biggs, Bird, Harries, and Salib looked at 100 YouTube videos providing information on rhinosinusitis, determining 27% to be completely misleading with 12% to have a mixture of some misleading and some accurate content (2013). One video demonstrated how to perform a potentially dangerous invasive medical procedure at home. Total viewership for the top 100 rhinosinusitis videos was over 2 million (Biggs, Bird, Harries, & Salib, 2013), raising concern about the number of viewers who could be exposed to inaccurate or dangerous content. However, most of the misleading rhinosinusitis content was from individual users, whereas the videos created by health information providers and medical professionals was found to be of higher quality (Biggs, Bird, Harries, & Salib, 2013). The problem of inaccurate health information in social media is not confined to the United States. For instance, Li, Zhang and Wang found that 57% of WeChat social media health information in their sample was inaccurate, with content containing rumors, business promotion or excessive hype and overstated claims (2017).

Experimental Systems for Automatic Judging of Health Information Credibility

Information professionals have attempted to address problems with website credibility and quality by developing experimental systems that either present better quality results or help the users evaluate the search results more easily. Li, Zhang, and

Wang identified lack of credibility, lack of accuracy, lack of reasonableness, lack of citations, or a combination of these issues as features that suggest inaccurate health information in social media (2017). The aspiration of Li, Zhang, and Wang is that these indicators for inaccurate health information can eventually be incorporated into a machine learning platform that would automatically filter the content for inaccurate health information (2017). Abbasi, Zahedi, and Kaza developed a machine learning algorithm, called recursive trust labeling, that can filter out fake health websites with 94% accuracy using some common features of low credibility health websites, including phrases like “no side effects” and links to other low quality health websites (2012). Ageev, Lagun, and Agichtein developed a machine learning algorithm to help improve document summaries shown in search engine results pages by using data from the behavior, such as mouse clicks and scrolling, of past users as an indicator of user feedback for the quality of the document summary (2013), but the algorithm by Ageev, Lagun, and Agichtein is not specifically targeted toward health information.

Relevance Assessments

Relevance is assessed by users with respect to topic, task, and context (Mizzaro, 1998). Relevance is dynamic over time (Mizzaro, 1998). With increased domain knowledge, relevance assessment is much quicker (Wildemuth, 2004). Bade argues that the terms “relevance” and “relevance ranking” are confusing because relevance rankings are based on algorithms that use estimates, measurements, and predictions to order surrogates that represent information, but relevance is a subjective decision made by a user, using context (2007). Chuklin and de Rijke posit that users make judgements based on "*presentation* of results: result attractiveness ('perceived relevance') and immediate usefulness of the snippets ('snippet relevance') (2014, p.1).

Over a period of time, users can and do change their relevance judgements, as demonstrated in the study by Zhitomirsky-Geffet, Bar-Ilan, and Levene in which university students were asked at two points during a semester to assess the relevance of individual documents and the search engine result set as a whole in reference to a query, in which over 30% of the changes were changes larger than one position in the rankings (2016). But, a combined crowdsourced relevance ranking for the search results was more consistent over time, leading Zhitomirsky-Geffet, Bar-Ilan, and Levene to wonder if crowdsourced search engines are a better information retrieval aim than personalized search engine results (2016).

In a study by Mao et al., relevance and usefulness were compared (2016). A group of assessors first manually rated documents for relevance, based on the query terms, and then a group of undergraduate users who volunteered for the study used a search engine to search for documents that would be useful for a specified task or situation (Mao et al., 2016). The study participants rated the documents found by the search engine for usefulness, after which Mao et al. compared the usefulness ratings to the relevance rankings, finding that relevance and usefulness are related but not aligned. For instance, 29.3% of documents marked as moderately relevant by the assessors were coded as not useful by the students (Mao et al., 2016). The authors note that relevance is more centered on the topic, but usefulness encompasses context and subjective factors, as usefulness includes, but goes beyond relevance (2016). Usefulness is a better predictor of user satisfaction with search results than relevance (Mao et al., 2016).

In a study with undergraduate and graduate student participants from an Asian university, Xu and Chen identified novelty (the “newness” and “originality” of the information to the user) and being “on topic” as the most important factors used by the participants to assess relevance, followed at a distance by understandability and reliability (2006). Spink and Greisdorf asked graduate students at an American university to search for information using a library database and then rate the retrieved items for relevance, finding that being “off topic” was enough to mark a result as irrelevant, but being “on topic” was not sufficient to conclude that the result is relevant (2001). An attribute often used to conclude that a search result is relevant is whether or not the document is useful for addressing the applicable information need (Spink & Greisdorf, 2001).

When asked to rate documents for relevance, the searchers do not universally agree on the relevance of the documents, leading Hariri to attribute the differences to “cognitive, affective, and even physical factors” (2011, p. 605). Bar-Ilan, Keenoy, Yaari, & Levene attribute user differences in relevance judgements to the process in which the users “evaluate information in their own context, which is influenced by cognitive, affective, and physical factors.” (2007, p. 1254).

Relevance and Traditional Search Engine Results

Popular general-purpose search engines like Google or Microsoft Bing commonly display the URL, page title, and a document summary as surrogates for individual websites in the results (Al-Maqbali, Scholer, Thom, & Wu, 2010; Capra, Arguello, & Scholer, 2013; Gwizdka, 2009; Kelly & Azzopardi, 2015). Search engine users do not have time to examine, compare, and contrast every one of the individual results presented on a search engine results page, so they must use the document summaries as surrogates to assess the relevance of the results (Lewandowski, 2008). Furthermore, document summaries are not enough. “Good surrogates should provide metadata that enable users to predict the relevance of the document quickly and accurately.” (Crystal & Greenberg 2006, p. 1370). There is evidence that the standard search result page presentation of URL, page title, and document summary does not provide enough metadata to judge relevance (Joho & Jose, 2006).

One problem with the traditional Google-style ranked list of search engine results is that the ranking does not align to other measures or constructs of relevance. For instance, when looking at gastric cancer results, researchers found that general-purpose search engine algorithms did provide high rankings to websites that presented good quality health information, but the more relevant sites with specialized expertise, like the Association of Upper Gastrointestinal Surgeons of Great Britain and Ireland, were not in the top 50 web results examined (Killeen et al., 2011). In one study of 47 Google queries, participants labeled 27.4% of the top-ten search results as irrelevant (Patil, Alpert, Karat, & Wolf, 2005). When searchers were asked to evaluate 40 Google search results (10 results per page) for relevance, the documents ranked as most relevant by at least 41% of the participants were distributed over the pages with documents being rated as most relevant being found on page 1 (documents 1,2,3,5,7,9); page 2 (documents 13, 18); page 3 (documents 20,25); and page 4 (36,37) (Hariri, 2011). In looking at crowdsourcing’s effects on relevance, Zhitomirsky-Geffet, Bar-Ilan, and Levene found that the crowdsourced relevance ranking for search results was consistent over time, but differed substantially from how Google ranked the documents (2016). When participants were asked to evaluate relevance for a set of search results, their relevance judgements not only did not agree with the rankings provided by MSN Search, Yahoo! or Google search engines, but were also dissimilar to rankings by the other participants, even though all the participants, as university computer science and information science students, had a similar educational background (Bar-Ilan, Keenoy, Yaari, & Levene, 2007). There was low correlation, although not statistically significant, between relevance rankings provided by search engines and the human relevance rankings created for these same documents by college students (Nowicki, 2002).

Another problem is that the document summaries in traditional search engine results pages do not always help users be able to predict whether the corresponding document will be relevant. A document summary contained in the search results that allows the user to conclude that the website will be irrelevant still provides valuable details to a searcher (Chuklin & de Rijke 2014). When researchers Bailey et al. compared relevance judgements of users who were asked to evaluate the relevance of documents based on the summaries displayed by a search engine against the relevance ratings assigned by users who viewed the actual documents, there were more documents that were rated as relevant based on their summaries than when these documents were assessed using the whole document (2010). The document summary from a search engine is often based on the query terms that match terms in the document and may not reflect the overall topic or relevance of the document (Bailey et al., 2010). "When searching, users interact first with a search engine results page (SERP) and then with the retrieved documents. Each document has a summary, which may overstate or understate its relevance..." (Bailey et al., 2010, p. 105). In a study by Turpin, Scholer, Järvelin, Wu, and Culpepper of relevance of document summaries and their respective documents, users erroneously graded 45% of relevant documents as irrelevant because the corresponding document summaries displayed in the search results caused the users to conclude that those documents would be irrelevant (2009). When university students were asked to evaluate document summaries from randomly ordered search engine results from one of five randomized search engines, Lewandowski concluded that the document summaries align with the document relevance 70% of the time (2008). Also, Lewandowski noted that 20% of the document summaries first appeared to lead to relevant results, but ultimately, when the documents themselves were retrieved, they were irrelevant (2008).

Xie and Benoit III compared how users evaluated document summaries from search engine results and how users evaluated the documents themselves, finding that users assessed credibility, depth, language, reputation, specificity, scope, cost, and layout in both the document summaries and the documents (2013). However, coverage, intended use, item type, speed, unique information, accuracy, currency, ease of use, item length, presence of pictures, presence of numbers or statistics, and availability were checked once the document was opened, while the number of results and their order were observed only at the search engine results level (Xie & Benoit III, 2013). Search result quality may be described as a complex mix of the 21 attributes, according to Xie & Benoit. The authors argue that search results pages should contain more metadata and better document summaries so that more attributes of the results could be reviewed on the search engine results page (Xie & Benoit III, 2013).

Searchers also tend to place too much weight on the search engine’s ranking of the results, even when there are cues that other documents might be better choices. Pan et al. (2007) conducted a study on the effects of search engine ranking using Google in which university undergraduates were presented with Google results that had their ordering altered. When the participants were shown search engine results with lower quality (per the original Google ranking of the documents) document summaries, the participants clicked on the results that were displayed at the top (Pan et al., 2007). Pan et al. used eye tracking software to record attention and noted that while the students in the study looked at the higher quality document summaries for a substantial time, they ultimately discounted their own judgements of relevance and selected less relevant, but top-displayed, results (2007). Using click analysis and eye-tracking, Joachims et al., conducted a study in which they manipulated the order of the results shown in a search engine results page, finding that the study participants (university students) generally picked the first result on the page, but looked at the document summaries for the first and second results about the same length of time (2007). Overall, Joachims et al. found that the students had a bias toward trusting the ranking of the results from the search engine, even when the documents summaries did not appear as relevant (2007). Haas and Unkel asked German university students to view search engine results with varied ordering for the search engine results, finding that the study participants chose to review the documents that were higher ranked (2017). Granka, Joachims, and Gay noted that users generally click on the first link in the search engine results page (2004). In a study by Bar-Ilan, Keenoy, Levene, and Yaari, evidence of presentation bias was found, in that users, when asked to pick the “best” search result were most likely to select results shown near the top of the page, even though the study design showed the results sorted in different orders (2009).

Other Topics in the Literature Concerning Traditional Search Engine Results

Google and Bing rely on the user to be at least partially adept at entering a query to start the search process. “Query formulation is known to be difficult for typical web users and even experts have problems in it.” (Käki, 2005, p. 132). When users input ambiguous query terms and do not provide enough detail in the query to match their information needs, then the search results or the algorithms that measure search engine performance may be flawed (Agrawal, Gollapudi, Halverson, & Ieong, 2009). Searchers who were provided a standard (Google-like) interface where they could enter search terms reported higher levels of cognitive workload and frustration in completing a search task than a comparison group who conducted the same task using an interface which provided predefined query refining suggestions after the initial query (Azzopardi, Kelly, and Brennan, 2013).

Major search engines now commonly suggest improved query terms, based on similarity to queries made by previous users (Fattahi, Parirokh, Dayyani, Khosravi, & Zareivenovel, 2016). Query suggestions can improve the relevance of the retrieved documents (Fattahi, Parirokh, Dayyani, Khosravi, & Zareivenovel, 2016). For a discussion of how search engines use query logs and graph algorithms to make query suggestions, see Anagnostopoulos, Becchetti, Castillo, and Gionis (2010).

Although query suggestions do improve the search experience, there are still unresolved problems for searchers. Like search engine results, users are substantially more likely to click on query suggestions at the top of the list, but researchers have not yet determined if the top ranked query suggestions are the most relevant to the searchers or if there is a strong position bias (Mitra, Shokouhi, Radlinski, & Hofmann, 2014).

Strohmaier, Kröll, and Körner argue that for query suggestions to be truly effective, algorithms must be developed to determine the searcher's query intent (2009). For instance, if the searcher uses query term "poker", the user could want to know where to play poker, how to learn poker, or even how to cheat at poker (Strohmaier, Kröll, & Körner, 2009).

One problem with the traditional ranked list of search results is that "the results on different subtopics or meanings of a query will be mixed together in the list" (Carpineto, Osiński, Romano, & Weiss, 2009, p. 17:2). Web page titles with more query terms displayed in bold in the search results page will be more likely to be preferred by users when asked to click on documents that might be relevant (Yue, Patel, & Roehrig, 2010). Similarly, there is a slight, but not significant, effect of having more query terms displayed in bold in the search results document summary resulting in more users selecting that document, believing it to be relevant (Yue, Patel, & Roehrig, 2010).

When using traditional search engine results pages, users may not examine all the results and may miss important results. Joachims et al. observed that most students read the search engine results page from top to bottom and that if the students decided to scroll down to results displayed in positions six-ten, they bypassed spending much time on the results displayed as ranked three-five (2007). However, student behavior varied with the quality of the search engine results presented, with students scrolling down more when shown results ranked with less relevant documents displayed first (Joachims et al., 2007). Another study noted that 73% of U.S. web searchers and 76% of European web searchers examine only the first page of search engine results (Jansen & Spink, 2006). Buscher, White, Dumais, and Huang examined Microsoft Bing query logs enhanced with scrolling logging software to calculate that around 70% of query sessions do not involve any scrolling and that in query sessions that do include scrolling, 55% are downward scrolls that do not return back upward (2012). One study using eye-tracking software suggested that users are unaccustomed to looking at search engine results that are ranked lower than the top 4-5 results, leading to slowness or errors in locating desired information that is not displayed in the top search results (Guan & Cutrell, 2007). Granka, Joachims, and Gay also had similar results when performing eye-tracking analysis on web searchers viewing search engine results, finding that mean time spent looking at the search results displayed as positions 1 or 2 was around 0.75-0.9 seconds for each and dropped off sharply with average time spent looking at search engine results in positions 5-10 was below 0.2 seconds for each (2004). Yet, it appears that users perform better if they have to scroll, given the alternative of not seeing the text summaries unless they hover (Dumais, Cutrell, & Chen, 2001). In a study comparing search engine results pages that always display text summaries and results pages that display text summaries only when the user hovers over the page title listed, Dumais, Cutrell, and Chen found that users were faster at completing search tasks when they didn't have to hover to see the text summary, even though displaying the text summary only on hover often eliminated the need to scroll to see all search results in the list (2001).

Kelly and Azzopardi conducted a study in which they varied the length of search engine results page to show either three, six or ten results per page to 36 undergraduate student volunteers (2015). In the Kelly and Azzopardi study, participants who were presented three results per page viewed 3.5 search result pages (about ten results) per query but reported the least amount of difficulty with their search tasks (2015).

Participants shown three search results per page, spent more time viewing the document summaries for the top ranked results, but didn't spend much time viewing the other document summaries and only accessed a few websites in the results (Kelly & Azzopardi, 2015). In another study of length of search engine results page length, users who were shown six search engine results and were then asked to select the best search result in a short time were more satisfied with their selection, when compared to the users who were shown 24 search engine results and asked to pick the best search result (Oulasvirta, Hukkinen, & Schwartz, 2009). Users might have less cognitive overload when presented with a smaller set of options (Kelly & Azzopardi, 2015; Oulasvirta, Hukkinen, & Schwartz, 2009).

In recent years, some major search engines have started to provide an "answer" to the query instead of, or along with, a list of search results for certain types of queries, such as requests for the weather or the time zone (Chilton & Teevan, 2011). Chilton and Teevan examined a query log and identified some methods that could be used to try to measure user satisfaction with the "answer" provided, such as whether or not the user clicked on the "answer", but also recommended that more research is needed on how to evaluate user experiences with an "answer" as a search result (2011).

Another development in search engine results presentation is the inclusion of results from specialized searches into the search results. "Aggregated search is the task of blending results from different search services, or *verticals*, into a set of web search results" (Arguello & Capra, 2014, p. 539). The vertical results are usually incorporated into the first page of search results (Arguello & Capra, 2014). Verticals can arise from expanding the search by genre (for instance, adding weather, encyclopedia entries, or horoscopes) or by media, such as inclusion of images, sound files, or video (Zhou, Demeester, Nguyen, Hiemstra, & Trieschnigg, 2014). For an algorithm to determine which media or genres to include, the algorithm must predict what the user will expect to see included for the particular query, along with selecting the most relevant items within the vertical (eg. including the two most relevant videos and 3 most relevant images) in the search results (Zhou, Demeester, Nguyen, Hiemstra, & Trieschnigg, 2014). According to Arguello and Capra, when the users receive aggregated search results that do not seem relevant to their meaning of the query, they are not as likely to be satisfied with the search results and vice versa (2014). When relevant images were shown in the verticals, users were 23% more likely to retain the search engine results page and click on a link from the results (Arguello & Capra, 2014). However, for news, video or shopping verticals, if they were enclosed in a border and different color background and/or displayed off to the right side of the search engine results, then they were not likely to influence the user interaction with the search engine results, even if the verticals were irrelevant (Arguello & Capra, 2016).

Categories/Clusters/Facets

Search engines can display search results arranged by category, which is also referred to as a faceted display (Hearst, 2006, p.60). According to the Concise Dictionary of Library and Information Science, a category is a "grouping of related documents; general concept that applies to a great deal of material which can be used to group other concepts." (Keenan & Johnston, 2011). A human can manually assign each document to a category or the categorization can be done automatically by a machine learning algorithm that does classification by category or by cluster of related topics. Clustering is usually done automatically using an algorithm that computes similarity (Hearst, 2006; Losee & Church, Jr., 2005). Search engines that offer categories can make the categories available from the beginning, supporting information retrieval done completely by browsing without any query terms being typed into a search box, or can display the categories only after the query is submitted, as part of the display, examination, and filtering of results (Niu & Hemminger, 2015). Faceted systems in common use, such as MeSH, contain categories that are not mutually exclusive, but the strict definition of a faceted system is one in which all of the categories are mutually exclusive (Niu & Hemminger, 2015; Tang 2007).

"Faceted category structure is one way to help people understand the composition of an information collection." (J. Zhang & Marchionini, 2005, p.179). When searchers are unable to develop query terms relating to their information need, they still have a chance for a successful search experience if they are fortunate enough for relevant content to appear on their screen, because users can often identify relevant information when they see it (Savolainen & Kari, 2006). A categorical display of search engine results or a categorical search system can allow the opportunity for searchers who struggle with appropriate query terms to still have access to relevant content. "Just as faceted analysis has been used to remind the indexer of the different aspects by which a document can be represented, a faceted display of the classification might also encourage users to articulate different aspects of their information needs." (Tang, 2007, p. 1998).

Chen and Dumais compared a list of search engine results against a categorized display of search engine results, created on the fly with Support Vector Machines, finding that participants of their study completed their search tasks 50% more quickly with the categorized search engine results, preferred the results from the categorized search engine results, and were less likely to give up on a search task when using the categorized search engine results (2000). In studies of coffee and magazine consumers, the presence of categories was found to increase consumer satisfaction with their choices and with their feelings of control, even when the categories are non-informative labels like "Category A" and "Category B" (Mogilner, Rudnick, & Iyengar, 2008).

J. Zhang and Marchionini studied an interface which allowed users to filter by categories such as genre and format (for films) and fuel type and sector (for an energy website) in comparison with an interface that was more like a traditional search engine, finding that users were satisfied with the categorized interface and were more efficient when using it (2005). However, because the study by J. Zhang and Marchionini was confined to two collections, and tested in isolation, this study may not be applicable to searching for health information on the broader Internet.

Tang developed a prototype information retrieval system for the PubMed bibliographic database that would allow users to both search by entering query terms into a box like a traditional search engine and also to browse via a directory of Medical Subject Heading (MeSH) terms, deliberately designed with dual information retrieval options because most searchers are now accustomed to entering terms into a search box (2007). After Tang's information retrieval system for the PubMed bibliographic database was developed, a naturalistic study of 19 students, health professionals, and researchers was conducted in which participants were asked to use the system to complete tasks from their real lives, with the freedom to use either the search box or the MeSH browsable directory. The preference for search interface varied by task and topic familiarity, with searchers looking for comprehensive understanding of a topic less likely to use the search box, but searchers desiring a quick answer on a familiar topic more likely to use the search box (Tang, 2007).

Multiple other researchers have attempted to study the effects of MeSH on health information retrieval. For instance, Y.H. Liu and Wacholder used MeSH as a controlled vocabulary, instead of a browsable tree, but still found that the interface which incorporated MeSH allowed domain experts to get better search results than just a standard search box (2017). Two additional studies comparing MeSH hierarchical, browsable trees and a standard search engine search box for health information retrieval, found that the MeSH hierarchical tree search interfaces allowed searchers to be more efficient and to overcome some of the disadvantages of topic unfamiliarity (Mu, Lu, & Ryu, 2014; Swetha, Uma, Suganya, Nivedhitha, & Saravanakumar, 2014).

In a study comparing a traditional search engine presentation of ranked results and a combined interface showing the results automatically clustered in a list of categories by similarity in the left sidebar with traditional search engine results in the center of the screen, results were mixed (Pu, 2010). For instance, when using the traditional search engine results presentation, users were faster and found a relevant result faster, but when using the category listing, the participants were able to find additional relevant pages (Pu, 2010). Some users found the categories and sub-categories helpful for reducing information overload and cognitive load for unfamiliar topics or for topics with a large amount of results, but some users felt that the categories and sub-categories induced anxiety as they worried about missing important results (Pu, 2010). Because of these examples and other contradictory results, Pu advocates for a dual interface for search results presentation (2010).

In that direction, Burt and Liew, conducted a study in which twelve participants were asked to search for a job and accommodations in a city of their choice by utilizing a search engine with both a traditional list of results and a categorized tree of browsable results that was created by a clustering algorithm (2012). Participants who liked the categorized results mentioned that the categories helped them rule out groupings of irrelevant results, assisted them in constructing query terms for re-query and empowered them to control the search toward the context where they wanted to proceed (Burt & Liew, 2012). But other participants in the study by Burt and Liew felt that seeing all the possible categories was tiring and that utilizing the categories was too time-consuming (2012). During interviews with the participants about their search experience, all participants mentioned Google as a comparison point, leading Burt and Liew to conclude that their study provides further support for a search engine offering both a categorized tree of browsable results and a traditional search engine list of results (2012). While some users will benefit from the categories, other users will likely not feel that categorized search results benefit them any or enough to migrate away from a traditional search engine list of results to which they are accustomed (Burt & Liew, 2012).

Carpineto, Osiński, Romano, and Weiss also argue that search engines should offer both a categorized tree of browsable results and a traditional search engine list of results, because the categorized tree of browsable results can help users without topic knowledge gain an overview, quickly find relevant (but not highly ranked results), and provide ideas for query refinement (2009). In a study which utilized an experimental search engine, Findex, that combined a traditional search engine list of results with categories in the left sidebar, participants used the categories about 25% of the time and were able to use the categories when their query terms did not return satisfactory results (Käki, 2005). The participants in Käki's study indicated that the categories were helpful in the cases of vague or broad topics (2005).

Dumais, Cutrell, and Chen tested seven interface designs to further investigate the effects of including categories in search engine results (2001). When using the interfaces with search engine results grouped by category, the participants were able to complete their search tasks more quickly than when using traditional search engine listings of results, even when the category interfaces were modified to not include the category name (Dumais, Cutrell, & Chen, 2001). The interface that was the most efficient for the participants in the Dumais, Cutrell, and Chen study was one which grouped the results under a specified category name and included the page title, URL, and text description (2001).

In a comparison of search engine logs from a search engine that provides traditional search results pages and the same search engine providing clusters as search results, Zamir and Etzioni concluded that once the users identified clusters of interest, finding documents within the clusters was easier than finding additional relevant documents using the traditional search engine results pages (1999). However, with users on average viewing seven documents in three separate clusters, Zamir and Etzioni speculate that the clustering algorithm should be improved so that one cluster will provide the information needed by the searcher.

You, DesArmo, Mu, Lee, and Neal presented a new health information retrieval system, at the 2014 ACM/IEEE-CS Joint Conference on Digital Libraries, which has a traditional Google-like search box and search results, but also includes a browsable tree of results organized by Medical Subject Heading (MeSH) categories and a visualization of how the query term(s) fit into the MeSH categories (2014). Unfortunately, results of studies in which this new health information retrieval system was tested by users do not seem to be available as of this writing.

Word Clouds

Another approach for representation of search engine results is the use of word clouds. According to the Oxford Dictionary of Journalism, a word cloud is “a graphic representation of the frequency with which certain words are used in a speech, statement, document, judgement, or similar. The higher the frequency, the bigger the depiction.” (Harcup, 2014). Word clouds are commonly used in social media websites, where they are often called tag clouds. In social media, “simple keywords called tags are used to categorize the information on the site (such as the photos or bookmarks), and tag clouds are frequently used as a way to give an overview [of the tags].” (Bateman, Gutwin, & Nacenta, 2008, p.193).

Word clouds can be used in the following information seeking tasks: searching, browsing, gaining an overview of a topic or collection, and distinguishing between entities or contexts (Rivadeneira, Gruen, Muller, & Millen, 2007). Word clouds are more flexible surrogates than document summaries because they can represent documents that are not standard written language like microblogs, texts, or video tags (Kaptein & Kamps, 2011).

Word clouds can also vary by how the terms within the word cloud are arranged, such as with the most frequent word in the center and the less frequent words on the outer edges (Rivadeneira, Gruen, Muller, & Millen, 2007). Other word cloud designs can include an alphabetic display format or arrangement by term frequency (Rivadeneira, Gruen, Muller, & Millen, 2007). In word clouds, the importance or frequency of the terms are visualized by the font color, font size, or font size of the terms shown in the cloud (Bateman, Gutwin, & Nacenta, 2008). A word cloud can be crossed with clustering to display multiple bunches of terms that are similar to each other, arranged in separate word clouds with importance of the terms visualized by font (Rivadeneira, Gruen, Muller, & Millen, 2007).

The literature includes discussion about design principles for optimal word clouds. Bateman, Gutwin, and Nacenta conducted a study of social media tags displayed in tag clouds with various properties (such as font size, bold fonts, font darkness, visual contrast, and term special position in the cloud), finding that terms in the cloud that had a bigger font size, had a darker font, and used bold fonts were more likely to receive attention and to be clicked by users (2008). Bateman, Gutwin, and Nacenta also found that users were more likely to click on terms in the center of the cloud, unless their attention was re-focused by terms with a bigger font, bolder font or darker font (2008). Venetis, Koutrika, and Garcia-Molina developed models and algorithms to measure the quality of a word cloud, suggesting that coverage, relevance, cohesiveness, and overlap are the most important hallmarks of a quality word cloud (2011). Findings from a study by Rivadeneira, Gruen, Muller, and Millen suggest that users may recollect the terms displayed in a word cloud better if the terms are in a larger font or if the terms are located in the upper right quadrant (2007).

Multiple studies have also been conducted to see if word clouds can improve search engine results presentation. By enhancing an experimental search engine results page with a word cloud representing the top 100 Yahoo! webpage, researchers were able to look at the interaction of word clouds, search engine results, and topic learning, (Wilson, Hurlock, & Wilson, 2012). In the study by Wilson, Hurlock, and Wilson, participants were asked to write a short document on one of six topics, then use either a search engine enhanced with the word cloud of the top 100 Yahoo! webpage results or a traditional search engine to learn about the selected topic, and finally, to write another short document on the selected topic. The findings of the study suggest that including a word cloud representing the top 100 Yahoo! webpage results in the search results pages has no effect on topic learning (Wilson, Hurlock, & Wilson, 2012). Kaptein and Kamps conducted a study that used word clouds to represent search results, grouped into categories (2011). Accuracy of relevance judgements when using word clouds was about 60% and sub-topics were correctly identified about 70% of the time (Kaptein & Kamps, 2011). Another component of the study by Kaptein and Kamps was the comparison of word clouds based the query terms found in the search results, word clouds based on the hyperlinks found in the websites in the search results, and word clouds based on the full text in the websites in the search results. When shown the word clouds based on the hyperlinks found in the websites in the search results, the participants were able to complete tasks more easily than when shown the other types of word clouds in the study

(Kaptein & Kamps, 2011). Kaptein and Kamps also reported that the participants preferred the word clouds based on the hyperlinks (2011).

Al-Maqbali, Scholer, Thom and Wu conducted a study to compare these five different methods of displaying search engine results lists: (a) traditional search engine results list, (b) traditional search engine results list with a visual enhancement of a thumbnail screenshot of the webpage in the results, (c) traditional search engine results list with visual enhancements of a tag cloud and a thumbnail of the webpage in the results, (d) traditional search engine results list with a visual enhancement of the top ranked image from the webpage in the results, and (e) traditional search engine results list with visual enhancements of the top ranked image from the webpage in the results along with the website logo (2010). None of the visual enhancements helped the users to predict which results would be relevant (Al-Maqbali, Scholer, Thom, & Wu, 2010). Additionally, the tag cloud led to users predicting non-relevant results as relevant, even though the users took longest, although not a statistically significant amount of time, when using the tag cloud version of the search engine (Al-Maqbali, Scholer, Thom, & Wu, 2010).

Sinclair and Cardew-Hall asked Australian undergraduate students from an engineering class to use a combined interface of a traditional search box with the tag clouds from a science news social media site to conduct searches for topics like applications of nanotechnology, Mars, and sustainable fuels (2008). The results suggest that the search box was more useful for specific searches, but the tag cloud was more useful for browsing for broad subjects, for finding query terms, and required less cognitive load (Sinclair & Cardew-Hall, 2008). Although this study involved social media tags and querying, rather than search results, it is still helpful because some of the findings could be applicable to displaying search results in a tag cloud to support browsing and query re-formulation.

Gwizdka studied the associations between search tasks and user cognitive ability in two presentations of search results using word clouds (2009). One of Gwizdka's search result displays included a word cloud instead of a text summary for each individual search result in the search engine results. The other search result display in Gwizdka's study was the same as just described, with the addition of a right sidebar with an overview word cloud of the important terms in the overall collection of search engine results. Terms in the individual document word clouds and the overview word cloud were clickable, allowing for a refined query and narrowed results (Gwizdka, 2009). In participants who had tested high on cognitive abilities testing during Gwizdka's study, performance on the search tasks was significantly faster when using the interface with both the overview word cloud and the individual document word clouds.

Kuo, Hentrich, Good, and Wilkinson conducted a study comparing two presentations of Pub-Med search engine results (2007). One presentation was a list of search engine results, showing the page title, text summary, and PubMed ID of the article, while the other presentation was an overview word cloud of the result set (Kuo, Hentrich, Good, & Wilkinson, 2007). Results from the study by Kuo, Hentrich, Good and Wilkinson were contradictory with participants rating the overview word cloud search results higher in user satisfaction but rating the list of search engine results as more helpful (2007). For finding descriptive questions, the users were faster and more accurate when using the word cloud search engine results, but for questions about relationships (such as finding genes involved in a given biological processes), the users were faster and more accurate when using the list of search engine results (Kuo, Hentrich, Good, & Wilkinson, 2007).

Halvey and Keane recruited university students in Ireland for a study of speed in finding and clicking on a country name when randomly presented with a horizontal list, an alphabetized horizontal list, a vertical list, an alphabetized vertical list, a word cloud, and an alphabetized word cloud. Mean times in seconds were 2.887 (alphabetized horizontal list), 2.892 (alphabetized vertical list), 2.94 (alphabetized word cloud), 3.199 (horizontal list), 3.241 (vertical list), and 3.409 for the word cloud. (Halvey & Keane, 2007). However, the authors note that confounding variables included font size of the country name in the word cloud and position of the country name in the word cloud, with country names in the upper right corner being found more quickly (Halvey & Keane, 2007).

Another study that focused on speed in finding terms in a word cloud was conducted by Schrammel, Leitner, and Tscheligi, but their study only compared different word cloud layouts and not lists (2009). In their study, Schrammel, Leitner, and Tscheligi found that the alphabetized word clouds were best for speed in finding a term in a word cloud (2009).

O'Grady et al. examined credibility, health information, social media tagging, and searching using a tag cloud (2012). Participants (Canadian adults with diabetes) were first asked to explore social media tagging by adding tags to health online forum posts using a modified interface for an online health forum, with the option to add content tags of their choosing (eg. terms such as "glucose" or "exercise") or metadata tags (from a defined list of terms including “author”, “references”, and “statistics”), or both types of tags (O’Grady et al., 2012). The six metadata tags were adapted from (Rains & Karmikel, 2009) in which indicators of health information credibility were discussed and studied. In the second phase of the study by O’Grady et al., participants were asked to use a tag cloud search interface to complete two search tasks (2009). The tag cloud search interface was pre-populated with content tags relevant to diabetes and with the six metadata tags (O’Grady et al., 2012). Around 80% of participants tagged content with either metadata tags or both metadata and content tags, although only one third of the participants, during post-task interviews, stated that they selected metadata tags because they were indicative of credibility (O’Grady et al., 2012). When specifically asked to search for credible forum posts, over 90% of the participants included at least one metadata tag in the query, but again, the participants did not attribute many of the metadata tags as being indicative of credibility (O’Grady et al., 2012).

Thumbnail Screenshots or Other Images as Part of Search Engine Results

Several studies examining the use of thumbnail screen shots or other images in search results are relevant to the research question. They address topics such as augmenting traditional Google search engine results, including browsable categories or comparing other search result presentations to Google as a control. Joho and Jose compared the Google baseline search engine results page with Google results augmented with top ranked sentences from each website in the results, Google results augmented with thumbnail screenshots of each website in the results, and Google results augmented with both the top ranked sentences and thumbnail screenshots (2006). Participants rated the combined interface as the most helpful for finding relevant documents in the search results, but also rated the page title as the most helpful information on average (Joho & Jose, 2006). For a search task involving obtaining background information on a topic, the participants rated the top ranked sentences as more helpful than the standard document summary (Joho & Jose, 2006). But, for a search task where participants were asked to find a list of resources on a topic, the participants rated the thumbnails as more helpful than the standard document summary (Joho & Jose, 2006). Joho and Jose concluded that some of their results are contradictory and speculated that the amount of augmenting needed should be customized (2006).

Teevan et al. contend that a specified web page should appear the same in the search engine results from one day to the next, so that users can remember that they have encountered that specific search result in the past (2009). In the Teevan et al. study, participants saw three types of search results: traditional search engine results (page title, URL, and document summary), thumbnail screenshots of the webpages in the search results and enhanced thumbnails, called visual snippets (2009). The visual snippet thumbnails consisted of a key image from the webpage in the results, the page title and the web page logo (Teevan et al., 2009). As hypothesized, mean time to find results from the previous day was shortest for the participants who were shown the visual snippet thumbnails to help them remember the webpages from the proceeding day (Teevan et al., 2009).

Traditional search engine results (page title, URL, and document summary), thumbnail screenshots of the webpages in the search results, and enhanced thumbnails (containing screenshots of webpages that were modified with larger headings and more prominent display of the query terms in the result webpage) were compared using participants at the Xerox Palo Alto, CA Research Center who were asked to perform four types of search tasks (Woodruff, Faulring, Rosenholtz, Morrsion, & Pirolli, 2001; Woodruff, Rosenholtz, Morrison, Faulring, & Pirolli, 2002). Using the enhanced thumbnails, the average time to complete search tasks was faster than the other search engine result presentations, but there were some variations based on the type of search task (Woodruff, Faulring, Rosenholtz, Morrsion, & Pirolli, 2001; Woodruff, Rosenholtz, Morrison, Faulring, & Pirolli, 2002).

In another study of thumbnail screenshots, traditional document summaries with page title, and URL were associated with overestimation of usefulness of the actual document in the results, but thumbnail screenshots of web pages in the search results were found to cause users to underrate the usefulness of the web pages in the search results (Aula, Khan, Guan, Fontes, & Hong, 2010). A search engine result display that combined both the traditional search engine results elements (document summary, URL, and page title) with a thumbnail led to more accurate relevance judgments (Aula, Khan, Guan, Fontes, & Hong, 2010).

Al Maqbali, Scholer, James, Thom, and Wu, compared two experimental search engine designs (2009). In their study, one search engine provided a traditional search engine list of results enhanced with a sidebar allowing browsing by topic, while one search engine provided the search results as screen shots of the website of each result with a small text area sidebar showing a traditional text-based list of the search engine results (2009). Study participants spent more time completing their tasks and made more errors when using the interface with the screen shots and the small sidebar containing the text-based list of results (Al Maqbali, Scholer, James, Thom, & Wu, 2009).

Loumakis, Stumpf, and Grayson were interested in how inclusion of an image for each of the search engine results would affect search efficiency and user satisfaction (2011). The authors searched for images on the web that were relevant to each search task in their study and then created an interface that would include these images in the search engine results (Loumakis, Stumpf, & Grayson, 2011). Among the study participants who expressed a preference regarding the interface, 70% preferred inclusion of an image for each result, but there was little effect from image inclusion on user ability to select relevant search results or time taken to complete tasks (Loumakis, Stumpf, & Grayson, 2011). Capra, Arguello, and Scholer conducted a similar study to that of Loumakis, Stumpf, and Grayson to test inclusion of an image with each search result. However, Capra, Arguello, and Scholer used an image taken directly from each document or website that appear in the search engine results (2013), but their results were somewhat similar to the results of Loumakis, Stumpf, and Grayson. Capra, Arguello, and Scholer found that, in general, adding an image to each search result did not improve the ability of the users to find relevant documents (2013). However, results suggest that a relevant image improved the searcher's success when the text summary provided insufficient information to make relevance judgements or when the results were diversified with many contexts for the query terms (Capra, Arguello, & Scholer, 2013).

Examples of Alternate Displays of Search Engine Results

Some other types of experimental systems have been developed to give users more details for assessing the document summaries. Although some of the display types are not specifically addressed in the research question, findings from the studies can still provide background related to the research question. For instance, Yamamoto and Tanaka included visualizations of attributes of credibility (such as accuracy, authority, and objectivity) using surrogates (such as Google PageRank, social media bookmarks for the site, and similarity to other documents) for each document summary, finding that the visualizations enhance the ability of users to select the relevant results from the search results page (2011).

González-Ibáñez, Proaño-Ríos, Fuenzalida, & Martinez-Ramirez developed a visual search engine called SERVS (Search Engine Results Visualization System) that displays 400 search results at a time in a spiral bubble chart, allowing for details about an individual result to be shown when hovering over a bubble (2017). In a study of 20 undergraduate and graduate students, in which subjects used SERVS and a search engine that displayed results in a standard list, SERVS was superior in reducing user effort, but there were no significant differences between SERVS and the traditional search engine list in performance, emotional experience, and usability (González-Ibáñez, Proaño-Ríos, Fuenzalida, & Martinez-Ramirez, 2017).

Chau conducted a study that compared (a) a traditional search engine list of results enhanced with an additional row showing metadata such as term counts and link counts for each result, (b) a traditional search engine list of results enhanced with a flower glyph representing the term counts, link counts, and document length for each result, and (c) a combined interface (2011). “Glyphs are graphical objects that represent the values of multiple dimensions by multiple visual parameters such as positions, colors, sizes, and shapes” (Chau, 2011, p. 2). All three systems were tested in a study of university students, who rated the combined interface as the most usable, but performance was better for simple, easy tasks when using the traditional search engine list of results enhanced with a row showing metadata such as term counts and link counts for each result (Chau, 2011).

Heo and Hirtle conducted a study comparing four types of search results displays: (a) a distorted visualization which enhances the topic map relevant to the query while de-emphasizing the irrelevant sections of the topic map, (b) a zoomed visualization presenting only the section of the topic map relevant to the query, (c) a hierarchical category tree, and (d) a control with no visualization of search results (2001). Participants using the zoomed visualization made more errors and spent significantly more time completing their searches than participants using the hierarchical category tree, with the participants using the zoomed visualization or the distorted visualization also performing significantly worse than the control group (Heo & Hirtle, 2001). Participants rated the hierarchical category tree favorably for understandability and manageability, but participant performance differences when using the hierarchical category tree and the control were insignificant (Heo & Hirtle, 2001).

Hoeber and Yang conducted a study to evaluate WordBars, an experimental information retrieval system which augments the search results with a sidebar showing an interactive histogram of the terms in the page title and document summary of the collection of the first 100 search results (2008), supporting filtering and query modification. The participants, computer science graduate and undergraduate students, reported higher satisfaction and higher confidence in the search results, compared with the default search results, after they utilized the interactive histogram to modify the query (Hoeber & Yang, 2008). Shani and Tractinsky conducted a unique study in which users were presented with search engine results in one of three formats: results only, results with a numeric relevance ranking, and results with a relevance graph (2013). In the study, users shown the results only format clicked on more results per query in order to accomplish their search tasks (Shani & Tractinsky, 2013).

Koshman conducted a study to test a prototype interactive visual search engine that represents the query terms and sets of results shown as glyphs on a graph, but participants who were not experts in the system before the study had some difficulties learning and using the system without errors (2005). Because users are accustomed to existing popular search engines, viewing search engine results presented in a non-traditional format may increase their cognitive load (Koshman, 2005).

Instead of looking at an alternate display of search engine results, Ageev, Lagun and Agichtein (2013) are trying to improve search engine results by improving the document summaries. After determining that document summaries from search engines are too centered around the query terms and which sections of the documents match the query term, Ageev, Lagun, and Agichtein used eye-tracking and mouse cursor movement analysis to try to identify which parts of the document summaries and documents are relevant to the users (2013). Using the results of the eye-tracking studies, mouse cursor studies, and other data about document summaries, Ageev, Lagun, and Agichtein used machine learning to create document summaries that were assessed to be of good quality by a group of users (2013). Liang, Devlin, and Tait devised a measure of document summary quality for the summaries shown on a search engine results page, defining document summary quality as an average of how well the summary represents the document and how well the summary assists the user in assessing relevance (2006).

Bailey et al. argue that search engine result pages should be reviewed collectively and separately. They created an interface called SASI that allows users to rate aspects of the entire page of results for attributes, such as overall satisfaction and overall authority of the results, plus allows users to rate each search result individually and the sidebars individually (2010).

Mobile Search Results Display

As mobile devices become more and more widespread, the question of how to best display search results so that users can easily identify the most relevant documents has now expanded to small screens. Research such as that of Guo, Jin, Lagun, Yuan, and Agichtein (2013) will help us move toward that end. Guo, Jin, Lagun, Yuan, and Agichtein explored methods to detect that users had located a relevant document when using touch screen mobile devices, finding that inactivity (no touches detected) is predictive of reading, which is a surrogate for relevance (2013). Traditional web page textual results of the page title, URL, and document summary take up a lot of screen space and might not as useful on mobile devices with their small screens (Teevan et al., 2009). Teevan et al. have studied how to used enhanced thumbnails as more compact surrogates in search engine results pages (2009).

Conclusions from the Literature

Although online searching for health information is common, there is a lot of conflicting information in the literature regarding whether health information searchers can effectively find credible and relevant health information online. An additional complication is that there are no universal standards for making credibility or relevance judgements. Although the Google search results presentation is common, it is not universally accepted as optimal. Some studies suggest that users need additional information in the search results display, but there is no consensus about what additional information, if any, to include in the search results or how to design the layout. There is a little support for retaining the Google ranked list of search results and augmenting it with additional information.

Some studies discussed in this literature review (Kuo, Hentrich, Good, & Wilkinson 2007; Mu, Lu, & Ryu, 2014; O'Grady et al., 2012; Swetha, Uma, Suganya, Nivedhitha, & Saravanakumar, 2014; Tang, 2007; Y.H. Liu & Wacholder, 2017; You, DesArmo, Mu, Lee, & Neal, 2014) have examined search results presentation for health information searches, but, the interaction of Internet health information and search results presentation is not fully known, especially because there are still so many unknowns on the two topics separately. Additionally, some of these studies of online health information and search results presentation involved searches on PubMed, not on a general-purpose search engine. In a study reviewing a potential model for selecting features by parts of speech for ordering documents in information retrieval, the model was discovered to be useful for social science documents but not medical documents (Losee, 2006), suggesting that health information may have undiscovered nuances that defy principles applicable to general-purpose information retrieval systems. Health information searchers may have different needs for their search results than searchers for other topics. Hence, all of these knowledge gaps indicate that a study of search results presentation for health information searches would be beneficial to better inform our field.

Scope and Limitations of this Literature Review

Although this literature review is comprehensive, the topics of search engine results, relevance, credibility, browsing, word clouds, information display, and health information seeking are all extremely broad. Therefore, a substantial, but filtered, collection of the literature has been selected and discussed. Interest in these fields continues to build, as do the variety and types of research. Hence, this literature review could not be exhaustive with becoming unreasonably long. Also, this literature review is limited to articles and other sources in English. Attempts were made to include some selections (written in English) from non-U.S. perspectives, but a truly global perspective on these topics is beyond the scope of this paper and remains an exciting avenue for future research.

Method

Overview

UNC-CH Non-Biomedical IRB Study Number 18-2487 was a study of a sample of online health information searchers to see what their preferences are with respect to four different display styles/layouts for search engine results on health topics. Screen shots of the control (Google traditional search engine results list) and three experimental search engine results displays were presented to the participants via a Qualtrics (www.qualtrics.com) online survey. For each search task, participants were asked to rate the search engine results displays for quality indicators, using Likert-type rating scales (as discussed in Babbie, 2004, pp. 169-70 and Wildemuth, 2016, pp. 292-293). At the end, in three concluding questions, the participants were asked to choose the display(s) that were best at meeting three specific criteria, based on overall impressions. A within-subjects design was used. Subjects saw multiple search engine displays to evaluate the displays for two health topics per participant.

Study Design Decisions

The search results that were presented to the participants are screen shots of actual search results (for display 1) or augmented screen shots (for displays 2, 3, and 4) from actual search results incorporated into the Qualtrics web survey. The users did not need to use Google to retrieve anything. They did not need to retrieve the documents contained in the search result. The focus of Non-Biomedical IRB Study Number 18-2487 is on how well the participants think that the search results display screens show the results. Non-Biomedical IRB Study Number 18-2487 is also about how the results screens help the participants to form impressions of which search results might be relevant and credible without looking at the documents themselves. Other studies have used search engine result visualization to predict relevance of the search results, without access to (Aula, Khan, Guan, Fontes, & Hong, 2010) or review of (Hoeber & Yang, 2008) the relevance of the actual documents in the search results.

As described by Broder (2002), Internet searches fall into 3 types of tasks, navigational (e.g. looking for the URL of your favorite band's official home page), transactional (e.g. going to a site of user generated video content to find a "funny" video), and informational (e.g. looking for side effects of a medication). Broder states that it is easy to evaluate search results for navigational queries because the user has either found the URL for the site or hasn't, while it is often difficult and subjective to evaluate search results for transactional queries (2002). Another consideration is that a study should focus on one type of query because each type of query could have a different process for determining relevance (Lewandowski, 2008). The study (Non-Biomedical IRB 18-2487) followed Lewandowski's recommendation to only employ one type of query, which was an informational query because of the guidance of Broder on the disadvantages of user studies employing navigational queries and transactional queries (2002).

Ideally, a study of preferences of search engine results would use queries from actual users (Koshman, 2005; Kules, Capra, Banta, & Sierra, 2009) and would allow the users to click on and examine the search engine results (Crystal & Greenberg, 2006, p.1381), but the current study (Non-Biomedical IRB 18-2487) was conducted for a master's paper and, thus, was under time and resource constraints. However, the researcher in Non-Biomedical IRB Study 18-2487 created a short search question to accompany the supplied search terms and search engine results in order to provide a context for the searches, which has been done in other studies in which the subjects are not using their own information seeking tasks (see Feild, Allan, & Jones, 2010 or Kules, Capra, Banta, & Sierra, 2009 as examples of more detailed information tasks). The search questions provided to the participants in the current study (Non-Biomedical IRB study 18-2487) were:

- Health question: What substances are common causes of outdoor allergies?
- Health question: When should I see a medical professional for an upset stomach?
- Health question: Can college students get high blood pressure?

which corresponded, respectively, to the three health topics used in the study (outdoor allergies, upset stomach, and high blood pressure in young adults).

A within-subjects design has advantages in that it allows the testing to be conducted with fewer participants ("Repeated Measures Designs," 2015; Sauro, 2015). For example, one can recruit 100 volunteers to test two systems (50 volunteers per system) or one can test the same two systems with only 50 total volunteers if the volunteers each test both systems (Sauro, 2015). Another advantage of a within-subjects design is that it reduces extremes associated with specific participants (Sauro, 2015). If, for instance, a given participant is very negative (or some other characteristic), he or she will rate all screens for all topics negatively, which will tend to moderate the negative effect (Sauro, 2015). Also, a within-subjects design reduces (but cannot fully eliminate) any effects caused by the media or methods of the study (Babbie, 2004, pp. 233-234). Finally, a within-subjects design helps research subjects to form opinions about the screens, because they have something to compare against instead of trying to form an impression about a screen in isolation from any other screens (Sauro, 2015).

A within-subjects design can also have some disadvantages. For instance, participants can suffer from fatigue or boredom more easily, because they are completing more tasks than if the study were implemented as a between-subjects design (Sauro, 2015), which is a risk with the Non-Biomedical IRB study 18-2487 in particular because the participants will determine ratings for two topics for four screens each, plus answer the concluding questions. Another disadvantage for a within-subjects design is that the participants might develop opinions when they are rating the screens for the first topic, then apply those opinions to the screens for the second topic, without considering the screens independently (Sauro, 2015). Another concern for a within-subjects study in the field of information science is that, as the participants become more familiar with how to use a system or how the information is presented, their cognitive load is reduced, which might influence the participants to issue higher ratings for easiness or understandability as the study progresses ("Repeated Measures Designs," 2015).

Ultimately, after considering the pluses and minuses of a between-subjects or a within-subjects design, the researcher chose a within-subjects design because it would allow more screens to be evaluated without recruiting as many research subjects. Another important benefit of the within-subjects design is that it can offset the participants who might be extreme in their ratings. However, asking the participants to rate four screens on six quality attributes for each of three health topics (outdoor allergies, upset stomach, high blood pressure in young adults) plus three overall concluding questions would involve 75 questions, could be too time-consuming, boring, or tiring for the participants. In order to respect the time of the participants and not cause them undue stress, the researcher decided that participants should only be assigned to evaluate screens for two out of the three health topics each. Therefore, in the Non-Biomedical IRB 18-2487 study, a given subject will only have to issue ratings for the six quality attributes for a total of eight screen shots (two health topics x four search engine results display layouts), totaling 48 ratings plus three concluding questions, for a grand total of 51 items in the study. The Qualtrics survey software has existing functionality to assign only two of the three health topics to any participant, on an evenly distributed basis. The design used in Qualtrics was set up to assign (approximately):

- Health topics outdoor allergies and upset stomach to one-third of the subjects
- Health topics upset stomach and high blood pressure in young adults to one-third of the subjects

- Health topics high blood pressure in young adults and outdoor allergies to one-third of the subjects

In order to prevent ordering effects, the researcher used randomization options in the Qualtrics software to randomize the order that the health topics were viewed by a study subject. For instance, some participants with health topics outdoor allergies and upset stomach would see outdoor allergies search engine results screen shots first to rate followed by upset stomach screen shots second. Other participants with health topics outdoor allergies and upset stomach would view upset stomach screen shots first to rate followed by screen shots of outdoor allergies search engine results. The order of the six quality attributes by which each type of search engine screen was rated were randomized as well. Randomization capabilities in Qualtrics were utilized, also, for the order of the three overall concluding questions. For all ratings and concluding questions, the order of the four types of search engine search results screens presented to the research subjects was also randomized.

At the end of the survey, the participants were asked to complete four brief multiple-choice demographic questions:

- What is your age?
- What is your ethnicity (check all that apply)?
- What is your gender identification?
- What is your university affiliation (check all that apply)?

For ethnicity and university affiliation, the participants were allowed to select as many options as needed. Racial categories match the categories recommended by the U.S. Office of Management and Budget [OMB] (1997), although the wording may differ slightly. The OMB recommends that presence of Hispanic/Latino/Latina ethnicity be asked in a separate question, but states that it is acceptable to include the option for Hispanic/Latino/Latina within the race question as long as the individual has the opportunity to select multiple race/ethnicities (1997). Another design decision for the demographic questions was to include them at the end of the study, rather than at the start. Answering demographic questions can enforce stereotypes which can carry over into the survey responses, causing self-bias in how the questions are answered (Sauro 2016a). In order to prevent the demographic questions from self-biasing the survey responses or from discouraging participation due to concerns over certain demographic groups being historically unwelcome (Sauro 2016a), the researcher placed the demographic questions at the end of the survey.

Demographic information was collected only in order to consider the generalizability of the study to the health information searcher population. The study design did not include any plans to compare the search engine results screen ratings by race/ethnicity or by gender identity.

Subjects

The target population for this study is health information searchers in the United States. With 80% of the population reporting having sought health information on the Internet (Fox, 2011), the population of health information searchers is large and was projected to be relatively easy to find in the general community. The researcher conducted the Non-Biomedical IRB 18-2487 study for a master’s paper. Recruitment for the study was via email sent via the UNC-CH mass email system (https://help.unc.edu/help/mass-email-requirements/), a bulk email distribution tool which sends bulk emails to members of the UNC-CH community who elect to receive mass emails. This distribution method is commonly used by UNC-CH researchers of many disciplines to recruit study participants. The researchers do not personally transmit the bulk emails. The bulk email request for Non-Biomedical UNC IRB Study 18-2487 was submitted through the UNC-CH Information Technology Services mass email website, approved, and then distributed by UNC-CH Information Technology Service experts. In order to obtain as broad and as large of a convenience sample as possible, the researcher for the Non-Biomedical IRB 18-2487 study did not utilize any filters to attempt to limit distribution among UNC-CH students, staff, faculty, or hospital employees.

The mass email distribution (Appendix C) for recruiting participants contained a brief description of the study, the incentive offered and a link to the survey. Screening criteria were kept to a minimum. The recruiting email mentioned that the intention of the study was to welcome a wide variety of individuals as participants in order to gather viewpoints representing a diverse society. When potential participants clicked on the link to the Qualtrics survey, they were first screened for being 18 years of age or older. If the potential participants met the age requirement, then they were asked if they had searched online for health information in the past (for themselves or for someone else). If the potential study subjects positively self-identified as an online health information searcher, then the Qualtrics software presented them with the Non-Biomedical IRB 18-2487 study consent form (Appendix D). Within Qualtrics, the participants read the consent form. The consent form was written with objective, non-coercive language. Consent consisted of the question in Qualtrics that stated:

> By clicking below to consent, you confirm and acknowledge all of the following:
> - you are 18 years of age or older
> - you are able to read, write and understand English
> - you want to participate in this study
> - your participation in the study is voluntary
> - you are aware that you may choose to terminate your participation in the study at any time
> - and for any reason

If the participant wanted to consent, he or she clicked the option for "I consent, begin the study". If the participant did not want to consent, he or she clicked the option for "I do not consent, I do not wish to participate" and then, the Qualtrics survey ended.

Using the survey size calculator from http://www.raosoft.com/samplesize.html, a sample size of 377 was the targeted sample size (see Appendix A Table A1), as it would have provided a 5% margin of error and 95% confidence level. However, a sample size of only 199 research subjects was obtained. Using the Raosoft sample size calculator, a sample size of 96 is sufficient to allow a 95% confidence level and the 10% margin of error, which is what was used for the Non-Biomedical IRB 18-2487 study.

Al Maqbali, Scholer, James, Thom, and Wu conducted their study comparing search engine results presentations with 50 participants (2009). Gwizdka tested a word cloud presentation of search results with 23 participants (2009).

Materials and Procedures

Ethics. The values of respect for persons, beneficence, and justice (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education and Welfare [DHEW], 1978) were identified by the Belmont report as essential to the ethical conduct of research involving human subjects. The following review of the ethics considerations in Non-Biomedical Study IRB 18-2487 will start with these three concepts and then end with discussion of privacy protections used in Non-Biomedical Study IRB 18-2487.

Respect for persons. An information sheet about the study was provided within the Qualtrics online survey and participants were given an online informed consent form within the Qualtrics online survey. This study involved only minimal risk, similar to the discomfort, risk or harm normally experienced in daily life, during university exams or during routine psychological tests. Participants were informed in the consent form that they could withdraw from the study at any point. The online survey was estimated to take around 25 minutes to complete, which is not a burdensome amount of time for the participants.

Non-Biomedical Study IRB 18-2487 was not a blind study. The subjects knew for which health topics they were seeing screen shots of search engine results. There was no deception involved in the study design.

During the survey, the participants were able to close their web browser and end their participation at any time as with any website. The option was enabled in Qualtrics for subjects to have the ability to skip any questions that they wanted to skip, as required by the UNC-CH IRB (The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, 2012). Non-Biomedical study IRB 18-2487 provided an extra webpage for participants to elect to submit their survey responses at the end instead of auto-submitting the responses when the last question is completed, which is the default for Qualtrics, allowing the participants one last chance to withdraw.

Beneficence. The participants could opt to remain anonymous. However, if participants decided to participate in the drawing for the gift cards, the participants needed to provide their names and email addresses. However, by creating two Qualtrics surveys (Qualtrics Support, 2018), one for the main study and one for the information collected for the drawing (names and email addresses), the researcher provided the participants with increased privacy because the study responses were recorded in a separate file from the names and email addresses. Data collected was the opinions of the participants with respect to their perceived usefulness of the four search engine result presentations, so the risk to the participants of embarrassing data being collected or accidentally released was minor.

The participants were also asked demographic questions in hopes of trying to assess the representativeness of the sample. However, the questions were structured to reduce the likelihood of a participant being identified. For instance, instead of asking a participant's age, only an age range was collected.

As an incentive, participants were offered a chance to win one of two $50.00 gift cards as an incentive for completing the entire study. The amount of $50.00 for a UNC-CH student, employee, faculty, staff or retiree was selected because it was not likely to cause unusual behavior outside of a participant's self-interest. Also, the amount of $50.00 is small enough that it should not place compromise or coercion on the ability to voluntarily provide informed consent. Participants only had the opportunity to register to win one of the two gift cards, not the assurance that they would win, which would further limit pressure to participate against one's will.

Justice. There were no benefits to the participants, but there were also no benefits that were gained by parties not included in the study. No vulnerable populations were recruited for this study. Only adults aged 18 and over were recruited for this study.

Because there is no direct benefit to participants, this was stated that in the consent form. The consent form did not cite the drawing for the gift cards as a benefit. The drawing was discussed in a separate section of the consent form.

Privacy. IP addresses of the study participants are normally collected by Qualtrics software, but the researcher disabled this collection in order to protect the anonymity and privacy of the participants as required by the UNC-CH IRB (The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, 2012). The Qualtrics survey was configured by the researcher to not collect identifying information with the survey responses. Once the survey is completed, the participants are electronically redirected to a separate survey, which collected, and then stored the names and UNC email addresses for the drawing in a separate data file from the survey responses (Qualtrics Support, 2018).

The demographic questions asked were broad categories, such as age ranges. The survey only asked participants for their university affiliation in high level categories like “Undergraduate Student”, “Graduate, Postdoc or Professional Program Student”, “Faculty”, “Staff” or “Hospital”. No departmental or job function information was collected to make it very unlikely that survey responses could be traced to any given employee. Similarly, the survey did not ask for department, school or major, lowering the chances of identifying a given student’s responses. The survey questions were all Likert-type items or multiple choice, which reduced risk of deducing the identity of a participant via their writing style or use of certain phrases.

Timeframe. The recruiting email and Qualtrics survey link for Non-Biomedical IRB study 18-2487 was distributed by bulk email on 11/7/2018 after UNC-CH IRB approval on 10/24/18. Due to only a moderate response, the recruiting email and Qualtrics survey link was re-distributed on 12/4/2018, after the mandatory waiting period required by UNC-CH mass email policy. Due to exams and holiday break, responses slowed during mid-December and the survey was closed in the Qualtrics software after ten days, marking the end of the data collection phase of the study.

Detailed list of the study procedure steps in the Qualtrics software. The Qualtrics extracts of the survey questions (for the main survey and the 2nd survey for entry into the drawing) are included as Appendix E (Qualtrics Survey Questions). The survey questions, the ratings scales for Part 1, the search result display screens shown to the participants, the reduced size screen shots shown to the participants in Part 2, the multiple-choice answers for the demographics questions in Part 3, and the 2nd survey data collection questions used for the participants to enter the raffle were all included in Appendix E. Appendix F shows the Qualtrics logic that was implemented by the researcher for Non-Biomedical IRB 18-2487 study in order to achieve the flow of the steps within Qualtrics. A list of all the steps in the study procedure within Qualtrics is outlined here:

1. The participant was screened and gave informed consent
2. Two (out of three) health topics were selected for each participant by the Qualtrics software.

3. For the first health topic selected for that participant by the Qualtrics software:

a. a screen shot of one of the four search engine results display layouts was randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

b. a screen shot of another of the four search engine results display layouts was randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

c. a screen shot of another of the four search engine results display layouts was randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

d. a screen shot of the remaining one of the four search engine results display layouts will be randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

4. For the second health topic selected for that participant by the Qualtrics software:

a. a screen shot of one of the four search engine results display layouts was randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

b. a screen shot of another of the four search engine results display layouts was randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

c. a screen shot of another of the four search engine results display layouts was randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

d. a screen shot of the remaining one of the four search engine results display layouts was randomly presented to the participant.

i. the participant was given six questions (in randomized order) asking the participant to rate the search engine display layout on six quality indicators.

ii. the participant answered the 6 questions using Likert-type rating scales.

5. For Part 2 of the study, participant was then asked the three concluding questions in randomized order, accompanied by reduced size screen shots of the four search result screen types that the participant saw in steps 3 and 4.

6. Participant was asked the 4 demographic questions as Part 3 of the study.

7. Participant was asked to submit the survey responses for the main survey.

8. Once the main survey responses were submitted, the participant was given the opportunity to opt-in for the gift card drawing.

9. If the participant chose to opt-in, the participant was redirected to another URL with a 2nd (completely separate) survey where he or she was able to input his or her name and UNC email address to register for the drawing.

Design and preparation of the search engine results screens. For each of the three health topics (outdoor allergies, upset stomach, and high blood pressure in young adults), the researcher had to create the four variations of search engine results display layouts:

Display 1: Standard Google (this is the control);

Display 2: Google enhanced with faceted browsable categories;

Display 3: Google enhanced with a word cloud for each search result

Display 4: Google enhanced with an overview word cloud for collection of search results

To start preparation, for each of the three search task scenarios, a query was entered into Google search engine. The top twenty search results were recorded. As an example, Lewandowski only used the top twenty search results in his study (2008). Twenty is also sufficient because most users only examine the first page of search engine results (Jansen & Spink, 2006). A screen shot of the first ten results was taken, using either the Microsoft Windows built-in screen shot functionality [accessed with Alt + Control + P keyboard buttons or the PrintScreen keyboard button] (Microsoft Windows, Version 1803) or Greenshot open source screen capture software (Braun, Klingen, & Krom, n.d.). Then, the html files for the top twenty search results were downloaded and saved to be used later to create the word clouds for Screens 3 and 4.

The Google logo and Google appearance are included in all four of the search result displays. Users trust the Google brand and see it as a benchmark against which to weigh other systems (Burt & Liew, 2012; Jansen, Zhang, and Schultz, 2009). Alternate representations of the search results can be implemented as an optional browser extension (Aula, Khan, Guan, Fontes, & Hong, 2010), rather than creating a whole new search engine.

Examples of the four search engine result displays are shown in Appendix G. Explanations of how they are constructed follows.

Screen 1: Standard Google. This was the screen shot of the first page of Google search results. All editing of screen shots for all screens was done using Microsoft Paint (Microsoft Windows, Version 1803). Paid ads marked with "Ad" or in the right sidebar were removed in order to minimize the effects of other variables. Other items removed included verticals (Arguello & Capra, 2014), such as videos, images, or news. The researcher also removed the query refining small ovals that Google was occasionally including in their search results in the summer and fall of 2018. These items were removed because they did not appear in the search results for every topic and it would be difficult to control for the effects of these elements. The element of "People also ask" at the top in which Google suggests similar search questions was retained because it appeared in the search results for all three topics. Similarly, the Google element "Searches related to ____________" at the bottom of the search results was included in Screen 1, as it appeared in the search results for all three topics. Screen 1 also serves as the control for the other screens.

The remaining three search engine results are augmentations of Screen 1. The traditional search engine results elements from Google (page title, URL, and document summary) are retained and augmented with other elements to include additional information. Using an unfamiliar search engine during a brief study may not allow enough time for the users to form a new mental model about the system and how the results are represented (Koshman, 2005). Building on the familiarity of Google will allow the participants in this short study to incorporate their existing framework into their evaluation of the screens. Also, some previous studies have concluded that combining a traditional search results presentation with enhanced display of information can lead to higher user satisfaction and better relevance judgements (Aula, Khan, Guan, Fontes, & Hong, 2010; Chau, 2011; Joho & Jose, 2006; Pu, 2010).

Screen 2: Google enhanced with faceted browsable categories. The Yippy search engine (www.yippy.com) displays its search engine results with a Google-style list (page title, URL, and document summary) of search engine results in the main panel and a list of faceted browsable categories in the left sidebar. The Yippy search engine is the concept underlying Screen 2. To create Screen 2, for each health topic, the same query terms that were used in Google were used in the Yippy search engine at www.yippy.com in order to generate the list of faceted categories. The Yippy search engine results were discarded, but, for each health topic, a screen shot of the left sidebar containing the faceted categories generated from the query was taken. The screen shot of the Yippy categories was added to the standard Google search results screenshot by pasting it in the left sidebar area to make a Screen 2 search results display for each health topic.

Screen 3: Google enhanced with word cloud for each search result. For each of the top ten search engine results, an individual word cloud was created and pasted into the Google search engine results screen shot below each individual result. Word clouds for each document in individual search engine results have been examined in other studies (Al Maqbali, Scholer, James, Thom, & Wu, 2009; Gwizdka, 2009).

The word clouds were created by using KNIME Analytics Platform version 3.5.3, an open source program (KNIME Analytics Platform, n.d.) As part of the text pre-processing for each full text individual webpage html file, markup tags were filtered, punctuation was removed, case was converted to all lowercase, and numbers were removed. Stop words were removed via both the built-in stop word list and an additional custom stop word list, which removed some terms used in the creation of the html for the pages (such as jquery, xhtml, and footer), which were not removed as expected when the markup tags were filtered by KNIME. Next, the file was converted to a Bag of Words within KNIME. Then, Term Frequencies (TF) and Relative Term Frequency were calculated by KNIME.

When creating the word clouds, KNIME was set to linear increase for the word cloud font. KNIME was set to be permitted to grow the word cloud font colors to 100% boldness and 100% intensity, because Bateman, Gutwin, and Nacenta found that word clouds with bolder fonts and more intensity in color were more informative to the users (2008). The cloud type created for the individual search engine results was an Inside Out cloud where the most frequent terms are in the center.

In the settings that the researcher chose for the KNIME word clouds, the word clouds created were specified to be 400 pixels x 200 pixels in size. The research also used the option in KNIME to request that the software limit the word clouds to the top 200 terms, because including more terms than 200 created a word cloud where the terms were too small to read (unless the size of the word cloud were to be increased beyond 400 pixels x 200 pixels).

For Screen 3, a complication was encountered where the word clouds took up so much screen space that the screen shot became excessively long when ten results were shown. The Qualtrics survey was unable to display such a long screen shot without reducing the dimensions, which made the search results too small to read. Therefore, the researcher decided to display only five search results for Screen 3, instead of ten as used in Screens 1, 2, and 4. For each of the three health topics, the Screen 3 screen shot was created with only five search results (and the five respective word clouds).

Screen 4: Google enhanced with overview word cloud for the collection of all 20 search results. For each of the three health topics, an overview word cloud for the collection of all 20 documents was created and pasted into the Google search engine results screen shot with placement in the right sidebar. An overview word cloud was examined in a study by Gwizdka (2009), in addition to individual word clouds for each search result.

For each health topic, the word cloud was created by using KNIME Analytics Platform software version 3.5.3, an open source program (KNIME Analytics Platform, n.d.). However, the word cloud for Screen 4 is an overview word cloud created using the full text of the html documents from the top 20 search results, rather than just one individual search engine result document. KNIME had to be set up to view the documents as a collection in order to make the word cloud. Similar text pre-processing that was done to create Screen 3 was performed in text pre-processing for Screen 4. Linear growth of the word cloud was selected by researcher again for Screen 4, as in Screen 3. Also, KNIME was permitted to grow the word cloud to 100% boldness in the word cloud font colors as in Screen 3, because Bateman, Gutwin, and Nacenta found that word clouds with bolder fonts and more intensity in color were more informative to the users (2008).

The cloud type selected for the overview search engine result word cloud was an alphabetic cloud. With alphabetic clouds, it may be easier for users to find the individual terms (Halvey & Keane, 2007). The overview word cloud was designed as size 500 pixels x 500 pixels for each of the three health topics.

Other quality control measures. Screens 1, 2, and 4 were all sized to be 1971 pixels in length. Screen 3 (because of the individual word clouds) was still slightly longer at 2549 pixels. Screens 1, 2, 3, and 4 were all the same uniform width. Particular care was also taken so that a specific type of screen (eg. Screen 4) had all elements displayed at the exact position (number of pixels from the top edge and from the side edge) on the screen, regardless of health topic, to control noise in the screen ratings of the study participants. Additionally, the individual word clouds were all horizontally centered under each corresponding search engine result.

Reduced size screen shots. For Part 2 of the survey, where the participants answer the three concluding questions, thumbnail (or more accurately, reduced size) screen shots were created to remind the participants of the screen shots that had been reviewed in Part 1. There were twelve reduced size screen shots created in total:

Health topics 1 and 2 Screen 1 (control)

Health topics 1 and 2 Screen 2

Health topics 1 and 2 Screen 3

Health topics 1 and 2 Screen 4

Health topics 2 and 3 Screen 1 (control)

Health topics 2 and 3 Screen 2

Health topics 2 and 3 Screen 3

Health topics 2 and 3 Screen 4

Health topics 3 and 1 Screen 1 (control)

Health topics 3 and 1 Screen 2

Health topics 3 and 1 Screen 3

Health topics 3 and 1 Screen 4

Each reduced size screen shot consisted of two reminder screens (one for each of the two assigned health topics) placed side by side for each of the four screen types (Screen 1, Screen 2, Screen 3, Screen 4) to allow the participant to only have to answer the three concluding questions once. The alternative would have been that the participants would have needed to view four screens for the first health topic from Part 1 and answer the three concluding questions, followed by viewing the four screens for the other health topic from Part 1 and answering the three concluding questions. The researcher ensured that each of the twelve screen shots were uniform in size, to prevent screen shot size from becoming an unwanted variable to address. The size for each reduced size screen shot is 720 pixels x 285 pixels.

Survey Design Considerations. Usability experts sometimes recommend using the ISO usability standards as a basis for developing questions for a user study (for example, Fidgeon, 2011; Franzreb & Franzreb, 2016). The ISO 9241 standard describes usability as a combination of effectiveness, efficiency, and satisfaction (as cited by Fidgeon, 2011 and Franzreb & Franzreb, 2016). The measures that the participants used in the Non-Biomedical IRB Study Number 18-2487 to rate the search result screens relate to the ISO definition in the follow ways:

- Relevant (effectiveness, efficiency, satisfaction)
- Credible (effectiveness, satisfaction)
- Quickly find (efficiency)
- Refine (efficiency, satisfaction)
- Visual appeal (satisfaction)
- Overall opinion (effectiveness, efficiency, satisfaction)

When designing the response values for the ratings scales, the researcher aligned the responses to the factor being measured as recommended by some experts (see Vannette, 2018a; Vannette, 2018b as examples). For instance, for the question "Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.", instead of including options for "Extremely unhelpful", "Very unhelpful", "Slightly unhelpful", the scale was built with the view that it is difficult to understand and conceptualize negative degrees of unhelpfulness. Therefore, the values used for the rating scale responses were "Not helpful at all", "Slightly helpful", "Moderately helpful", "Very helpful", and Extremely helpful". All points on the Likert-type scales were labeled with descriptive text labels, which is a best practice (Vannette, 2018a; Vannette, 2018b). The levels chosen for the design were:

- For measure "relevant", rating levels were [Not helpful at all, Slightly helpful, Moderately helpful, Very helpful, Extremely helpful]
- For measure "credible", rating levels were also [Not helpful at all, Slightly helpful, Moderately helpful, Very helpful, Extremely helpful]
- For measure "quickly find", rating levels were [Extremely useless, Moderately useless, Slightly useless, Neither useful nor useless, Slightly useful, Moderately useful, Extremely useful]
- For measure "refine", rating levels were [Extremely difficult, Moderately difficult, Slightly difficult, Neither easy nor difficult, Slightly easy, Moderately easy, Extremely easy]
- For measure "visual appeal", rating levels were [Terrible, Poor, Average, Good, , Excellent]
- For measure "overall opinion", rating levels were [Extremely negative, Moderately negative, Slightly negative, Neither positive nor negative, Slightly positive, Moderately positive, Extremely positive]

One design challenge faced by the researcher was how to communicate to the participants how the query for each type of search result display would be refined in actual use. The participants needed to be informed that the categories in Display 2 and the individual terms (O'Grady et al., 2012) in the word clouds (Display 3 and Display 4) should be assumed to be clickable. It was required for the participants to know this in order to evaluate the screen types in general and also for answering the question about how easy it would be to refine the query. Also, there was a concern about whether or not to include a description of how to refine Screen 1 (control), given that it is based on Google, which is so commonly used. Plus, the researcher also wanted to avoid introducing extra variables into the study, such as “refine message received or not”. Ultimately, the researcher designed the Qualtrics study so that every type of search result screen included the following default message with refining instructions:

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

For Screen 2, Screen 3 and Screen 4, the Qualtrics software also displayed an additional message, in italic type to distinguish it from the default refining instructions, customized to the respective screen:

Screen 2 message: *In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud in the right sidebar.*

Screen 3 message: *In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud below each listing in the search results.*

Screen 4 message: *In addition, this search engine results display would allow you to refine a query by using the folders in the left sidebar.*

Survey Questions. The survey questions are shown in full detail in Appendix E. In Qualtrics, the ratings scales were presented horizontally and the scale items were visually evenly spaced. However, not all electronic versions of this paper may render the ratings horizontally.

Part 1 Survey Questions. In Part 1, the participant viewed one of the search topics for their study block and then answered these six questions, which were presented in random order, for one of the four types of display screens, which were also presented in random order:

- Please rate this search engine results display on how helpful the display would be in helping you choose relevant results.
- Please rate this search engine results display on how helpful the display would be in helping you choose credible results.
- Please rate this search engine results display on how useful it would be in helping you quickly find what you need.
- Please rate this search engine results display on how easy it would be to refine your query. (For each type of search engine results display, a message was included with the question, explaining how this specific type of screen display could be used for query refining.)
- Please rate this search engine results display on visual appeal.
- Overall, what is your opinion of this search engine results display?

Then, the participant sees the other search topic assigned and answers the same six questions.

Part 2 Survey Questions. Although ratings for different aspects of each of the four search engine result displays can provide insights useful for studying granularity of screen preferences, this study also incorporated ratings at a composite level, in order to allow the study subjects the opportunity to consider the screen as a whole and in comparison to the other screens. In Part 2, participants were asked to answer these three questions (presented in random order):

- Overall, which type of display did you prefer the most for viewing search engine results from health information searches?
- Which type of display did you dislike the most for viewing search engine results from health information searches?
- If a browser extension or other customization was available to ensure that your search engine results from health information searches would appear like any of these displays, which option would you pick?

To assist them in remembering the types of search engine result screens, the subjects were shown small summaries of the four displays, in random order, which they viewed during questions they completed for their specific block in Part 1. In Part 2, it was only necessary for the questions to be answered once because they applied to both health topics assigned.

Part 3 Survey Questions. In Part 3, participants were asked the basic multiple-choice demographic questions discussed earlier. They are repeated here for convenience of the reader:

- What is your age?

- What is your ethnicity (check all that apply)?
- What is your gender identification?
- What is your university affiliation (check all that apply)?

Evaluation Criteria. In Part 1, the search engine results pages were evaluated by user assessments of: (a) perceived helpfulness in selecting relevant documents [variable relevant], (b) perceived helpfulness in selecting credible documents [variable credible], (c) perceived ability to help the user find needed information quickly [variable quickly find], (d) perceived ease in helping the user refine their query [variable refine], (e) visual appeal [variable visual], (f) overall opinion of the search engine results display [variable opinion]. The ratings scales used in Part 1 were Likert-type (for further discussion, see Babbie, 2004, pp. 169-70 and Wildemuth, 2016, pp. 292-293). The ratings scales used are shown with each question in Appendix E, which contains all of the Qualtrics survey. Both relevance and credibility were rated on a five-point scale, ranging from “Not helpful at all” to “Extremely helpful”. Perceived ability to help the user find needed information quickly was evaluated by the users using a seven-point rating scale ranging from “Extremely useless” to “Extremely useful”. The refine variable used a seven-point rating scale from “Extremely difficult” to “Extremely easy”. Visual appeal of each display screen type was assessed on a five-point scale, with range of “Terrible” to “Excellent”. The research subjects rated their overall opinion of each search results display screen type, using a seven-point scale of “Extremely negative” to “Extremely positive”. In Part 2, the participants selected their overall preferred search engine results display screen, their most disliked search engine results display screen, and which search engine results display screen they would utilize if given the ability to customize or use a browser extension.

Data analysis. After export from the Qualtrics website, data was cleaned and prepared for analysis using Microsoft Excel (Microsoft Office Professional Plus 2016 en-us, Version 16.011029.20108). Descriptive statistics were calculated using either Microsoft Excel or SAS Enterprise Guide (Version 7.15 HF7 7.100.5.6177 64-bit). Chi-Square Test for Equal Proportions and Pearson's Chi-Square Test were computed using SAS Enterprise Guide (Version 7.15 HF7 7.100.5.6177 64-bit). As per the survey size calculator from http://www.raosoft.com/samplesize.html (or see Appendix A Table A1), the sample size of 199 participants allows a 95% confidence level with a 10% margin of error.

Results

Counts and Demographics of Participants

As shown in Appendix A Table A2, there were a total of 279 clicks on the link to visit the Qualtrics webpage with this survey. However, only 199 participants completed the survey. Since the survey questions were completed anonymously and without IP address tracking to protect privacy, it is possible that someone could have started the survey, left, and completed it at a different time, resulting in him or her being captured in the totals for both completing and not completing the survey. For purposes of further discussion, terms such as "participants", "research subjects", and "respondents" will refer to the 199 participants who completed the survey.

It is useful to discuss how "not completed" was determined for the purposes of the present study. The 63 people who started and did not finish the survey completed some or all of Part 1, but none of Part 2, and none of the demographic questions at the end. Also, these 63 individuals did not submit their survey at the end. Hence, all of these surveys were counted as "not completed". As seen in Appendix A Table A2, another 17 prospective participants ended their survey visit at various points before completing any Part 1 questions.

Participants were allowed to skip questions, as required by the UNC-CH IRB to avoid psychological distress for the participants. One participant completed the study, but skipped or did not complete ratings for two questions (evaluating Health Topic 2 Display Screen 4 for the credible quality measure and evaluating Health Topic 2 Display Screen 1 for the refine quality measure). The data of this participant was used and results have notations for the counts of ratings being reduced for those variables (credible and refine). Some participants also skipped or did not complete one or more demographic questions at the end, but their surveys were still counted as completed because they completed all questions in Part 1 and Part 2.

As shown in Appendix B Figure B1, the participants were overwhelmingly female. While Fox and Duggan (2013) noted equal rates of Internet usage among men and women, women were found to be more likely to have used the Internet for health searches, with 62% of women searching for information online about a specific medical condition, compared to only 48% of men (Fox & Duggan, 2013). Additionally, more women (49%) than men (37%) reported searching online for information about specific treatments (Fox & Duggan, 2013). Therefore, the gender demographics of the present study are not inconsistent with the population tendency of more women than men to search for health information online.

The participant ages were surprisingly evenly divided among age groups as seen in Appendix B Figure B2. As the survey was distributed through a university mass email system, it was expected by the researcher that young adult students would be the primary respondents. Neither age group 65-74 nor age group 75 and older is well-represented, but this would be anticipated in a convenience sample using a university mass email system. Fox and Duggan noted that age groups 18-29, 30-49, and 50-64 use the Internet for searching for health information at rates of 76%, 75%, and 71%, respectively, while only 58% of older adults (age 65 and over) use the Internet for searching for health information (2013). The age demographics of the health information searcher population, then, are not altogether dissimilar from the age demographics of the survey sample.

Appendix A Table A3 summarizes the race and ethnicity characteristics of the study participants. When participants selected multiple definitions for their race/ethnicity, the combinations are listed in the results, instead of rolling up into a “Two or More Races Category”, to align with best practices suggested by the U.S. Office of Management and Budget [OMB] (1997). Although the recruitment email specifically mentioned the need for perspectives from a wide variety of races, the Caucasian or White racial/ethnic group still predominated the survey respondents. The Caucasian or White racial/ethnic group is 61.72% of the U.S. population, the Black or African American racial/ethnic group is 12.38%, the Hispanic or Latino/Latina population in the U.S. is 17.66%, and the Asian population is 5.28% of the U.S. population (U.S. Census Bureau, 2017), but the convenience sample in the present study did not closely resemble the U.S. population racial/ethnic distributions. Although it is somewhat difficult to tell the magnitude of the gap, because of methodologies surrounding counting of the category of two or more races, it appears that there is a particularly large gap between the percentage of Hispanic or Latino/Latina persons in the convenience sample in this study and the target population of online health information searchers.

In Appendix A Table A4, the university affiliation of the respondents is listed. As mentioned when the age demographics were discussed, a lower number of students than expected participated in this study.

Topics Assigned

In the Qualtrics software, the survey flow was designed so that the software randomly assigned two health topics (out of three total) for each participant that would be used in Part 1 and Part 2 for evaluating the four types of search engine result screens. Qualtrics has some functionality that attempts to keep the random assignment of topics to participants approximately equal, which aligns to the design of this study in trying to have approximately equal numbers of participants see each combination of health topics:

- outdoor allergies and upset stomach (Topics 1 and 2)
- upset stomach and high blood pressure in young adults (Topics 2 and 3)
- outdoor allergies and high blood pressure in young adults (Topics 3 and 1)

Appendix B Figure B3 shows the actual distribution of the health topics among the survey participants. During the study, 68 participants were shown Topics 1 and 2 to use in rating the search engine results displays, while 69 participants were shown Topics 2 and 3 and 62 participants were shown Topics 3 and 1 for their ratings tasks. A Chi-Square Test for Equal Proportions was conducted. As shown in Appendix B Figure B4, the Chi-Square was 0.4322 and the p value was 0.8057 using a significance level of 0.05.

Therefore, the null hypothesis cannot be rejected. The survey is assumed to have equal proportions between the blocks of two assigned health topics. Appendix B Figure B4 also shows the frequency distribution of the health topics assigned.

Part 1 Results

In Part 1, participants were asked to perform the following ratings for each type of search engine results display for both health topics:

- Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.
- Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.
- Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.
- Please rate this search engine results display on how easy it would be to **refine your query.**
- Please rate this search engine results display on **visual appeal**.
- Overall, what is **your opinion** of this search engine results display?

Appendix A Table A5 lists the summary statistics for each of the quality measures. However, the reader should be aware that it is somewhat controversial to compute summary statistics for Likert-type items, as some theorize that Likert-type items should be treated as categorical only (Sauro, 2016b; Lindeløv, 2018). Relevant, credible, and visual variables were scored by the participants on a five-point scale. The mean for all of the ratings for credible is a little lower than the means for relevant and visual. Quickly find, refine and opinion were rated by the participants using a seven-point scale.

The mean rating for opinion is slightly lower than the means for the refine and quickly find variables. Perhaps the overall opinion of the screens is being brought down by other factors considered by the participants or some of the other quality measures.

Appendix B Figures B5-B10 show the frequency distribution of the ratings for each variable (relevant, credible, visual appeal, quickly find, refine and opinion) respectively. The variables relevant (see Appendix B Figure B5) and visual appeal (see Appendix B Figure B9) are normally distributed. While the other variables all resemble a normal distribution, they have differences from a normal distribution. As seen in Appendix B Figure B6, credible is skewed, which might be why its mean is lower than the other five-point-scale variables. Quickly find (see Appendix B Figure B7) and refine (see Appendix B Figure B8) are also skewed. When looking at Appendix B Figure B10, one can see that opinion is close to resembling a bimodal distribution.

Appendix B Figures B11-B16 show the relative frequency distributions for each of the six measures, by display screen type and by each level of each variable. For the variable relevant (see Appendix B Figure B11), the ratings of “4-Very helpful” and “5-Extremely helpful” appear more frequently for Screen 1 and Screen 2 than for Screen 3 and Screen 4. For the credible variable (Appendix B Figure B12), none of the display screens had a high frequency of better ratings. As shown in Appendix B Figure B13 for quickly find, the participants often gave high ratings to Screens 1, 2, and 4, but did not often give high ratings to Screen 3. For the refine variable (see Appendix B Figure B14), the participants tended to give better ratings to Screens 1 and 2. As seen in Appendix B Figure B15 for visual appeal, the ratings seemed to center around average for each screen, but Screen 3 had a high frequency of low ratings. In Appendix B Figure B16, for overall opinion, one can see that Screen 3 again commonly received low ratings, while Screen 1 and 2 both had more instances of higher ratings than Screen 4.

One method used to analyze Likert-type ratings is to aggregate them into a total score in order to get some sort of a composite rating for comparing the experimental conditions against one another conditions (Babbie, 2004, p. 169-70; Lindeløv, 2018; Sauro, 2016b;). Some statisticians who believe that Likert-type data should be always treated as categorical (Sauro, 2016b; Lindeløv, 2018) would obviously not support any sort of arithmetic aggregation. But, some statisticians do advocate combining Likert-type items for composite analysis under some conditions (Babbie, 2004, p. 169-70; Lindeløv, 2018; Sauro, 2016b;). In this case, the researcher for Non-Biomedical IRB Study Number 18-2487 has combined the data for visualization purposes. No attempts should be made to interpret the "totals" and "averages" as meaningful numbers. They are just an interim tool to get to a visualization of the composite ratings of the participants for the screen types by viewing their relative positions against each other.

Appendix B Figure B17 is the visualization showing the composite rating for each screen. Each "set" of ratings (the ratings for one screen for one health topic from one participant) was added up to get a total rating for that "set". Then, all of the composites were averaged to find the relative total measure for each search result display screen. Based on Appendix B Figure B17, Screen 1 (control) and Screen 2 appear to have received better ratings than Screen 3 and Screen 4, with Screen 3 having the lowest ratings of all.

The researcher also recoded the Likert-type questions to collapse into just the levels of negative, neutral and positive to allow further analysis (Babbie, 2004, pp. 408-

409). Appendix A Table A6 shows how the data from the present study was re-coded in the Part 1 codebook, which is a list of variables and possible responses (Babbie, 2004, pp. 399-400). Appendix A Table A7 shows the ratings by screen for the variables using the collapsed response values. For the relevant variable, it is easy to see that Screen 1 (control) and Screen 2 received more positive ratings, but the ratings for Screen 3 and were more evenly distributed. Appendix B Figure B18 is a visualization of the recoded data. Screens 1 and 2 have far more positive ratings than negative ratings. Screen 4 is equally balanced.

In Appendix A Table A8, the Pearson's Chi-Square is reported for all six variables. The distributions of the ratings for each of the six variables are statistically significant at the 0.05 significance level. The null hypothesis of random results can be rejected. The distributions of the ratings are associated with the type of screen, instead of just being by random chance.

Part 2 Results

In Part 2, participants were asked to answer these three questions (presented in random order):

- Overall, which type of display did you prefer the most for viewing search engine results from health information searches?
- Which type of display did you dislike the most for viewing search engine results from health information searches?
- If a browser extension or other customization was available to ensure that your search engine results from health information searches would appear like any of these displays, which option would you pick?

Appendix A Table A9 shows the results for all three concluding questions. No participants skipped any questions in Part 2, so all counts total 199. Results are mostly similar across assigned topics. For instance, Screen 1 (control) was the selection for overall preferred search engine result display by 44.12% of participants assigned Topics 1 and 2, by 37.68% of participants assigned Topics 2 and 3, and by 43.55% of users assigned Topics 3 and 1. But, Screen 2 was the most preferred screen by participants assigned Topics 2 and 3 and by participants with Topics 3 and 1, but not by participants with Topics 1 and 2.

Appendix B Figure B19, Appendix B Figure B20, and Appendix B Figure B21 are relative frequency distributions for the prefer, browser extension and the dislike questions, respectively. Participants who saw the Topics 2 and 3 block or the Topics 3 and 1 block selected Screen 2 as the display screen that they would prefer, but the rate of participants who selected Screen 1 (control) was only slightly lower. For the overall totals, Screen 2 was selected as the overall preferred search results display screen (48.24% vs. 41.71%). Because Google is a well-regarded search engine in the United States (Jansen, Zhang, and Schultz 2009), any search results display screen that could be identified as preferred about the same as Google would be noteworthy. Screen 2 received the most responses when participants were asked to pick a display screen that they would want to use as a browser extension or other customization of their search results. However, some users might have noticed that Screen 1 was the default Google search results (with ads, videos, and images removed), so they might have felt that a customization or browser extension was not needed to obtain search engine results similar to Screen 1.

Appendix B Figure B21 summarizes the results of the question asking participants to select the search engine results screen that they dislike the most. Screen 3 was especially disliked. The other screen containing a word cloud, Screen 4, is in second place for dislike. Some studies have found that word clouds enhanced the search process either in speed in completing tasks (Gwizdka, 2009) or in user satisfaction (Kuo, Hentrich, Good, & Wilkinson, 2007), but the present study suggests that users may not want word clouds in their search engine results. Another explanation could be that screen shots of Screen 3 had fewer search results than Screens 1, 2, or 4 (because the search results become too long with both a word cloud and a standard search engine result listing for each result). Perhaps the participants reacted to having fewer search results displayed. Appendix B Figure B21 also depicts the percentage of participants who disliked either Search 1 (control) or Screen 2, which are roughly equal in number of participants who selected these two screens as the most disliked. Again, the counts for dislike of Screen 2 and Screen 1 (control) are very close, suggesting that users might view both screen types in similar lights.

As shown in Appendix A Table A10, 63.86% of the 83 participants who preferred Screen 1 (control) also selected Screen 1 (control) for their browser extension or customization. Co-occurrence of Screen 2 as both the preferred screen and as the search results display screen selected for a browser extension or customization was 94.79% (91 out of 96 participants). In Appendix A Table A11, co-occurrence of the screen display type selected for the browser extension or customization screen type and selected as the overall preferred screen is shown. Of the 58 participants who selected Screen 1 (control) for their browser extension, 91.38% (53 participants) also selected Screen 1 as their overall preferred display screen type for presentation of search results. Although Screens 3 and 4 were not commonly chosen for either the preferred screen type or the browser extension screen type, it is important to note that they follow the general trend of co-occurring together. For instance, 81.82% of the participants who selected screen 3 for their browser extension or customization also selected screen 3 for their preferred search engine results screen type.

In looking at co-occurrences surrounding the dislike variable, five observations were noted as unusual because the participant selected the same display screen as the search result screen that they preferred and as the search result screen that they disliked. Three of these instances involved participants who selected the same display screen for all three variables (prefer, dislike and browser). A possible interpretation was that the participant did not read the questions carefully and thought that all questions were asking to pick the display screen for which he/she had positive feelings toward. One participant chose Display Screen 2 for both the preferred screen and the disliked screen, but selected Display Screen 4 for a potential browser extension. Similarly, one participant chose Screen 2 for a browser extension, but picked Screen 1 (control) for the preferred and disliked browser. The researcher did not exclude any observations or adjust any data when these unusual responses were encountered. They are all included in the dataset as valid data. While analyzing the dislike variable, the researcher noted that out of the nine participants who disliked Screen 1 (control), four selected Display Screen 3, which was otherwise mostly unpopular, as the display screen that they prefer overall.

Another way to represent the data from Part 2 of the survey is in binomial (or dummy variable) format. Appendix A Table A12 includes the summary frequencies in this format and gives an extra way to consider the data. For instance, 116 participants did not select Screen 1 (control) as their overall preferred screen type, but only 103 participants did not select Screen 2 as their overall preferred screen type. Also, 187 out of 199 participants (93.97%) did not select Screen 3 as their overall preferred screen display type, which is another indication that Screen 3 was not well-liked by the participants.

In looking at the cumulative frequencies for the preferred display screens and the screen(s) desired as a browser extension or customization in Appendix A Table A13, one can note that most of the responses were distributed in Screens 1 and 2 (around 89-90% of the responses). Similarly, for the dislike question, only around 14.5% of the cumulative frequency appears in screens 1 and 2 as seen in Appendix A Table A13.

In Appendix A Table A14, results of the Chi-Square Test for Equal Proportions are shown for the prefer, browser extension or customization, and dislike variables. At significance level 0.05, the null hypothesis of equal proportions can be rejected for each of the three variables (prefer, browser, and dislike). In other words, the distribution of the search result screen type selections was not due to random chance. Appendix B Figure B22, Appendix B Figure B23, and Appendix B Figure B24 show the frequency distributions of the variables prefer, browser, and dislike respectively.

Discussion

Limitations of Results and Analysis

When using a within-subjects design, the standard practice is to use a repeated measures statistical analysis technique (C. Wiesen, The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, personal communication, January 4, 2019; "Repeated Measures Designs," 2015; Wildemuth, 2016, p. 398). A repeated measures technique should have been used for analyzing the Likert-type ratings (C. Wiesen, The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, personal communication, January 4, 2019). A repeated measures technique was not applicable for the overall ratings for the preferred screen, the screen to use as a browser extension and the disliked screen, because each participant only answered each question once, which pertained to both health topics on which he or she viewed the screens (C. Wiesen, The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, personal communication, January 4, 2019). However, due to time constraints and limited statistical training, the repeated measures statistical technique was not used. In the case of the data from the present study, the results are unlikely to change if a repeated measures technique was added to the statistical analysis (C. Wiesen, The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, personal communication, January 4, 2019). However, it is a limitation to the results analysis for Part 1.

The variables were all assumed to be independent, but it is likely that they are not independent. Correlation of the variables was not considered or addressed, which could have led to valuable insights. A very preliminary correlation analysis was performed in SAS Enterprise Guide (Version 7.15 HF7 7.100.5.6177 64-bit). Results of the preliminary correlation analysis are shown in Appendix H. However, the methods used may not be the optimal algorithms for the data. Another challenge for evaluating the correlation more thoroughly included how the data is structured (binomial vs. categorical vs. ordinal) from this study.

The researcher strived to utilize sound statistical concepts and employ appropriate statistical techniques. However, the researcher does not have an extensive background in statistical analysis of survey data. It is possible that better algorithms could have been used to provide more accuracy of the analysis. The data was not all normally distributed, which might have required alterations to the analytical techniques.

It was beyond the scope of this project to go beyond identifying that a statistically significant association exists between the search engine display types and the evaluation variables. However, the best practice would be to use additional techniques to attempt to identify the specific relationships that exist in the data and to try to measure the magnitude of the relationship.

Procedural Limitations

Because this was a web survey instead of an in-person evaluation of the search engine results with the ability to interact with the search engine results, the study was somewhat unnatural "… and therefore may not fully reflect individuals' natural searching behavior." (Crystal & Greenberg, 2006, p.1381). The design of this study did not allow the participants to create their own queries, which is a best practice (Koshman, 2005). The refining instructions that were necessary to provide to the participants, despite efforts to not introduce other variables, could still have impacted the results. Ideally, the participants could have used the search engine display screens themselves in order to learn and comprehend for themselves how to refine the queries.

The question asking participants to indicate which search engine results search screen they would select for a browser extension or customization might have been interpreted by participants who recognized Screen 1 as the default Google search results, as indicating that Screen 1 was not available as a customization or that it was unnecessary as a browser extension. Screen shots of Screen 3 had fewer search results than Screens 1, 2, or 4, because the screen results become too long with both a word cloud and a standard search engine result listing for each result. Perhaps the participants reacted to having fewer search results displayed. Ideally, the same number of search engine results could have been displayed for all screen types while still allowing the screen shot to be the same size (but not zoomed out or in).

Another limitation was that study participants may not have had interest in the health topics which they were shown. Freeman and Spyridakis found that greater interest in the topic of the health information led to higher ratings of credibility (2004). Cunningham and Johnson found that participants judged health information by whether they could relate or identify with it (2016), but it is unknown if that effect could carry over to their evaluations of the search results screens.

Although Qualtrics has functionality that attempts to keep random assignments of elements such as search topics approximately equal, sometimes when participants start the study and don't complete it, the elements can become unbalanced in the final results of completed surveys. There is additional functionality where the survey designer can manually rebalance the elements, such as search topics, during the course of the survey, but this was not utilized during this survey. A procedural improvement might utilize the manual rebalancing.

By using a convenience sample, a fully representative sample of the target population of health information searchers was not achieved. Ideally, a sampling method other than a convenience sample would have been used.

Other Study Weaknesses

As discussed in the section about the demographics of the participants, the survey sample was not racially and ethnically representative of the population of Internet health information searchers, which could affect the results and make them less representative of the views of the actual target population. In American adults, there are not broad differences in the rates by which ethnic groups use the Internet to search for health information online, with 73% of the Caucasian or White racial/ethnic group, 69% of the Black or African American racial/ethnic group, and 66% of the Hispanic or Latino/Latina ethnic groups conducting searches for health information online (Fox & Duggan, 2013). If the recruiting for the study had successfully obtained a representative distribution of the U.S. population, it might have been fairly easy obtain a representative distribution of Internet health information searchers.

Additionally, the experience of participants in Part 1 of the study could have affected their responses in Part 2, but this was not factored into the data analysis and is another area for future study.

Summary of Results

In Part 1, the distributions of the ratings for each of the six variables (relevant, credible, quickly find, refine, visual appeal, and overall opinion) are statistically significant at the 0.05 significance level. The distributions of the ratings are associated with the type of screen, instead of just being by random chance. In Part 2, at significance level 0.05, the null hypothesis of equal proportions can be rejected for each of the three variables (prefer, browser, and dislike). In other words, the distribution of the search result screen type selections was not due to random chance.

The participants seemed to like Screen 1 (control) and Screen 2 better than Screen 3 or Screen 4, as evidenced by the ratings for the six measures in Part 1 and by the choices of the preferred display screen and desired browser extension in Part 2. Screen 3 received the lowest ratings in Part 1 and was chosen as the disliked search result display screen most often in Part 2. Screen 4 appeared to be rated neutrally in Part 1. In Part 2, Screen 4 was rarely chosen as the preferred display screen or for the browser extension, but it was also chosen as the disliked display screen only a moderate amount of times. Perhaps Screen 4 was also viewed neutrally in Part 2. Overall, the researcher did not detect any large inconsistencies between the results in Part 1 and Part 2.

In some ways, the results agree with similar studies, but there are a few contradictions. In a study in which participants used a faceted browsing/searching tool similar to Screen 2 and a tool with a word cloud similar to Screen 4, participants gave both the faceted searching tool and the system with the sidebar word cloud mostly positive ratings for being helpful in finding information and easy to learn (Hernández, Sharit, Pirolli, & Czaja, 2018), which is consistent with the high ratings and frequent preferences for Screen 2 noted in Non-Biomedical IRB Study Number 18-2487. However, the participants' assessments of Screen 4 perhaps were inconsistent with the Hernández, Sharit, Pirolli, and Czaja study. Dunaiski, Greene, and Fischer tested a search interface with a tag cloud used for filtering and exploring search results, finding that many participants rated the tag cloud useful and easy to use, although some felt neutral or negative (2017). However, the results from Non-Biomedical IRB Study Number 18-2487 suggested a strong negative assessment of the word cloud screens, especially Screen 3. The positive evaluations of Screen 2 by participants in Non-

Biomedical IRB Study Number 18-2487 are consistent with the results from a study in which health information searchers responded positively to faceted browsing elements that allowed them to consider related, but previously unconsidered topics (Pang, Chang, Verspoor, & Pearce, 2016).

Directions for Future Study

To the researcher, the results from Non-Biomedical IRB Study Number 18-2487 answered some questions, but brought many new potential research questions to light, including:

- Although Google is dominant, would or could an alternative search engine results display type be accepted by consumers in general and specifically, by Internet health information searchers?
- If additional statistical testing and further studies found that Screen 2 is liked equally as well as Screen 1, then what are the implications? Should industry and academic experts attempt to steer away from the Google way of presenting search engine results? Or present both options (Screen 1 and Screen 2) for display choices?
- Are these results specific to health information searching?
- Do health information searchers have special or different needs in search engine results presentation?
- What quality measure should be used for evaluating search engine results?
- What combination of factors leads to picking a screen as the best? Or the worst?
- Why did some participants pick a display screen as their preferred screen overall, but not choose it for their desired browser extension? And, why did some

participants pick a display screen for their desired browser extension, but not as their preferred screen overall?

- In the observations where a screen type was both chosen as a browser extension and as the disliked screen type, were these results just errors or noise?
- How did Screen 3 showing fewer search results than Screens 1, 2, or 4 affect the participants' evaluations of Screen 3?
- What did the participants dislike so much in Screen 3?
- What comments and insights would the participants who selected Screen 3 for their preferred screen overall and their browser extension offer? What did they see in Screen 3 that other participants might have overlooked or not valued as much?
- If Screen 1 (control) was not an option, what display screen would have been the preferred display screen and the preferred browser extension for the participants who chose Screen 1 (control) in this study?
- Were the nine participants who disliked Screen 1 (control), and who picked display screen 3, which was otherwise mostly unpopular, as the display screen that they prefer overall, part of a trend or just noise in the data?
- Could machine learning be applied to search engine display preferences? For instance, could screen types be clustered together in a recommendation system, such as "If you like screen display X and screen display Y, then you might also like screen display Z"?

References

Abbasi, A., Zahedi, F. “Mariam,” & Kaza, S. (2012). Detecting Fake Medical Web Sites Using Recursive Trust Labeling. *ACM Transactions on Information Systems*, *30*(4), 1–36. https://doi.org/10.1145/2382438.2382441

Adams, S. A. (2010). Revisiting the online health information reliability debate in the wake of “web 2.0”: An inter-disciplinary literature and website review. *International Journal of Medical Informatics*, *79*(6), 391–400. https://doi.org/10.1016/j.ijmedinf.2010.01.006

Ageev, M., Lagun, D., & Agichtein, E. (2013). Improving Search Result Summaries by Using Searcher Behavior Data. In *SIGIR ’13: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28–August 1, 2013, Dublin, Ireland* (pp. 13–22). ACM. https://doi.org/10.1145/2484028.2484093

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results (p. 5). ACM Press. https://doi.org/10.1145/1498759.1498766

Aguirre, P. E. A., Coelho, M. M., Rios, D., Machado, M. A. A. M., Cruvinel, A. F. P., & Cruvinel, T. (2017). Evaluating the Dental Caries-Related Information on Brazilian Websites: Qualitative Study. *Journal of Medical Internet Research*, *19*(12), e415. https://doi.org/10.2196/jmir.7681

Al Maqbali, H., Scholer, F., James, S., Thom, J., & Wu, M. (2009). Do Users Find Looking at Text More Useful than Visual Representations? A Comparison of Three Search Result Interfaces. In Judy Kay, Paul Thomas, & Andrew Trotman (Eds.), *Proceedings of the 14th Australasian Document Computing Symposium (ADCS 2009), Sydney, Australia,* *December 4, 2009*. Retrieved from http://www.cs.otago.ac.nz/homepages/andrew/involvement/2009-ADCS.pdf

Al-Maqbali, H., Scholer, F., Thom, J. A., & Wu, M. (2010). Evaluating the Effectiveness of Visual Summaries for Web Search. In F. Scholer, A. Trotman, & A. Turpin (Eds.), *Proceedings of the 15th Australasian Document Computing Symposium (ADCS 00), Melbourne, Australia, December 10, 2010*. Melbourne, Australia: RMIT University. Retrieved from http://www.cs.rmit.edu.au/adcs2010/proceedings/

AlGhamdi, K. M., & Moussa, N. A. (2012). Internet use by the public to search for health-related information. *International Journal of Medical Informatics*, *81*(6), 363–373. https://doi.org/10.1016/j.ijmedinf.2011.12.004

Al-Maskari, A., Sanderson, M., & Clough, P. (2007). The Relationship Between IR Effectiveness Measures and User Satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), Amsterdam, The Netherlands, July 23-27, 2007* (pp. 773–774). ACM. https://doi.org/10.1145/1277741.1277902

Allam, A., Schulz, P. J., & Nakamoto, K. (2014). The Impact of Search Engine Selection and Sorting Criteria on Vaccination Beliefs and Attitudes: Two Experiments Manipulating Google Output. *Journal of Medical Internet Research*, *16*(4), e100, 1–20. https://doi.org/10.2196/jmir.2642

Anagnostopoulos, A., Becchetti, L., Castillo, C., & Gionis, A. (2010). An optimization framework for query recommendation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10), New York City, New York,* *USA, February 4–6, 2010* (pp. 161–170). ACM Press. https://doi.org/10.1145/1718487.1718508

Anderson, M., & Perrin, A. (2017). *Tech adoption climbs among older adults*. Washington, D.C., USA: Pew Research Center. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/14/2017/05/16170850/PI_2017.05.17_Older-Americans-Tech_FINAL.pdf

Arguello, J., & Capra, R. (2014). The Effects of Vertical Rank and Border on Aggregated Search Coherence and Search Behavior. In J. Li & X. S. Wang (Eds.), *CIKM'14: proceedings of the 23rd ACM International Conference on Information and Knowledge Management: November 3-7, 2014, Shanghai, China* (pp. 539–548). ACM. https://doi.org/10.1145/2661829.2661930

Arguello, J., & Capra, R. (2016). The Effects of Aggregated Search Coherence on Search Behavior. *ACM Transactions on Information Systems*, *35*(1), 1–30. https://doi.org/10.1145/2935747

Aslam, R., Gibbons, D., & Ghezzi, P. (2017). Online Information on Antioxidants: Information Quality Indicators, Commercial Interests, and Ranking by Google. *Frontiers in Public Health*, *5*. https://doi.org/10.3389/fpubh.2017.00090

Atsaros, G., Spinellis, D., & Louridas, P. (2008). Site-Specific versus General-purpose Web Search Engines: A Comparative Evaluation. In S. Gritzalis, D. Plexousakis, & D. Pnevmatikatos (Eds.), *PCI 2008: 12th Panhellenic Conference on Informatics* (pp. 44–48). Los Alamitos, CA: IEEE Computer Society. https://doi.org/10.1109/PCI.2008.47

Aula, A., Khan, R. M., Guan, Z., Fontes, P., & Hong, P. (2010). A comparison of visual and textual page previews in judging the helpfulness of web pages. In *WWW '10:*

Proceedings of the 19th International Conference on World Wide Web (pp. 51–60). ACM Press. https://doi.org/10.1145/1772690.1772697

Azzopardi, L., Kelly, D., & Brennan, K. (2013). How Query Cost Affects Search Behavior. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 23–32). Dublin, Ireland: ACM. https://doi.org/10.1145/2484028.2484049

Babbie, E. R. (2004). *The practice of social research*. Belmont, CA: Thomson/Wadsworth.

Bade, D. (2007). Relevance ranking is not relevance ranking or, when the user is not the user, the search results are not search results. *Online Information Review*, *31*(6), 831–844. https://doi.org/10.1108/14684520710841793

Bailey, P., Craswell, N., White, R. W., Chen, L., Satyanarayana, A., & Tahaghoghi, S. M. M. (2010). Evaluating search systems using result page context. In *IIiX '10: Proceedings of the third symposium on Information interaction in context* (p. 105). New Brunswick, New Jersey, USA. https://doi.org/10.1145/1840784.1840801

Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results—A user study. *Journal of the American Society for Information Science and Technology*, *60*(1), 135–149. https://doi.org/10.1002/asi.20941

Bar-Ilan, J., Keenoy, K., Yaari, E., & Levene, M. (2007). User rankings of search engine results. *Journal of the American Society for Information Science and Technology*, *58*(9), 1254–1266. https://doi.org/10.1002/asi.20608

Bateman, S., Gutwin, C., & Nacenta, M. (2008). Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections. In *Proceedings of the Nineteenth ACM*

Conference on Hypertext and Hypermedia (pp. 193–202). New York, NY, USA: ACM. https://doi.org/10.1145/1379092.1379130

Bates, B. R., Romina, S., Ahmed, R., & Hopson, D. (2006). The effect of source credibility on consumers' perceptions of the quality of health information on the Internet. *Medical Informatics and the Internet in Medicine*, *31*(1), 45–52. https://doi.org/10.1080/14639230600552601

Bates, M. J. (1989). The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, *13*(5), 407–424.

Bedell, S. E., Agrawal, A., & Petersen, L. E. (2004). A systematic critique of diabetes on the world wide web for patients and their physicians. *International Journal of Medical Informatics*, *73*(9), 687–694. https://doi.org/10.1016/j.ijmedinf.2004.04.011

Benotsch, E. G., Kalichman, S., & Weinhardt, L. S. (2004). HIV-AIDS Patients' Evaluation of Health Information on the Internet: The Digital Divide and Vulnerability to Fraudulent Claims. *Journal of Consulting and Clinical Psychology*, *72*(6), 1004–1011. https://doi.org/10.1037/0022-006X.72.6.1004

Bernstam, E. V., Walji, M. F., Sagaram, S., Sagaram, D., Johnson, C. W., & Meric-Bernstam, F. (2008). Commonly cited website quality criteria are not effective at identifying inaccurate online information about breast cancer. *Cancer*, *112*(6), 1206–1213. https://doi.org/10.1002/cncr.23308

Biggs, T. C., Bird, J. H., Harries, P. G., & Salib, R. J. (2013). YouTube as a source of information on rhinosinusitis: the good, the bad and the ugly. *The Journal of Laryngology and Otology; Devon*, *127*(8), 749–754. http://dx.doi.org.libproxy.lib.unc.edu/10.1017/S0022215113001473

Blignaut, P., & McDonald, T. (2012). The effect of experience and socio-economic status on web searching performance: a South African perspective. *Behaviour & Information Technology*, *31*(4), 401–412. https://doi.org/10.1080/01449290903488140

Brady, E., Segar, J., & Sanders, C. (2016). “You get to know the people and whether they’re talking sense or not”: Negotiating trust on health-related forums. *Social Science & Medicine*, *162*, 151–157. https://doi.org/10.1016/j.socscimed.2016.06.029

Braun, T., Klingen, J., & Krom, R. (n.d.). Greenshot (Version 1.2.10 64 bit) [Software]. Retrieved from http://getsgreenshot.org

Brinich, M. A., Mercer, M. B., & Sharp, R. R. (2013). An analysis of online messages about probiotics. *BMC Gastroenterology*, *13*(1), 5. https://doi.org/10.1186/1471-230X-13-5

Broder, A. (2002). A Taxonomy of Web Search. *SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552

Burt, M., & Liew, C. L. (2012). Searching with clustering: An investigation into the effects on users’ search experience and satisfaction. *Online Information Review*, *36*(2), 278–298. https://doi.org/10.1108/14684521211229075

Buscher, G., White, R. W., Dumais, S., & Huang, J. (2012). Large-scale Analysis of Individual and Task Differences in Search Result Page Examination Strategies. In *Proceedings of WSDM’12 Fifth ACM International Conference on Web Search and Data Mining, Seattle, Washington, USA, February 8–12, 2012* (pp. 373–382). ACM. https://doi.org/10.1145/2124295.2124341

Capra, R., Arguello, J., & Scholer, F. (2013). Augmenting web search surrogates with images. In *Proceedings of the 22nd ACM International Conference on Information and*

Knowledge Management (CIKM ’13), San Francisco, CA, USA, October 27-November 1, 2013 (pp. 399–408). ACM Press. https://doi.org/10.1145/2505515.2505714

Carpineto, C., Osiński, S., Romano, G., & Weiss, D. (2009). A survey of Web clustering engines. *ACM Computing Surveys (CSUR)*, *41*(3), 17:1–17:38. https://doi.org/10.1145/1541880.1541884

Chang, P., Hou, I.-C., Hsu, C.-L., & Lai, H.-F. (2006). Are Google or Yahoo a Good Portal for Getting Quality Healthcare Web Information? In *AMIA Annual Symposium (AMIA 2006) Proceedings, Washington, DC, USA, November 11-15, 2006* (p. 878). American Medical Informatics Association. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839683/

Charnock, D., Shepperd, S., Needham, G., & Gann, R. (1998). *The DISCERN handbook: quality criteria for consumer health information ; user guide and training resource*. Abingdon: Radcliff. Retrieved from http://www.discern.org.uk/discern.pdf

Charnock, D., Shepperd, S., Needham, G., & Gann, R. (1999). DISCERN: An Instrument for Judging the Quality of Written Consumer Health Information on Treatment Choices. *Journal of Epidemiology and Community Health*, *53*(2), 105–111.

Chau, M. (2011). Visualizing web search results using glyphs: Design and evaluation of a flower metaphor. *ACM Transactions on Management Information Systems*, *2*(1, Article 2), 1–27. https://doi.org/10.1145/1929916.1929918

Chen, H., & Dumais, S. (2000). Bringing order to the Web: automatically categorizing search results. In *CHI ’00: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, The Hague, The Netherlands, April 1-6, 2000* (pp. 145–152). ACM Press. https://doi.org/10.1145/332040.332418

Chilton, L. B., & Teevan, J. (2011). Addressing People's Information Needs Directly in a Web Search Result Page. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web* (pp. 27–36). Hyderabad, India. https://doi.org/10.1145/1963405.1963413

Christensson, P. (2008, January 14). Web 2.0 Definition. *TechTerms*. Retrieved from https://techterms.com/definition/web20

Chuklin, A., & de Rijke, M. (2014). The Anatomy of Relevance: Topical, Snippet and Perceived Relevance in Search Result Evaluation. Presented at the SIGIR 2014 Workshop on Gathering Efficient Assessments of Relevance, Gold Coast, Queensland, Australia. Retrieved from https://arxiv.org/pdf/1501.06412.pdf

Commission of the European Communities, Brussels. (2002). eEurope 2002: Quality Criteria for Health related Websites. *Journal of Medical Internet Research*, *4*(3), e15. https://doi.org/10.2196/jmir.4.3.e15

Costello, K. L. (2015). Similarity as a credibility cue in online support groups for chronic kidney disease. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–4. https://doi.org/10.1002/meet.2014.14505101105

Crystal, A., & Greenberg, J. (2006). Relevance Criteria Identified by Health Information Users During Web Searches. *Journal of the American Society for Information Science & Technology*, *57*(10), 1368–1382. https://doi.org/10.1002/asi.20436

Cunningham, A., & Johnson, F. (2016). Exploring trust in online health information: a study of user experiences of patients.co.uk. *Health Information & Libraries Journal*, *33*(4), 323–328. https://doi.org/10.1111/hir.12163

Dalmer, N. K. (2017). Questioning reliability assessments of health information on social media. *Journal of the Medical Library Association*, *105*(1), 61–68. http://dx.doi.org.libproxy.lib.unc.edu/10.5195/jmla.2017.108

Diviani, N., van den Putte, B., Meppelink, C. S., & van Weert, J. C. M. (2016). Exploring the role of health literacy in the evaluation of online health information: Insights from a mixed-methods study. *Patient Education and Counseling*, *99*(6), 1017–1025. https://doi.org/10.1016/j.pec.2016.01.007

Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01), Seattle, Washington, USA, March 31-April 05, 2001* (pp. 277–284). ACM Press. https://doi.org/10.1145/365024.365116

Dunaiski, M., Greene, G. J., & Fischer, B. (2017). Exploratory search of academic publication and citation data using interactive tag cloud visualizations. *Scientometrics*, *110*(3), 1539–1571. https://doi.org/10.1007/s11192-016-2236-3

Dutta-Bergman, M. J. (2004). The Impact of Completeness and Web Use Motivation on the Credibility of e-Health Information. *Journal of Communication*, *54*(2), 253–269. https://doi.org/10.1111/j.1460-2466.2004.tb02627.x

Dutton, W. H., & Blank, G. (2013). *Cultures of the Internet: The Internet in Britain* (with Groselj, D.). Oxford Internet Survey (OxIS) 2013. Oxford Internet Institute, University of Oxford. Retrieved from http://oxis.oii.ox.ac.uk/wp-content/uploads/sites/43/2014/11/OxIS-2013.pdf

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ*, *324*(7337), 573–577. https://doi.org/10.1136/bmj.324.7337.573

Fattahi, R., Parirokh, M., Dayyani, M. H., Khosravi, A., & Zareivenovel, M. (2016). Effectiveness of Google keyword suggestion on users' relevance judgment: A mixed method approach to query expansion. *The Electronic Library*, *34*(2), 302–314. https://doi.org/10.1108/EL-03-2015-0035

FDA Office of Media Affairs. (2018, February 13). BeSafeRx: Know Your Online Pharmacy - For the Media [WebContent]. Retrieved from https://www.fda.gov/Drugs/ResourcesForYou/Consumers/BuyingUsingMedicineSafely/BuyingMedicinesOvertheInternet/BeSafeRxKnowYourOnlinePharmacy/ucm294170.htm#FAQs

Feild, H. A., Allan, J., & Jones, R. (2010). Predicting searcher frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10), Geneva, Switzerland, July 19–23, 2010* (p. 34). ACM Press. https://doi.org/10.1145/1835449.1835458

Feufel, M. A., & Stahl, S. F. (2012). What do Web-Use Skill Differences Imply for Online Health Information Searches? *Journal of Medical Internet Research*, *14*(3), e87. https://doi.org/10.2196/jmir.2051

Fidgeon, T. (2011, March 14). Usability Testing: How to statistically compare the usability of different designs. Retrieved from https://www.spotless.co.uk/insights/comparing-usability/

Flanagin, A. J., Hocevar, K. P., & Samahito, S. N. (2014). Connecting with the user-generated Web: how group identification impacts online information sharing and evaluation.

Information, Communication & Society, *17*(6), 683–694. https://doi.org/10.1080/1369118X.2013.808361

Fogg, B. J., Swani, P., Treinen, M., Marshall, J., Laraki, O., Osipovich, A., … Shon, J. (2001). What makes Web sites credible?: a report on a large quantitative study. In *CHI '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA, March 31-April 4, 2001* (pp. 61–68). ACM Press. https://doi.org/10.1145/365024.365037

Fox, S. (2011). *The Social Life of Health Information, 2011*. Washington, D.C., USA: Pew Research Center's Internet & American Life Project. Retrieved from www.pewinternet.org/files/old-media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf

Fox, S., & Duggan, M. (2013). *Health Online 2013*. Washington, D.C., USA: Pew Research Center's Internet & American Life Project. Retrieved from http://www.pewinternet.org/files/old-media//Files/Reports/PIP_HealthOnline.pdf

Fox, S., & Rainie, L. (2002). *Vital decisions: How Internet users decide what information to trust when they or their loved ones are sick: Plus a Guide from the Medical Library Association about Smart Health-search Strategies and Good Web Sites*. Washington, D.C., USA: Pew Internet & American Life Project. Retrieved from http://www.pewinternet.org/files/old-media//Files/Reports/2002/PIP_Vital_Decisions_May2002.pdf.pdf

Franzreb, D., & Franzreb, P. (2016, June 21). Setting Standards for Usability Testing | UX Booth. Retrieved January 7, 2019, from https://www.uxbooth.com/articles/setting-standards-for-usability-testing/

Freeman, K. S., & Spyridakis, J. H. (2004). An Examination of Factors That Affect the Credibility of Online Health Information. *Technical Communication; Washington*, *51*(2), 239–263.

Gauld, R., & Williams, S. (2009). Use of the internet for health information: A study of Australians and New Zealanders. *Informatics for Health and Social Care*, *34*(3), 149–158. https://doi.org/10.1080/17538150903102448

González-Ibáñez, R., Proaño-Ríos, V., Fuenzalida, G., & Martinez-Ramirez, G. (2017). Effects of a visual representation of search engine results on performance, user experience and effort. In *Proceedings of the 80th Annual Meeting of the Association for Information Science & Technology, Washington, DC, USA, Oct. 27-Nov. 1, 2017* (Vol. 54, pp. 128–138). https://doi.org/10.1002/pra2.2017.14505401015

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking Analysis of User Behavior in WWW Search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 478–479). Sheffield, United Kingdom: ACM. https://doi.org/10.1145/1008992.1009079

Guan, Z., & Cutrell, E. (2007). An Eye Tracking Study of the Effect of Target Rank on Web Search. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California, USA, April 30 - May 03, 2007* (pp. 417–420). ACM. https://doi.org/10.1145/1240624.1240691

Guo, Q., Jin, H., Lagun, D., Yuan, S., & Agichtein, E. (2013). Towards Estimating Web Search Result Relevance from Touch Interactions on Mobile Devices. In *CHI EA '13: CHI '13 Extended Abstracts on Human Factors in Computing Systems, Paris, France,* *April 27-May 02, 2013* (pp. 1821–1826). ACM. https://doi.org/10.1145/2468356.2468683

Gwizdka, J. (2009). What a difference a tag cloud makes: effects of tasks and cognitive abilities on search results interface use. *Information Research*, *14*(4). Retrieved from http://www.informationr.net.libproxy.lib.unc.edu/ir/14-4/paper414.html

Haas, A., & Unkel, J. (2017). Ranking versus reputation: perception and effects of search result credibility. *Behaviour & Information Technology*, *36*(12), 1285–1298. https://doi.org/10.1080/0144929X.2017.1381166

Halvey, M. J., & Keane, M. T. (2007). An assessment of tag presentation techniques. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07), Banff, Alberta, Canada, May 08-12, 2007* (pp. 1313–1314). ACM Press. https://doi.org/10.1145/1242572.1242826

Harcup, T. (2014, September 18). Word cloud. *A Dictionary of Journalism*. Oxford University Press. Retrieved from http://www.oxfordreference.com.libproxy.lib.unc.edu/view/10.1093/acref/9780199646241.001.0001/acref-9780199646241-e-1499

Hariri, N. (2011). Relevance ranking on Google: Are top ranked results really considered more relevant by the users? *Online Information Review*, *35*(4), 598–610. https://doi.org/10.1108/14684521111161954

Health On the Net Foundation (HON). (2017, May 3). Health On the Net Foundation. Retrieved from http://www.hon.ch/

Hearst, M. A. (2006). Clustering Versus Faceted Categories for Information Exploration. *Communications of the ACM*, *49*(4), 59–61. https://doi.org/10.1145/1121949.1121983

Heo, M., & Hirtle, S. C. (2001). An empirical comparison of visualization tools to assist information retrieval on the web. *Journal of the American Society for Information Science and Technology*, *52*(8), 666–675. https://doi.org/10.1002/asi.1117

Hernández, M. A., Sharit, J., Pirolli, P., & Czaja, S. J. (2018). Adapting Information Search Tools for use by Health Consumers: Challenges and Lessons for Software Designers. *International Journal of Human–Computer Interaction*, *34*(5), 445–456. https://doi.org/10.1080/10447318.2017.1358546

Hoeber, O., & Yang, X. D. (2008). Evaluating WordBars in exploratory Web search scenarios. *Information Processing & Management*, *44*(2), 485–510. https://doi.org/10.1016/j.ipm.2007.07.003

Horrigan, J. B., & Duggan, M. (2015). *Home Broadband 2015*. Washington, D.C., USA: Pew Research Center. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/14/2015/12/Broadband-adoption-full.pdf

Hu, Y., & Sundar, S. S. (2010). Effects of Online Health Sources on Credibility and Behavioral Intentions. *Communication Research*, *37*(1), 105–132. https://doi.org/10.1177/0093650209351512

Humenberger, H. (2011). How does Google come to a ranked list?--making visible the mathematics of modern society. *Teaching Mathematics and Its Applications*, *30*(3), 107–119. https://doi.org/10.1093/teamat/hrr007

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248–263. https://doi.org/10.1016/j.ipm.2004.10.007

Jansen, B. J., Zhang, M., & Schultz, C. D. (2009). Brand and its effect on user perception of search engine performance. *Journal of the American Society for Information Science and Technology*, *60*(8), 1572–1595. https://doi.org/10.1002/asi.21081

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (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), 1–27. https://doi.org/10.1145/1229179.1229181

Joho, H., & Jose, J. M. (2006). A Comparative Study of the Effectiveness of Search Result Presentation on the Web. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in Information Retrieval: 28th European Conference on IR Research (ECIR 2006), London, UK, April 10-12, 2006, Proceedings* (Vol. 3936, pp. 302–313). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11735106_27

Jung, E. H., Walsh-Childers, K., & Kim, H.-S. (2016). Factors influencing the perceived credibility of diet-nutrition information web sites. *Computers in Human Behavior*, *58*, 37–47. https://doi.org/10.1016/j.chb.2015.11.044

Käki, M. (2005). Findex: Search Result Categories Help Users when Document Ranking Fails. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05), Portland, OR, USA, April 02-07, 2005* (pp. 131–140). ACM. https://doi.org/10.1145/1054972.1054991

Kaptein, R., & Kamps, J. (2011). Word Clouds of Multiple Search Results. In A. Hanbury, A. Rauber, & A. P. de Vries (Eds.), *Multidisciplinary Information Retrieval: Proceedings of Second Information Retrieval Facility Conference IRFC 2011, Vienna, Austria, June 6,* *2011* (pp. 78–93). Springer, Berlin, Heidelberg. Retrieved from https://link-springer-com.libproxy.lib.unc.edu/chapter/10.1007/978-3-642-21353-3_7

Keenan, S., & Johnston, C. (2011, November 30). Category. *Concise Dictionary of Library and Information Science*. Berlin/Boston, GERMANY: Walter de Gruyter GmbH. Retrieved from http://ebookcentral.proquest.com/lib/unc/detail.action?docID=3040998

Kelly, D., & Azzopardi, L. (2015). How many results per page?: A Study of SERP Size, Search Behavior and User Experience. In *SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 09-13, 2015* (pp. 183–192). ACM Press. https://doi.org/10.1145/2766462.2767732

Killeen, S., Hennessey, A., El Hassan, Y., Killeen, K., Clarke, N., Murray, K., & Waldron, B. (2011). Gastric cancer–related information on the Internet: incomplete, poorly accessible, and overly commercial. *The American Journal of Surgery*, *201*(2), 171–178. https://doi.org/10.1016/j.amjsurg.2009.12.015

Kitchens, B., Harle, C. A., & Li, S. (2014). Quality of health-related online search results. *Decision Support Systems*, *57*, 454–462. https://doi.org/10.1016/j.dss.2012.10.050

KNIME Analytics Platform. (n.d.). (Version 3.5.3). Zurich, Switzerland: KNIME AG. Retrieved from http://www.knime.com

Koshman, S. (2005). Testing user interaction with a prototype visualization-based information retrieval system. *Journal of the American Society for Information Science and Technology*, *56*(8), 824–833. https://doi.org/10.1002/asi.20175

Kules, B., Capra, R., Banta, M., & Sierra, T. (2009). What do exploratory searchers look at in a faceted search interface? In *JCDL'09 Proceedings of the 9th ACM/IEEE-CS joint* *conference on Digital libraries, June 15–19, 2009, Austin, Texas, USA* (p. 313). ACM Press. https://doi.org/10.1145/1555400.1555452

Kuo, B. Y.-L., Hentrich, T., Good, B. M. ., & Wilkinson, M. D. (2007). Tag clouds for summarizing web search results. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07), Banff, Alberta, Canada, May 08-12, 2007* (pp. 1203–1204). ACM Press. https://doi.org/10.1145/1242572.1242766

LaValley, S. A., Kiviniemi, M. T., & Gage-Bouchard, E. A. (2017). Where people look for online health information. *Health Information & Libraries Journal*, *34*(2), 146–155. https://doi.org/10.1111/hir.12143

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, *64*(6), 915–937. https://doi.org/10.1108/00220410810912451

Li, Y., Zhang, X., & Wang, S. (2017). Fake vs. real health information in social media in China. *Proceedings of the Association for Information Science & Technology*, *54*(1), 742–743. https://doi.org/10.1002/pra2.2017.14505401139

Liang, S. F., Devlin, S., & Tait, J. (2006). Evaluating Web Search Result Summaries. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings* (Vol. 3936, pp. 96–106). Berlin, Heidelberg: Springer, Berlin, Heidelberg. https://doi.org/10.1007/11735106_10

Liao, Q. V., & Fu, W.-T. (2014). Age differences in credibility judgments of online health information. *ACM Transactions on Computer-Human Interaction*, *21*(1), 1–23. https://doi.org/10.1145/2534410

Lindeløv, Jonas Kristoffer. (2018, April 6). Can I use parametric analyses for my Likert scales? A brief reading guide to the evidence-based answer. Retrieved from http://lindeloev.net/can-i-use-parametric-analyses-for-my-likert-scales-a-brief-reading-guide/

Liu, J. M., Xu, R. X., Hu, Y. S., Ren, L. K., Qiao, H., Ding, H., & Liu, Z. L. (2015). Chinese Internet Searches Provide Inaccurate and Misleading Information to Epilepsy Patients. *Chinese Medical Journal; Mumbai*, *128*(24). http://dx.doi.org.libproxy.lib.unc.edu/10.4103/0366-6999.171425

Liu, Y. H., & Wacholder, N. (2017). Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. *Information Processing & Management*, *53*(4), 851–870. https://doi.org/10.1016/j.ipm.2017.03.004

Losee, R. M. (2006). Is 1 noun worth 2 adjectives? Measuring relative feature utility. *Information Processing & Management*, *42*(5), 1248–1259. https://doi.org/10.1016/j.ipm.2005.11.002

Losee, R. M., & Church, Jr., L. (2005). Are two document clusters better than one? The Cluster Performance Question for information retrieval. *Journal of the American Society for Information Science and Technology*, *56*(1), 106–108. https://doi.org/10.1002/asi.20068

Loumakis, F., Stumpf, S., & Grayson, D. (2011). This image smells good: effects of image information scent in search engine results pages. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM' 11), Glasgow, Scotland, UK, October 24–28, 2011* (pp. 475–484). ACM Press. https://doi.org/10.1145/2063576.2063649

Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., … Luo, H. (2016). When does Relevance Mean Usefulness and User Satisfaction in Web Search? In *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy, July 17-21, 2016* (pp. 463–472). ACM Press. https://doi.org/10.1145/2911451.2911507

Marchionini, G. (1995). Information-seeking perspective and framework. In *Information Seeking in Electronic Environments* (Online publication date: January 2010, pp. 27–60). Cambridge, UK: Cambridge University Press. Retrieved from doi:10.1017/CBO9780511626388.004

Mitra, B., Shokouhi, M., Radlinski, F., & Hofmann, K. (2014). On user interactions with query auto-completion. In *SIGIR '14: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, QLD, Australia, July 06-11, 2014* (pp. 1055–1058). ACM Press. https://doi.org/10.1145/2600428.2609508

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers*, *10*(3), 303–320. https://doi.org/10.1016/S0953-5438(98)00012-5

Mogilner, C., Rudnick, T., & Iyengar, S. S. (2008). The Mere Categorization Effect: How the Presence of Categories Increases Choosers' Perceptions of Assortment Variety and Outcome Satisfaction. *Journal of Consumer Research*, *35*(2), 202–215. https://doi.org/10.1086/588698

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., & Hoving, C. (2013). A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and

Limitations of Social Media for Health Communication. *Journal of Medical Internet Research*, *15*(4), e85. https://doi.org/10.2196/jmir.1933

Mu, X., Lu, K., & Ryu, H. (2014). Explicitly integrating MeSH thesaurus help into health information retrieval systems: An empirical user study. *Information Processing & Management*, *50*(1), 24–40. https://doi.org/10.1016/j.ipm.2013.03.005

Nadasan, V., Vancea, G., Georgescu, P.-A., Tarcea, M., & Abram, Z. (2011). The credibility, completeness and accuracy of information about first aid in case of choking on the romanian websites. *Journal of Applied Quantitative Methods*, *6*(3), 18–26.

National Association of Boards of Pharmacy. (2017). *Internet Drug Outlet Identification Program Progress Report for State and Federal Regulators: August 2017*. Retrieved from https://nabp.pharmacy/wp-content/uploads/2016/08/Internet-Drug-Outlet-Report-August-2017.pdf

National Center for Biotechnology Information (US). (2018). *PubMed Help*. Bethesda, MD, USA: National Center for Biotechnology Information (US). Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK3827/

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education and Welfare (DHEW). (1978). The Belmont Report (PDF) (No. OS 78-0012). Washington, DC: United States Government Printing Office. Retrieved from https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf

Nguyen, H. T. L., Nakamura, K., Seino, K., & Vo, V. T. (2017). Association Between a Wider Availability of Health Information and Health Care Utilization in Vietnam: Cross-Sectional Study. *Journal of Medical Internet Research*, *19*(12), e405. https://doi.org/10.2196/jmir.8328

Niu, X., & Hemminger, B. (2015). Analyzing the interaction patterns in a faceted search interface. *Journal of the Association for Information Science & Technology*, *66*(5), 1030–1047. https://doi.org/10.1002/asi.23227

Nowicki, S. A. (2002). *Students' success with World Wide Web search engines: Retrieving relevant results with respect to end -user relevance judgments* (Ph.D.). Nova Southeastern University, United States -- Florida. Retrieved from http://search.proquest.com/pqdtglobal/docview/275875780/abstract/B6E59D13B91147B9PQ/1

O'Grady, L., Wathen, C. N., Charnaw-Burger, J., Betel, L., Shachak, A., Luke, R., … Jadad, A. R. (2012). The use of tags and tag clouds to discern credible content in online health message forums. *International Journal of Medical Informatics*, *81*(1), 36–44. https://doi.org/10.1016/j.ijmedinf.2011.10.001

Oulasvirta, A., Hukkinen, J. P., & Schwartz, B. (2009). When More is Less: The Paradox of Choice in Search Engine Use. In *SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, July 19-23, 2009* (pp. 516–523). ACM. https://doi.org/10.1145/1571941.1572030

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication*, *12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Pang, P. C. I., Chang, S., Verspoor, K., & Pearce, J. (2016). Designing Health Websites Based on Users' Web-Based Information-Seeking Behaviors: A Mixed-Method Observational

Study. *Journal of Medical Internet Research*, *18*(6), e145. https://doi.org/10.2196/jmir.5661

Patil, S., Alpert, S. R., Karat, J., & Wolf, C. (2005). “THAT’s What I Was Looking For”: Comparing User-Rated Relevance with Search Engine Rankings. In M. F. Costabile & F. Paternò (Eds.), *Human-Computer Interaction - INTERACT 2005, IFIP TC13 International Conference Proceedings* (Vol. 3585, pp. 117–129). Rome, Italy: Springer, Berlin, Heidelberg. https://doi.org/10.1007/11555261_13

Perry, J. W. (1963). Defining the Query Spectrum-The Basis for Developing and Evaluating Information-Retrieval Methods. *IEEE Transactions on Engineering Writing and Speech*, *6*(1), 20–27. https://doi.org/10.1109/TEWS.1963.4322498

Peters, E., Lipkus, I., & Diefenbach, M. A. (2006). The Functions of Affect in Health Communications and in the Construction of Health Preferences. *Journal of Communication*, *56*(s1), S140–S162. https://doi.org/10.1111/j.1460-2466.2006.00287.x

Pirolli, P., & Card, S. (1995). Information foraging in information access environments (pp. 51–58). ACM Press. https://doi.org/10.1145/223904.223911

Pu, H. (2010). User evaluation of textual results clustering for web search. *Online Information Review*, *34*(6), 855–874. https://doi.org/10.1108/14684521011099379

Qualtrics Support. (2018, June 21). Creating an Anonymized Raffle. Retrieved from https://www.qualtrics.com/support/survey-platform/common-use-cases-rc/creating-an-anonymized-raffle/

Rains, S. A., & Karmikel, C. D. (2009). Health information-seeking and perceptions of website credibility: Examining Web-use orientation, message characteristics, and structural features of websites. *Computers in Human Behavior*, *25*(2), 544–553. https://doi.org/10.1016/j.chb.2008.11.005

Repeated Measures Designs: Benefits, Challenges, and an ANOVA Example. (2015, September 17). Retrieved from http://blog.minitab.com/blog/adventures-in-statistics-2/repeated-measures-designs-benefits-challenges-and-an-anova-example

Rivadeneira, A. W., Gruen, D. M., Muller, M. J., & Millen, D. R. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, San Jose, California, USA, April 30 - May 03, 2007* (pp. 995–998). ACM Press. https://doi.org/10.1145/1240624.1240775

Robins, D., Holmes, J., & Stansbury, M. (2009). Consumer health information on the Web: The relationship of visual design and perceptions of credibility. *Journal of the American Society for Information Science and Technology*, *61*(1), 13–29. https://doi.org/10.1002/asi.21224

Sample Size Calculator by Raosoft, Inc. (2004). Retrieved May 3, 2018, from http://www.raosoft.com/samplesize.html

Sauro, J. (2015, August 25). MeasuringU: Comparing Between and Within Subjects Studies. Retrieved from https://measuringu.com/between-within/

Sauro, J. (2016, May 24). MeasuringU: Can You Take the Mean of Ordinal Data? Retrieved January 5, 2019, from https://measuringu.com/mean-ordinal/

Sauro, J. (2016, September 27). MeasuringU: 9 Biases That Affect Survey Responses. Retrieved from https://measuringu.com/survey-biases/

Savolainen, R., & Kari, J. (2006). Facing and bridging gaps in Web searching. *Information Processing & Management*, *42*(2), 519–537. https://doi.org/10.1016/j.ipm.2005.01.009

Schrammel, J., Leitner, M., & Tscheligi, M. (2009). Semantically Structured Tag Clouds: An Empirical Evaluation of Clustered Presentation Approaches. In *Proceedings of the 27th International SIGCHI Conference on Human Factors in Computing Systems (CHI 2009), Boston, MA, USA, April 4-9, 2009* (pp. 2037–2040). ACM. https://doi.org/10.1145/1518701.1519010

Shani, G., & Tractinsky, N. (2013). Displaying relevance scores for search results (p. 901). Dublin, Ireland. https://doi.org/10.1145/2484028.2484112

Shneiderman, B., Byrd, D., & Croft, W. B. (1997, January). Clarifying Search: A User-Interface Framework for Text Searches. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/january97/retrieval/01shneiderman.html

Sinclair, J., & Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, *34*(1), 15–29. https://doi.org/10.1177/0165551506078083

Spink, A., & Greisdorf, H. (2001). Regions and levels: Measuring and mapping users' relevance judgements. *Journal of the American Society for Information Science and Technology; Hoboken*, *52*(2), 161–173.

Strohmaier, M., Kröll, M., & Körner, C. (2009). Intentional query suggestion: making user goals more explicit during search. In *Proceedings of the 2009 Workshop on Web Search Click Data (WSCD '09), Barcelona, Spain, February 9, 2009* (pp. 68–74). ACM Press. https://doi.org/10.1145/1507509.1507520

Stvilia, B., Mon, L., & Yi, Y. J. (2009). A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology*, *60*(9), 1781–1791. https://doi.org/10.1002/asi.21115

Swetha, S., Uma, D., Suganya, P., Nivedhitha, V., & Saravanakumar, K. (2014). On the Performance of Medical Information Retrieval using MeSH Terms – A Survey. *Journal of Engineering Science and Technology Review*, *7*(4), 137–142.

Tang, M.-C. (2007). Browsing and searching in a faceted information space: A naturalistic study of PubMed users' interaction with a display tool. *Journal of the American Society for Information Science and Technology*, *58*(13), 1998–2006. https://doi.org/10.1002/asi.20689

Tao, D., LeRouge, C. M., Deckard, G., & De Leo, G. (2012). Consumer Perspectives on Quality Attributes in Evaluating Health Websites. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (pp. 2675–2684). Maui, HI, USA: IEEE. https://doi.org/10.1109/HICSS.2012.180

Taylor, R. S. (1968). Question-Negotiation and Information Seeking in Libraries. *College & Research Libraries*, *29*(3), 178–194. https://doi.org/10.5860/crl_29_03_178

Teevan, J., Cutrell, E., Fisher, D., Drucker, S. M., Ramos, G., André, P., & Hu, C. (2009). Visual snippets: summarizing web pages for search and revisitation. In *Proceedings of the 27th International SIGCHI Conference on Human Factors in Computing Systems (CHI 2009), Boston, MA, USA, April 4-9, 2009* (pp. 2023–2032). ACM Press. https://doi.org/10.1145/1518701.1519008

The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill,. (2012, March). Qualtrics Tips and Tricks for UNC-CH Users. Retrieved from https://odum.unc.edu/files/2017/05/qualtricstipsandtricks4-1.pdf

Turpin, A., Scholer, F., Järvelin, K., Wu, M., & Culpepper, J. S. (2009). Including Summaries in System Evaluation. In *SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, July 19-23, 2009* (pp. 508–515). ACM. https://doi.org/10.1145/1571941.1572029

U.S. Census Bureau. (2017). Percentage distribution of population in the United States in 2015 and 2060, by race and Hispanic origin. (Current Population Survey 2014 National Population Projections). In Statista - The Statistics Portal. Retrieved from http://www.statista.com/statistics/270272/percentage-of-us-population-by-ethnicities/

U.S. Food and Drug Administration. (2018, January 8). Products Claiming to "Cure" Cancer Are a Cruel Deception [WebContent]. Retrieved from https://www.fda.gov/ForConsumers/ConsumerUpdates/ucm048383.htm

U.S. Office of Management and Budget [OMB]. (1997). *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity* (October 30, 1997 Federal Register Notice). Executive Office of the U.S. President. Retrieved from https://obamawhitehouse.archives.gov/node/15626

van Deursen, A. J. A. M. (2012). Internet skill-related problems in accessing online health information. *International Journal of Medical Informatics*, *81*(1), 61–72. https://doi.org/10.1016/j.ijmedinf.2011.10.005

Vannette, D. (2018, June 19). How to Label Response Scale Points in Your Survey. Retrieved from https://www.qualtrics.com/blog/how-to-label-response-scale-points-in-your-survey-to-avoid-misdirecting-respondents/

Vannette, D. (2018, June 28). Three Tips for Effectively Designing Rating Scales. Retrieved from https://www.qualtrics.com/blog/three-tips-for-effectively-using-scale-point-questions/

Venetis, P., Koutrika, G., & Garcia-Molina, H. (2011). On the Selection of Tags for Tag Clouds. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11), Hong Kong, China, February 09-12, 2011* (pp. 835–844). ACM. https://doi.org/10.1145/1935826.1935855

Walther, J. B., Wang, Z., & Loh, T. (2004). The Effect of Top-Level Domains and Advertisements on Health Web Site Credibility. *Journal of Medical Internet Research*, *6*(3), e24. https://doi.org/10.2196/jmir.6.3.e24

Wang, Z., Walther, J. B., Pingree, S., & Hawkins, R. P. (2008). Health Information, Credibility, Homophily, and Influence via the Internet: Web Sites Versus Discussion Groups. *Health Communication*, *23*(4), 358–368. https://doi.org/10.1080/10410230802229738

White, R. W., & Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09), Hong Kong, China, November 2–6, 2009* (p. 87). ACM Press. https://doi.org/10.1145/1645953.1645967

White, R. W., & Hassan, A. (2014). Content Bias in Online Health Search. *ACM Transactions on the Web (TWEB)*, *8*(4), 25:1–25:33. https://doi.org/10.1145/2663355

Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246–258. https://doi.org/10.1002/asi.10367

Wildemuth, B. M. (Ed.). (2016). *Applications of Social Research Methods to Questions in Information and Library Science* (2nd Edition). Santa Barbara, California: ABC-CLIO.

Wilson, M., Hurlock, J., & Wilson, M. (2012). Keyword Clouds: Having Very Little Effect on Sensemaking in Web Search Engines. In *CHI EA '12: CHI'12 ACM SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts, Austin, Texas, USA, May 5–10, 2012* (pp. 2069–2074). ACM. https://doi.org/10.1145/2212776.2223754

Wirth, W., Sommer, K., von Pape, T., & Karnowski, V. (2016). Success in online searches: Differences between evaluation and finding tasks: Success in Online Searches: Differences Between Evaluation and Finding Tasks. *Journal of the Association for Information Science and Technology*, *67*(12), 2897–2908. https://doi.org/10.1002/asi.23389

Woodruff, A., Faulring, A., Rosenholtz, R., Morrsion, J., & Pirolli, P. (2001). Using thumbnails to search the Web. In *CHI '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA, USA, March 31-April 4, 2001* (pp. 198–205). ACM Press. https://doi.org/10.1145/365024.365098

Woodruff, A., Rosenholtz, R., Morrison, J. B., Faulring, A., & Pirolli, P. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. *Journal of the American Society for Information Science and Technology*, *53*(2), 172–185. https://doi.org/10.1002/asi.10029

www.yippy.com. (n.d.). www.yippy.com [www.yippy.com]. Retrieved from www.yippy.com

Xie, I., & Benoit III, E. (2013). Search result list evaluation versus document evaluation: similarities and differences. *Journal of Documentation*, *69*(1), 49–80. https://doi.org/10.1108/00220411311295324

Xu, Y. (Calvin), & Chen, Z. (2006). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, *57*(7), 961–973. https://doi.org/10.1002/asi.20361

Yamamoto, Y., & Tanaka, K. (2011). Enhancing credibility judgment of web search results. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, May 7–12, 2011* (pp. 1235–1244). ACM Press. https://doi.org/10.1145/1978942.1979126

You, S., DesArmo, J., Mu, X., Lee, S., & Neal, J. C. (2014). Visualized Related Topics (VRT) system for health information retrieval. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '14), London, United Kingdom, September 08-12, 2014* (pp. 429–430). IEEE. https://doi.org/10.1109/JCDL.2014.6970209

Yue, Y., Patel, R., & Roehrig, H. (2010). Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, April 26–30, 2010* (pp. 1011–1018). ACM. https://doi.org/10.1145/1772690.1772793

Zamir, O., & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. *Computer Networks*, *31*(11–16), 1361–1374. https://doi.org/10.1016/S1389-1286(99)00054-7

Zhang, J., & Marchionini, G. (2005). Evaluation and Evolution of a Browse and Search Interface: Relation Browser++. In *Proceedings of the 2005 National Conference on Digital Government Research* (pp. 179–188). Atlanta, Georgia, USA: Digital Government Society of North America.

Zhang, Y., Sun, Y., & Xie, B. (2015). Quality of health information for consumers on the web: A systematic review of indicators, criteria, tools, and evaluation results: Quality of Health Information for Consumers on the Web. *Journal of the Association for Information Science and Technology*, *66*(10), 2071–2084. https://doi.org/10.1002/asi.23311

Zhitomirsky-Geffet, M., Bar-Ilan, J., & Levene, M. (2016). Testing the stability of "wisdom of crowds" judgments of search results over time and their similarity with the search engine rankings. *Aslib Journal of Information Management*, *68*(4), 407–427. https://doi.org/10.1108/AJIM-10-2015-0165

Zhou, K., Demeester, T., Nguyen, D., Hiemstra, D., & Trieschnigg, D. (2014). Aligning Vertical Collection Relevance with User Intent. In J. Li & X. S. Wang (Eds.), *CIKM'14: proceedings of the 23rd ACM International Conference on Information and Knowledge Management: November 3-7, 2014, Shanghai, China* (pp. 1915–1918). ACM. https://doi.org/10.1145/2661829.2661941

Appendix A

Tables

Table A1

Results from http://www.raosoft.com/samplesize.html *Calculator*

Sample Size for Confidence Level	Margin of Error = 10%		Margin of Error = 5%	
Confidence Level	90%	95%	90%	95%
Minimum Sample Size	68	96	267	377

Note. Response Distribution set to 50% for most conservative (largest) estimate

Table A2

Counts of People Who Clicked on the Qualtrics Survey Link

Status	**Number Completed**	**Number Not Completed**	**Total**
Completed study			199
Screened in, consented and completed study	199		
Did not complete study			80
Self-screened out (Not a health searcher)		1	
Did not consent		9	
Consented but did not answer any questions		7	
Started survey but did not finish		63	
Total	**199**	**80**	**279**

Table A3

Race and Ethnicity of Survey Participants

Race/ethnicity	**Count**	**Percentage**
One race/ethnicity selected		
Asian	6	3.02%
Black or African American	18	9.05%
Caucasian or White	152	76.38%
Hispanic or Latino/Latina	10	5.03%
Native Hawaiian or Pacific Islander	0	0.00%
More than one race/ethnicity selected		
American Indian or Alaska Native and Caucasian or White	2	1.01%
American Indian or Alaska Native and Hispanic or Latino/Latina	1	0.50%
Asian and Caucasian or White	1	0.50%
Black or African American and Caucasian or White	1	0.50%
Black or African American and Hispanic or Latino/Latina	2	1.01%
Caucasian or White and Hispanic or Latino/Latina	3	1.51%
Other	2	1.01%
Did not answer question	1	0.50%
Total	**199**	**100.00%**

Table A4

University Affiliation of Survey Participants

University affiliation	**Count**	**Percentage**
Only one affiliation		
Faculty	23	11.56%
Graduate, Postdoc or Professional Program Student	26	13.07%
Hospital	4	2.01%
Retiree	1	0.50%
Staff	109	54.77%
Undergraduate Student	24	12.06%
More than one affiliation		
Faculty and Hospital	1	0.50%
Graduate, Postdoc or Professional Program Student and Staff	2	1.01%
Faculty and Staff	1	0.50%
Staff and Hospital	2	1.01%
Staff and Other	1	0.50%
Undergraduate Student, Graduate, Postdoc or Professional Program Student and Staff	1	0.50%
Undergraduate Student and Hospital	2	1.01%
Visiting Student or Other Type of Student and Staff	1	0.50%
Did not answer question	1	0.50%
Total	**199**	**100.00%**

Table A5

Summary Statistics for the Six Ratings Variables in Part 1 as Ordinal Variables

Variable	Mean	Std Dev	Variance	Min	Max	N	N Miss	Corrected SS
Relevant	2.79	1.09	1.20	1	5	1592	0	1902.18
Credible	2.53	1.12	1.25	1	5	1591	1	1991.88
Quickly Find	4.38	1.73	2.98	1	7	1592	0	4748.73
Refine	4.58	1.68	2.82	1	7	1591	1	4482.16
Visual	2.82	1.11	1.23	1	5	1592	0	1964.75
Opinion	4.05	1.74	3.02	1	7	1592	0	4811.35

Note: 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings. The variable credible is missing one rating and the variable refine is missing one rating because one participant missed or skipped a rating for each of these variables for one health topic.

Table A6

Recoded Values for the Collapsed Ratings Scales for Part 1

Variable	Original Value for the Level	Recoded Value
Relevant	Not helpful at all	Negative
Relevant	Slightly helpful	Neutral
Relevant	Moderately helpful	Positive
Relevant	Very helpful	Positive
Relevant	Extremely helpful	Positive
Credible	Not helpful at all	Negative
Credible	Slightly helpful	Neutral
Credible	Moderately helpful	Positive
Credible	Very helpful	Positive
Credible	Extremely helpful	Positive
Quickly_Find	Extremely useless	Negative
Quickly_Find	Moderately useless	Negative
Quickly_Find	Slightly useless	Neutral
Quickly_Find	Neither useful nor useless	Neutral
Quickly_Find	Slightly useful	Neutral
Quickly_Find	Moderately useful	Positive
Quickly_Find	Extremely useful	Positive
Refine	Extremely difficult	Negative
Refine	Moderately difficult	Negative
Refine	Slightly difficult	Neutral
Refine	Neither easy nor difficult	Neutral
Refine	Slightly easy	Neutral
Refine	Moderately easy	Positive
Refine	Extremely easy	Positive
Visual	Terrible	Negative
Visual	Poor	Negative
Visual	Average	Neutral
Visual	Good	Positive
Visual	Excellent	Positive
Opinion	Extremely negative	Negative
Opinion	Moderately negative	Negative
Opinion	Slightly negative	Neutral
Opinion	Neither positive nor negative	Neutral
Opinion	Slightly positive	Neutral
Opinion	Moderately positive	Positive
Opinion	Extremely positive	Positive

Table A7

Ratings by Variable by Screen (for Collapsed Response Levels)

Variable	Collapsed Values for Display Screen Ratings	Screen 1 (control)	Screen 2	Screen 3	Screen 4	Grand Total
Relevant	Negative	18	13	120	59	210
Relevant	Neutral	90	77	123	144	434
Relevant	Positive	290	308	155	195	948
Relevant	Grand Total	398	398	398	398	1592
Credible	Negative	50	41	151	101	343
Credible	Neutral	92	96	132	122	442
Credible	Positive	256	261	115	174	806
Credible	Grand Total	398	398	398	397	1591
Quickly Find	Negative	28	27	174	88	317
Quickly Find	Neutral	203	165	177	233	778
Quickly Find	Positive	167	206	47	77	497
Quickly Find	Grand Total	398	398	398	398	1592
Refine	Negative	16	23	120	84	243
Refine	Neutral	198	151	198	220	767
Refine	Positive	183	224	80	94	581
Refine	Grand Total	397	398	398	398	1591
Visual	Negative	30	71	296	230	627
Visual	Neutral	200	134	55	107	496
Visual	Positive	168	193	47	61	469
Visual	Grand Total	398	398	398	398	1592
Opinion	Negative	18	22	204	114	358
Opinion	Neutral	223	206	163	235	827
Opinion	Positive	157	170	31	49	407
Opinion	Grand Total	398	398	398	398	1592

Note. 199 participants x 2 topics x 4 screens rated per topic = 1592 total ratings.
The variable credible is missing one rating and the variable refine is missing one rating because one participant missed or skipped a rating for each of these variables for one topic.

Table A8

Pearson's Chi-Square of all Six Measures for Display Screen Ratings

Variable	**Pearson's Chi-Square**	**p value**	**Sample size**	**Significance Level**
Relevant	270.7484	<.0001	1592	0.05
Credible	199.9493	<.0001	1591 (missing = 1)	0.05
Quickly find	365.9399	<.0001	1592	0.05
Refine	275.4335	<.0001	1591 (missing = 1)	0.05
Visual	581.1004	<.0001	1592	0.05
Opinion	521.7513	<.0001	1592	0.05

Note: 199 participants x 2 topics x 4 screens rated per health topic = 1592 sample size. The variable credible is missing one rating and the variable refine is missing one rating because one participant missed or skipped a rating for each of these variables for one topic.

Table A9

Part 2 Results by Percentage of Participants (Overall Prefer, Browser Extension or Customization, and Dislike)

Topics Assigned	Screen 1 (control)	Screen 2	Screen 3	Screen 4	Count
Overall, which type of display did you prefer the most…?					
Topics 1 and 2	44.12%	39.71%	7.35%	8.82%	n=68
Topics 2 and 3	37.68%	55.07%	5.80%	1.45%	n=69
Topics 3 and 1	43.55%	50.00%	4.84%	1.61%	n=62
All Topics Total	41.71%	48.24%	6.03%	4.02%	n=199
If a browser extension or other customization was available…?					
Topics 1 and 2	27.94%	57.35%	7.35%	7.35%	n=68
Topics 2 and 3	27.54%	63.77%	5.80%	2.90%	n=69
Topics 3 and 1	32.26%	59.68%	3.23%	4.84%	n=62
All Topics Total	29.15%	60.30%	5.53%	5.03%	n=199
Which type of display did you dislike the most...?					
Topics 1 and 2	2.94%	5.88%	58.82%	32.35%	n=68
Topics 2 and 3	7.25%	4.35%	72.46%	15.94%	n=69
Topics 3 and 1	3.23%	6.45%	70.97%	19.35%	n=62
All Topics Total	4.52%	5.53%	67.34%	22.61%	n=199

Table A10

Co-occurrence of Preferred screen and Desired Browser Extension

Search Results Display Screens	Overall, which type of display did you prefer the most for viewing search engine results from health information searches?	What % selected the same screen for a browser extension or customization as well?
Screen 1 (control)	83	63.86%
Screen 2	96	94.79%
Screen 3	12	75.00%
Screen 4	8	87.50%

Note. n=199 participants

Table A11

Co-occurrence of Desired Browser Extension and Preferred Screen

Search Results Display Screens	If a browser extension or other customization was available to ensure that your search engine results from health information searches would appear like any of these displays, which option would you pick?	What % selected the same screen for their preferred screen type as well?
Screen 1 (control)	58	91.38%
Screen 2	120	75.83%
Screen 3	11	81.82%
Screen 4	10	70.00%

Note. n=199 participants

Table A12

Binomial Representation of the Prefer, Browser and Dislike Variables

Screen	Value	Overall Prefer	Browser or Customization	Dislike
Screen 1 (control)	Not Selected (0)	116	141	190
Screen 1 (control)	Selected (1)	83	58	9
Screen 1 (control)	**Total**	**199**	**199**	**199**
Screen 2	Not Selected (0)	103	79	188
Screen 2	Selected (1)	96	120	11
Screen 2	**Total**	**199**	**199**	**199**
Screen 3	Not Selected (0)	187	188	65
Screen 3	Selected (1)	12	11	134
Screen 3	**Total**	**199**	**199**	**199**
Screen 4	Not Selected (0)	191	189	154
Screen 4	Selected (1)	8	10	45
Screen 4	**Total**	**199**	**199**	**199**

Note. n=199 participants

Table A13

Cumulative Frequencies for Part 2 (Overall Prefer, Browser Extension or Customization, and Dislike)

Screen	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Overall, which type of display did you prefer the most…?				
Display Screen 1 (control)	83	41.71	83	41.71
Display Screen 2	96	48.24	179	89.95
Display Screen 3	12	6.03	191	95.98
Display Screen 4	8	4.02	199	100
If a browser extension or other customization was available…?				
Display Screen 1 (control)	58	29.15	58	29.15
Display Screen 2	120	60.3	178	89.45
Display Screen 3	11	5.53	189	94.97
Display Screen 4	10	5.03	199	100
Which type of display did you dislike the most...?				
Display Screen 1 (control)	9	4.52	9	4.52
Display Screen 2	11	5.53	20	10.05
Display Screen 3	134	67.34	154	77.39
Display Screen 4	45	22.61	199	100

Note. n=199 participants

Table A14

Chi-Square Test for Equal Proportions for Part 2 (Overall Prefer, Browser Extension or Customization, and Dislike)

Variable	Chi-Square	p value	Sample Size	Significance Level
Prefer	128.8995	<.0001	199	0.05
Browser	162.5075	<.0001	199	0.05
Dislike	206.6884	<.0001	199	0.05

Appendix B

Figures

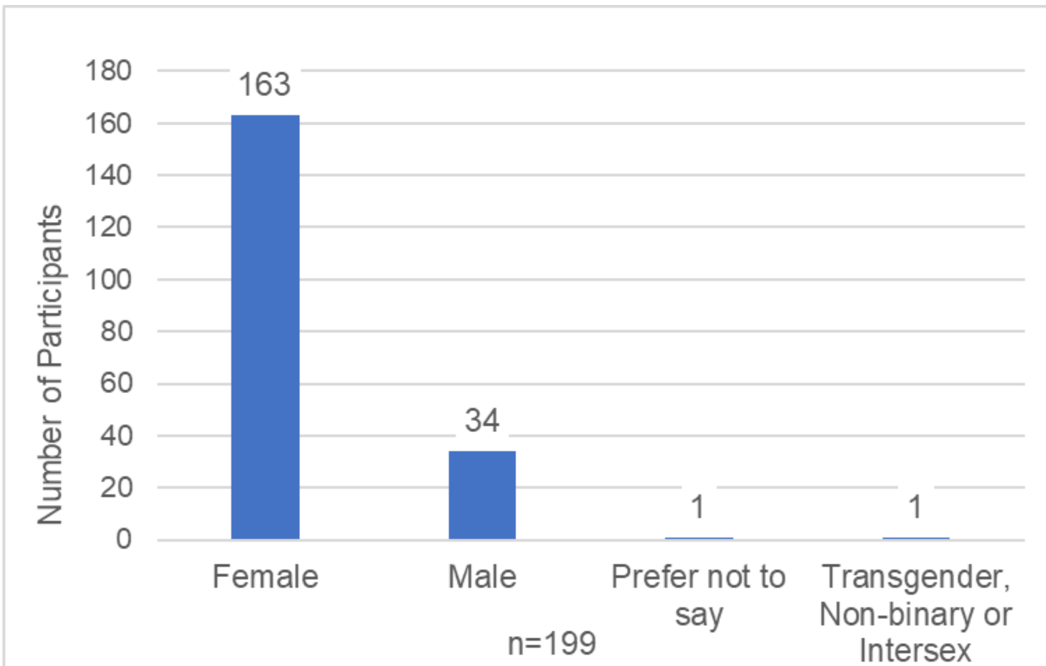


Figure B1. Frequency distribution of gender in survey participants.

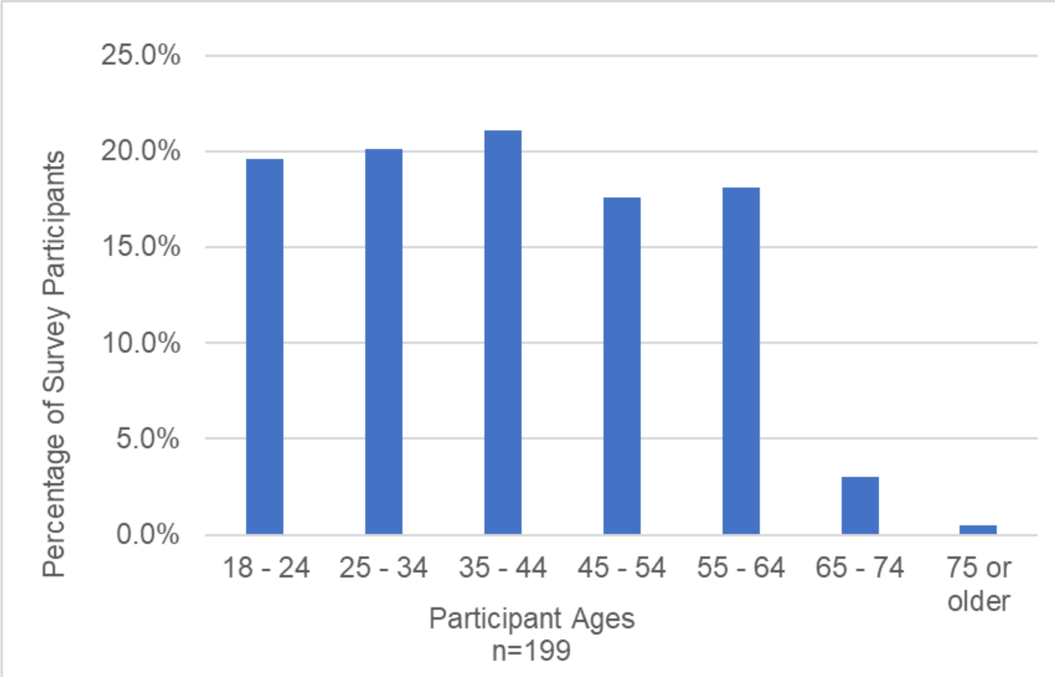


Figure B2. Relative frequency distribution of age in survey participants.

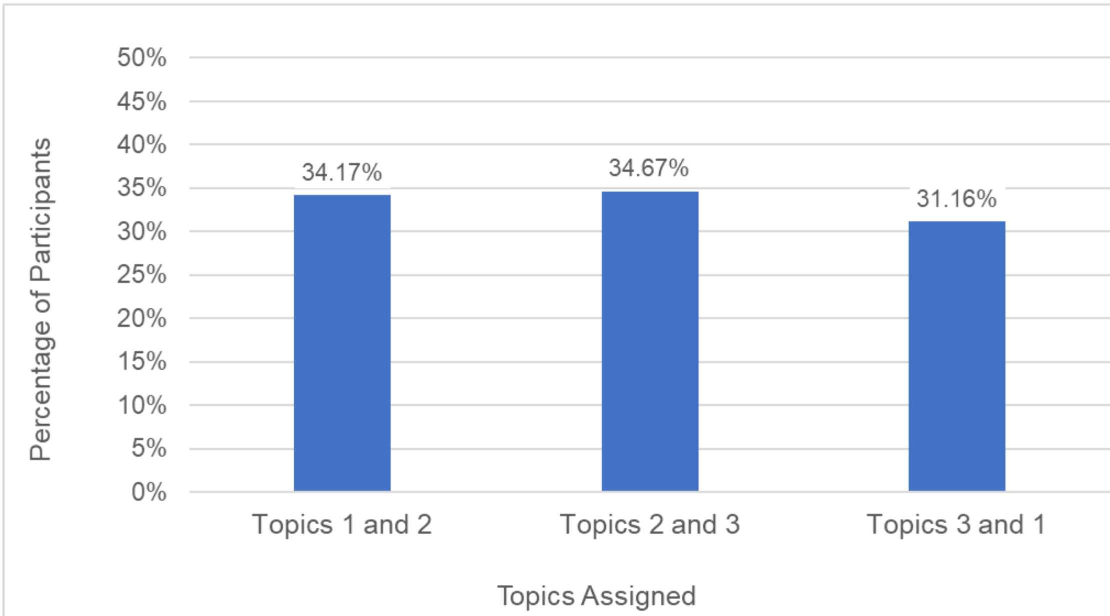


Figure B3. Relative frequency distribution of health topics by percentage of participants. n=199 participants.

Chi-Square Test for Equal Proportions	
Chi-Square	0.4322
DF	2
Pr > ChiSq	0.8057

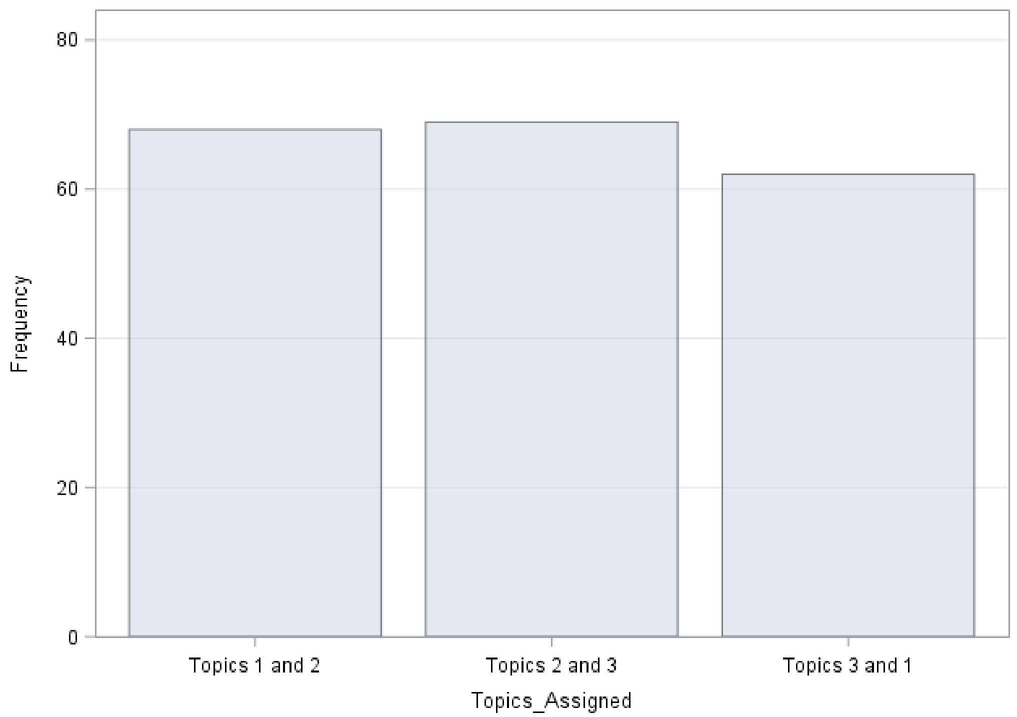


Figure B4. Chi-square test for equal proportions and frequency distribution of the health topics assigned.

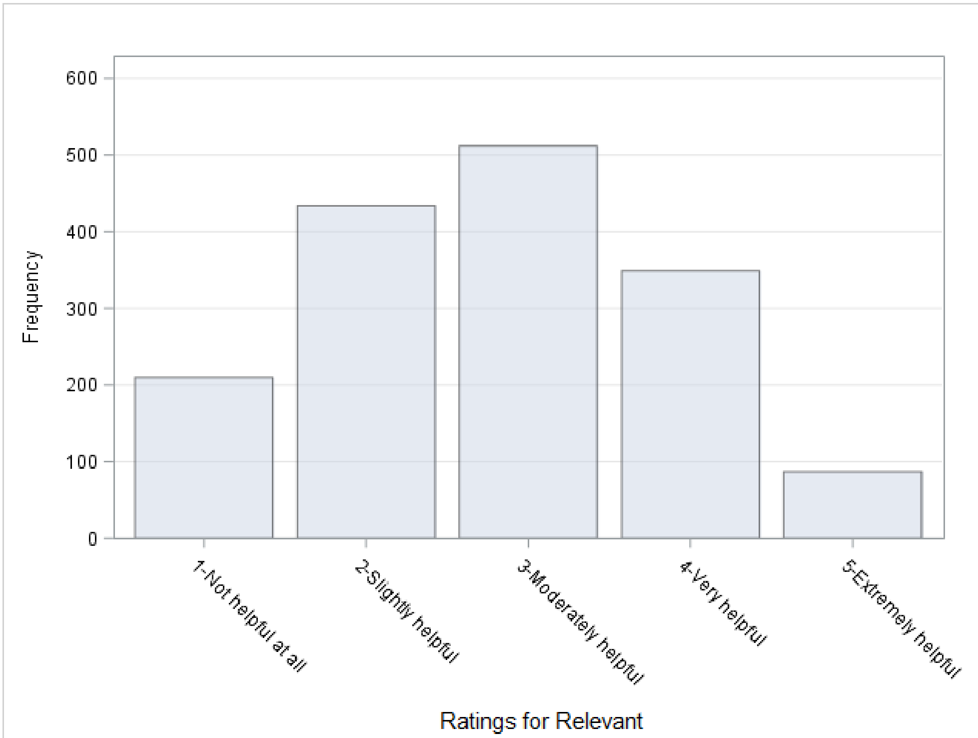


Figure B5. Frequency distribution of relevant as a categorical variable. 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings.

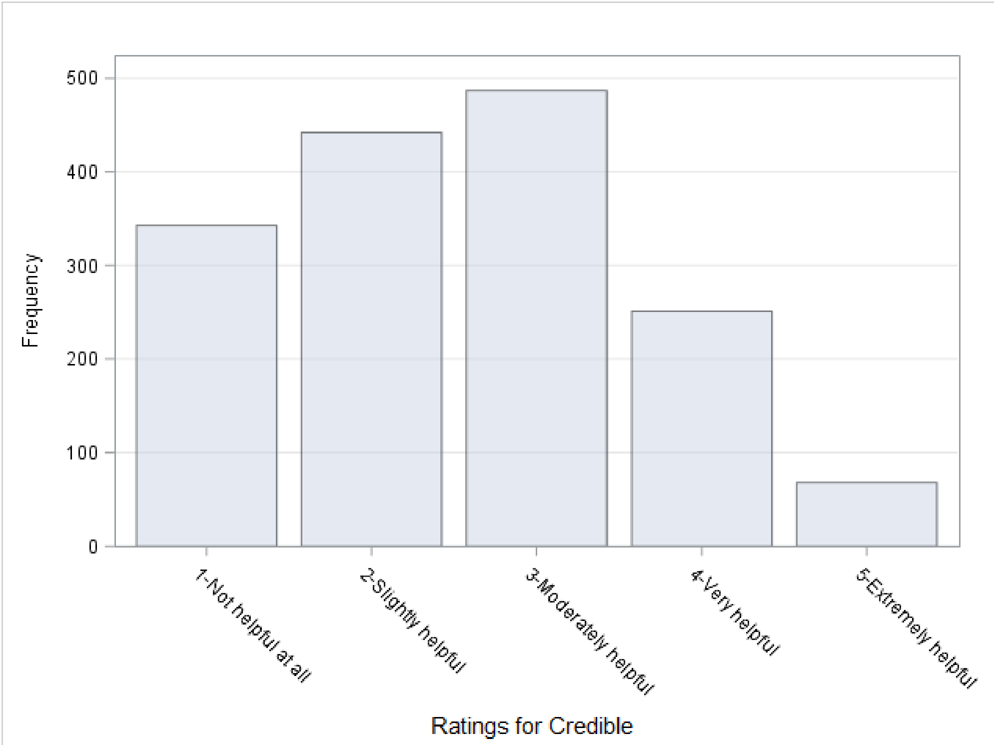


Figure B6. Frequency distribution of credible as a categorical variable. 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings. The variable credible is missing one rating because one participant missed or skipped a rating, so this frequency distribution only has 1591 ratings.

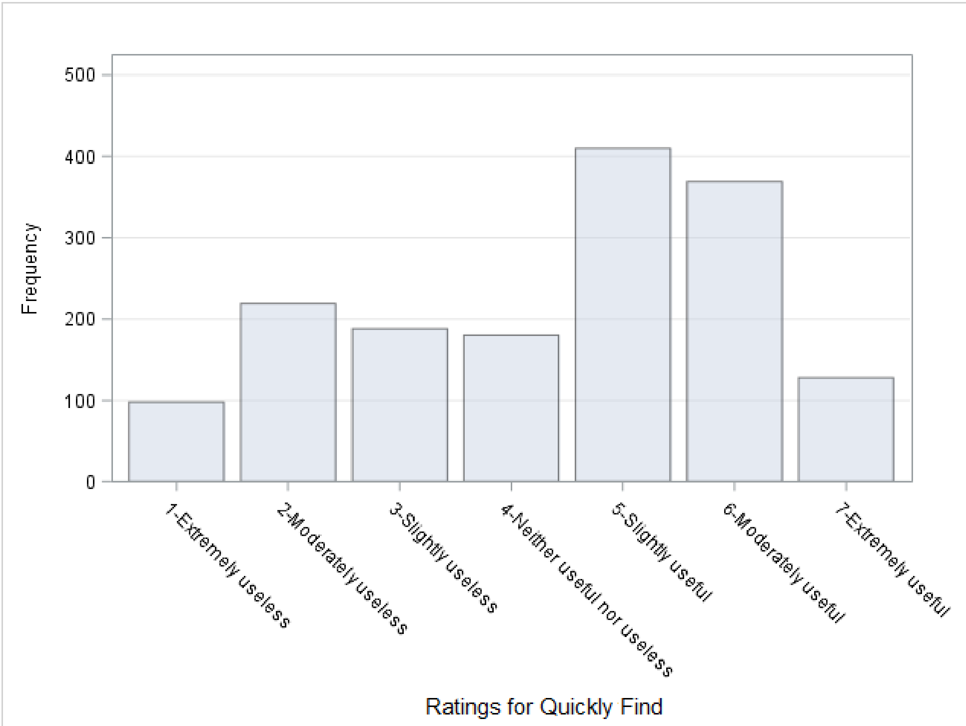


Figure B7. Frequency distribution of quickly find as a categorical variable. 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings.

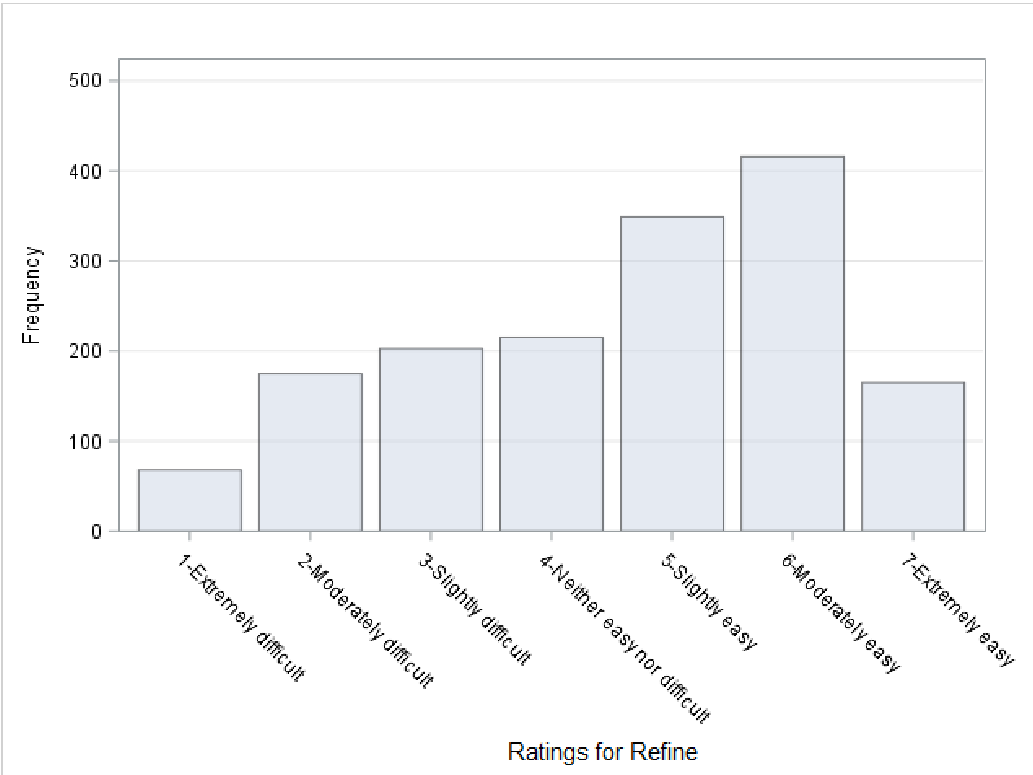


Figure B8. Frequency distribution of refine as a categorical variable. 199 participants x 2 health topics x 4 screens rated per topic = 1592 ratings. The variable refine is missing one rating because one participant missed or skipped a rating for each of these variables for one health topic, so there are only 1591 ratings in this frequency distribution.

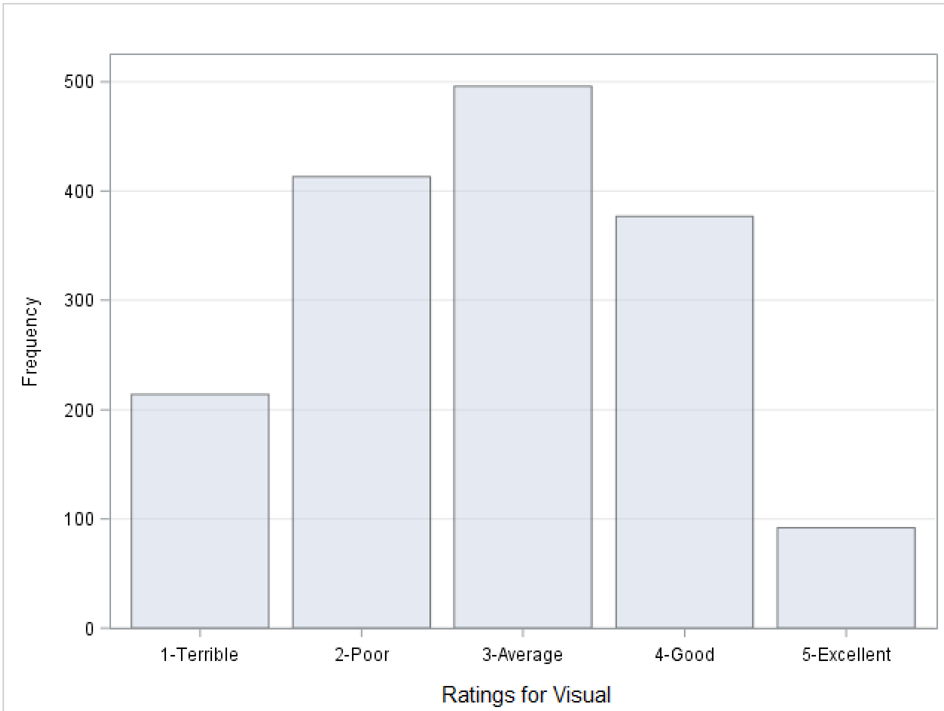


Figure B9. Frequency distribution of visual appeal as a categorical variable .199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings.

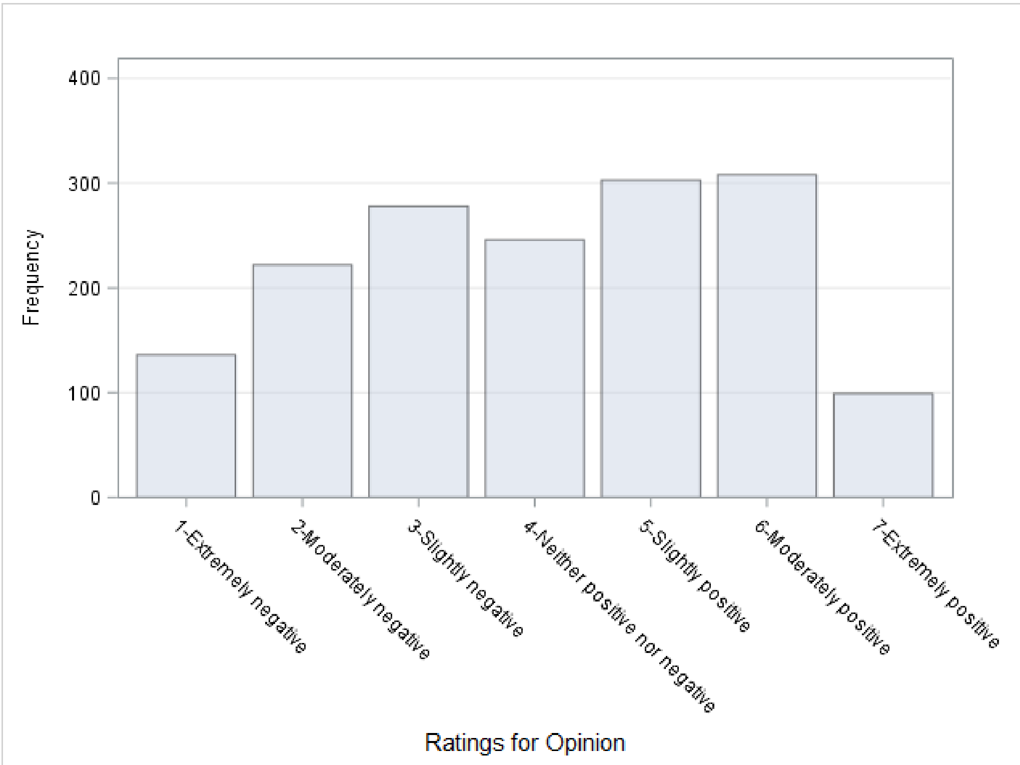


Figure B10. Frequency distribution of opinion as a categorical variable. 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings.

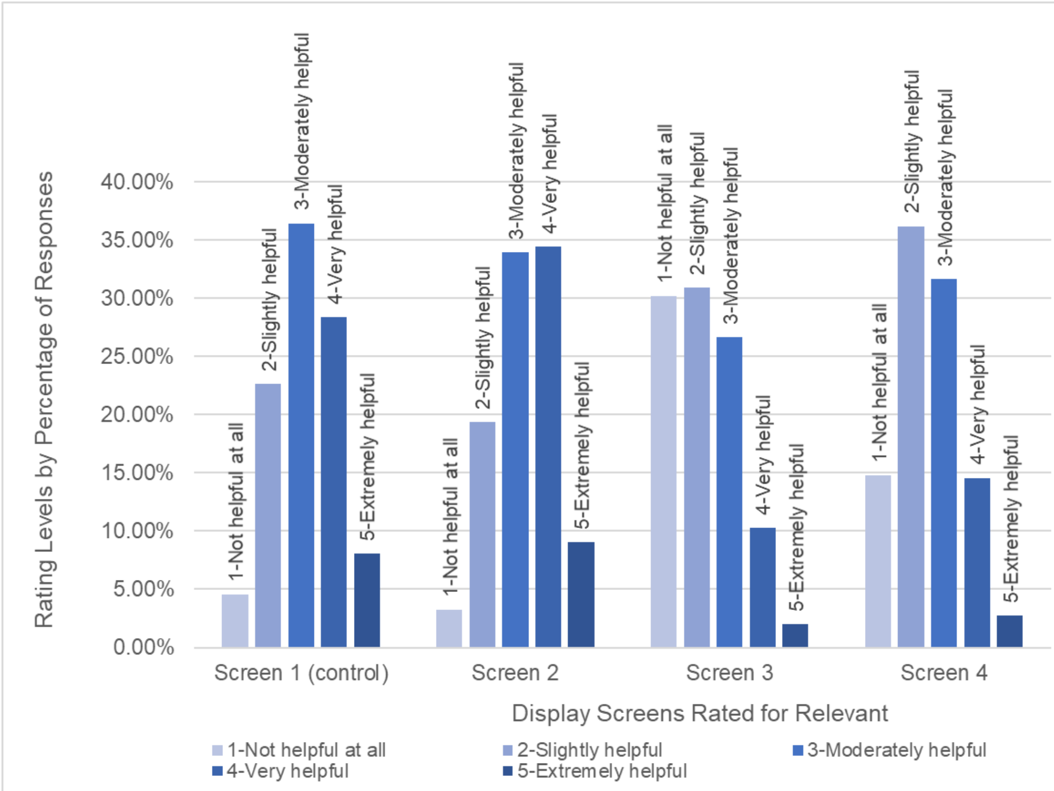


Figure B11. Relative frequency distribution of display screen ratings for variable relevant. n=398 per screen (199 participants rating each screen x 2 health topics).

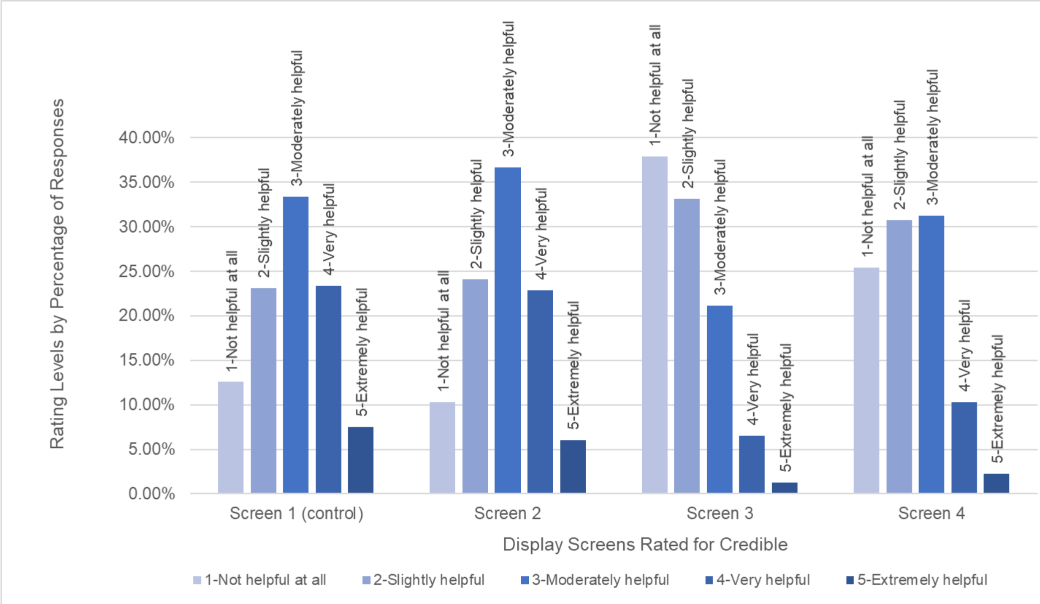


Figure B12. Relative frequency distribution of display screen ratings for variable credible. n=398 per screen (199 participants rating each screen x 2 health topics). [n=397 for Screen 4 because one participant skipped or did not complete one rating].

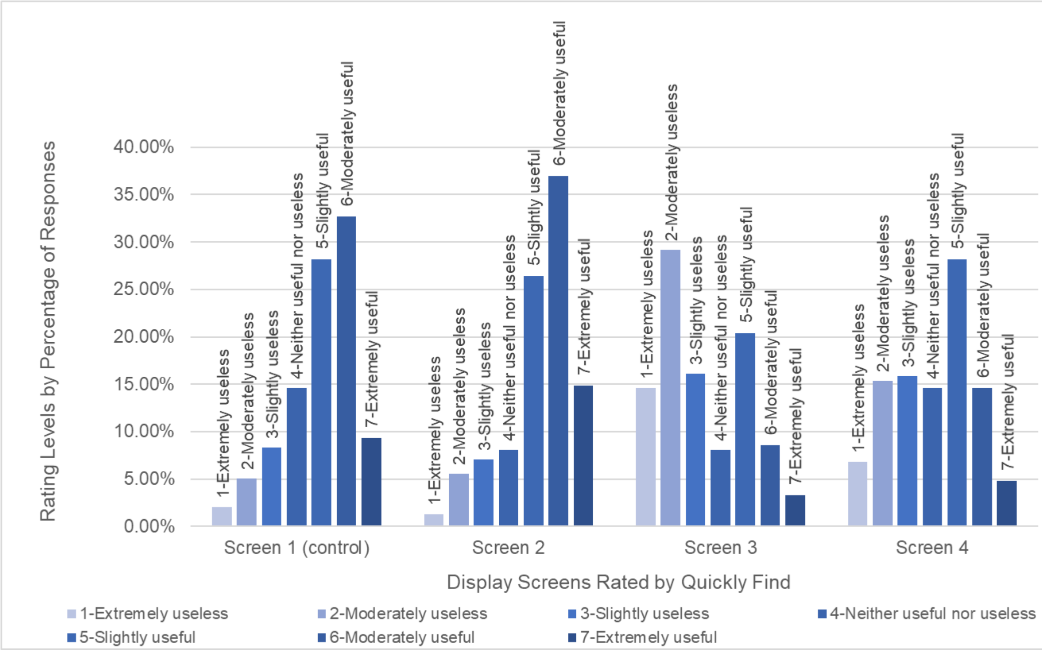


Figure B13. Relative frequency distribution of display screen ratings for variable quickly find. n=398 per screen (199 participants rating each screen x 2 health topics).

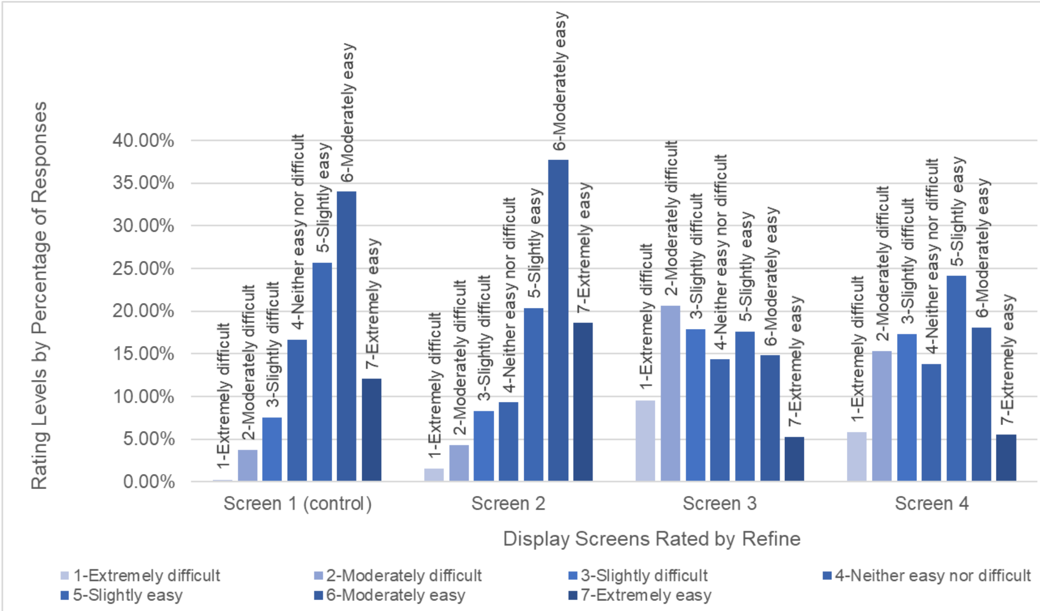


Figure B14. Relative frequency distribution of display screen ratings for variable refine. n=398 per screen (199 participants rating each screen x 2 health topics). [n=397 for Screen 1 (control) because one participant skipped or did not complete one rating].

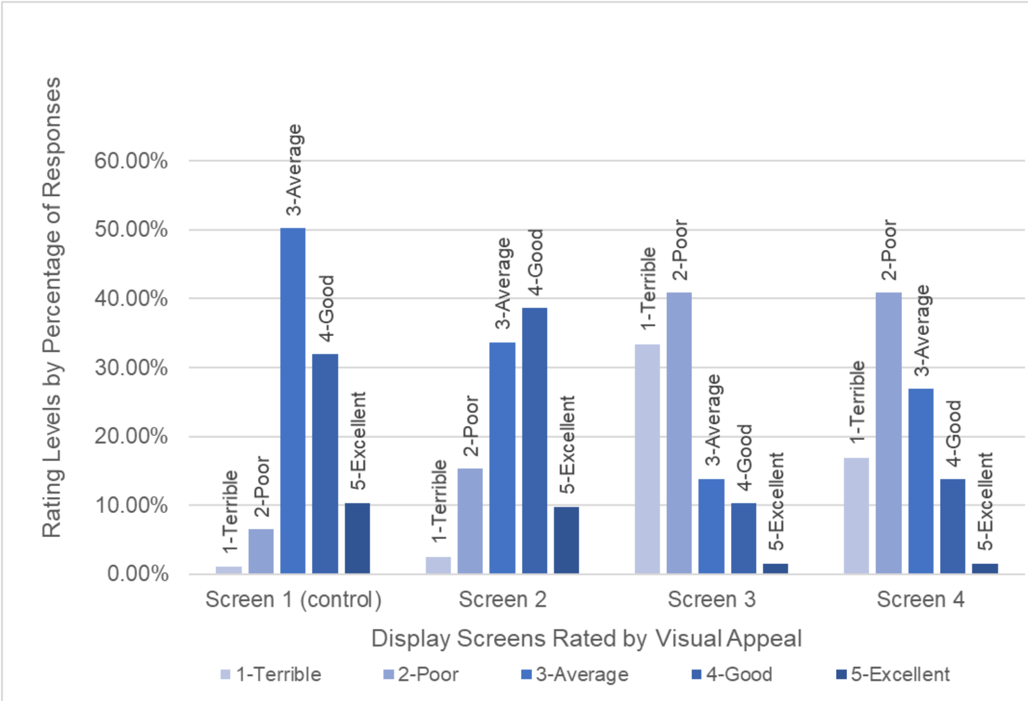


Figure B15. Relative frequency distribution of display screen ratings for variable visual appeal. n=398 per screen (199 participants rating each screen x 2 health topics).

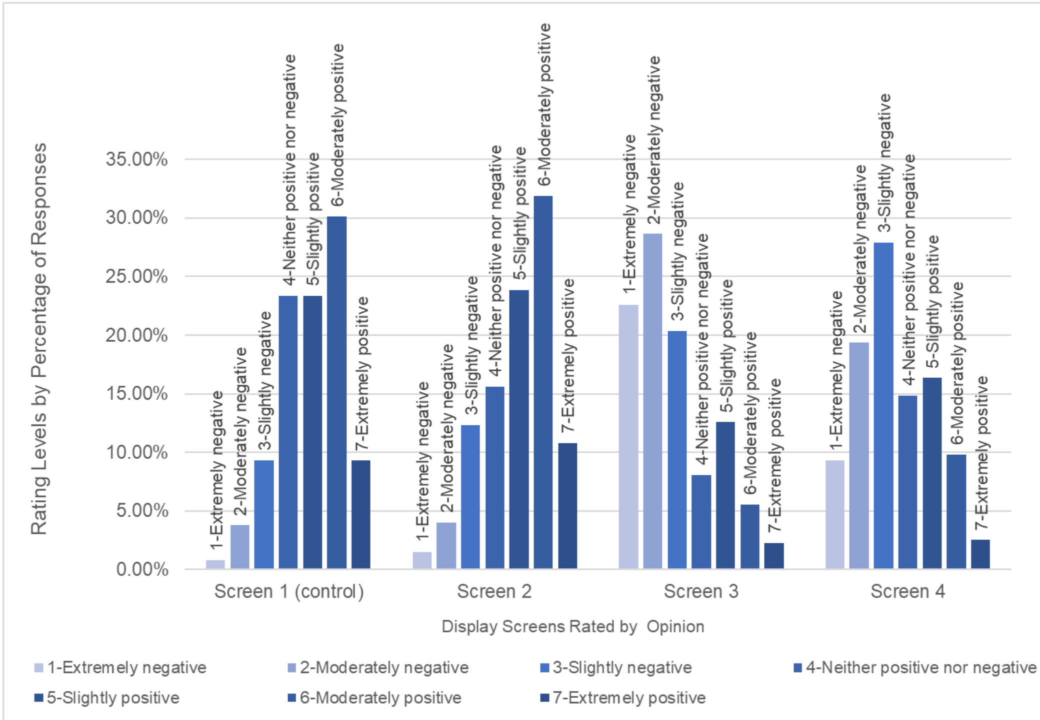


Figure B16. Relative frequency distribution of display screen ratings for variable opinion. n=398 per screen (199 participants rating each screen x 2 health topics).

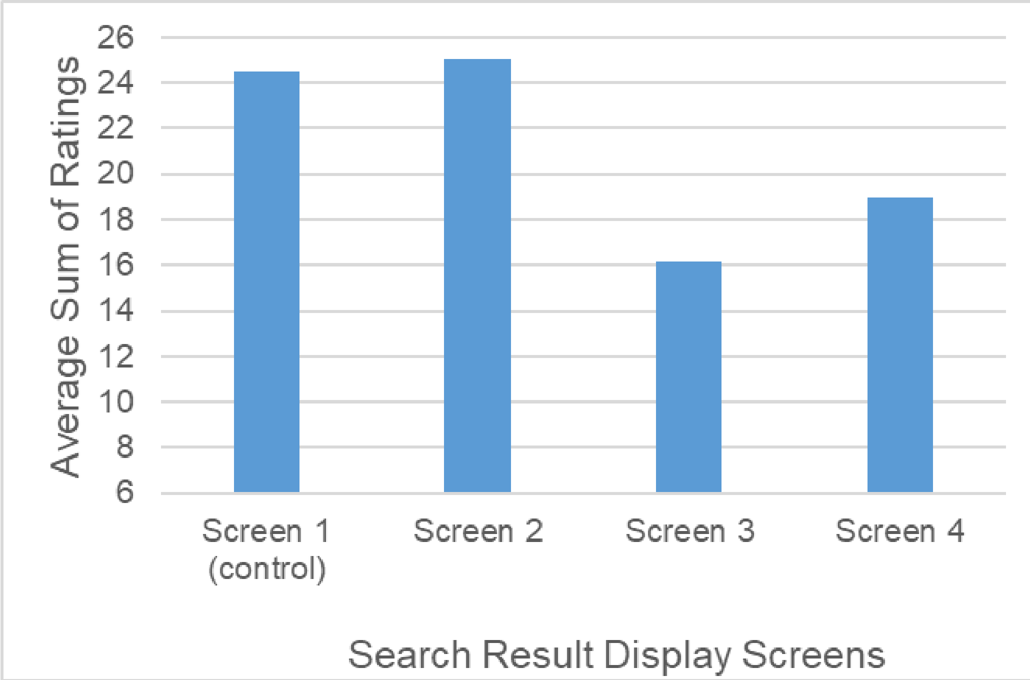


Figure B17. Average Sum of Ratings for all 6 measures for both Health Topics Assigned. Minimum Average is 6 because any screen given all "1" ratings would still have a sum of 6 and average of 6. n=398 ratings per screen (199 participants rating each screen for 2 health topics). [n=397 for Screen 1 because one participant skipped or did not complete one rating]. [n=397 for Screen 4 because one participant skipped or did not complete one rating]. Some rating scales were seven-point and some were five-point.

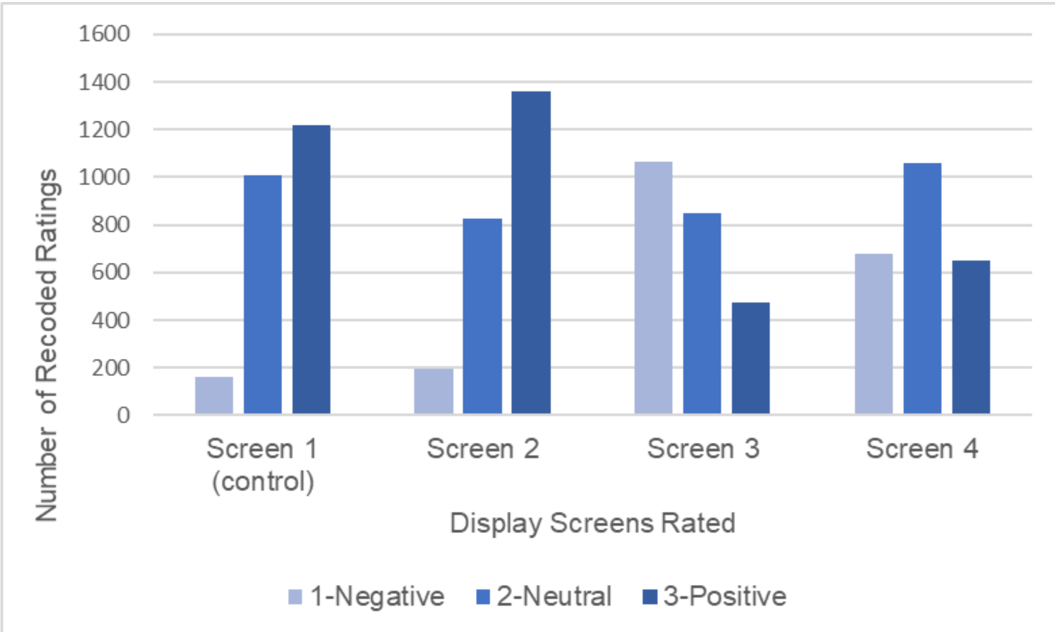


Figure B18. Frequency distribution of ratings collapsed into only negative, neutral and positive response levels. 199 participants x 2 topics x 6 ratings = 2,388 ratings per screen. Screen 1 (control) and Screen 4 have 2,387 ratings each because one participant didn't enter one rating each for those screens.

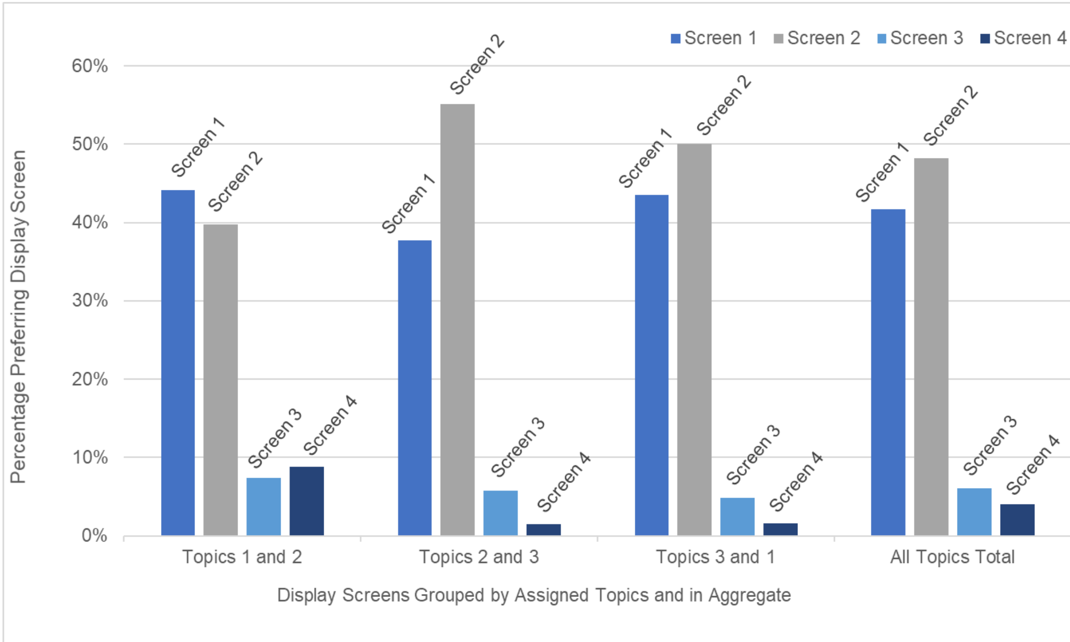


Figure B19. Relative frequency distribution of participant preferred search engine display screens by health topics assigned and by all health topics combined. Topics 1 and 2 Participants n=68. Topics 2 and 3 Participants n=69. Topics 3 and 1 Participants n=62. Total Participants n=199.

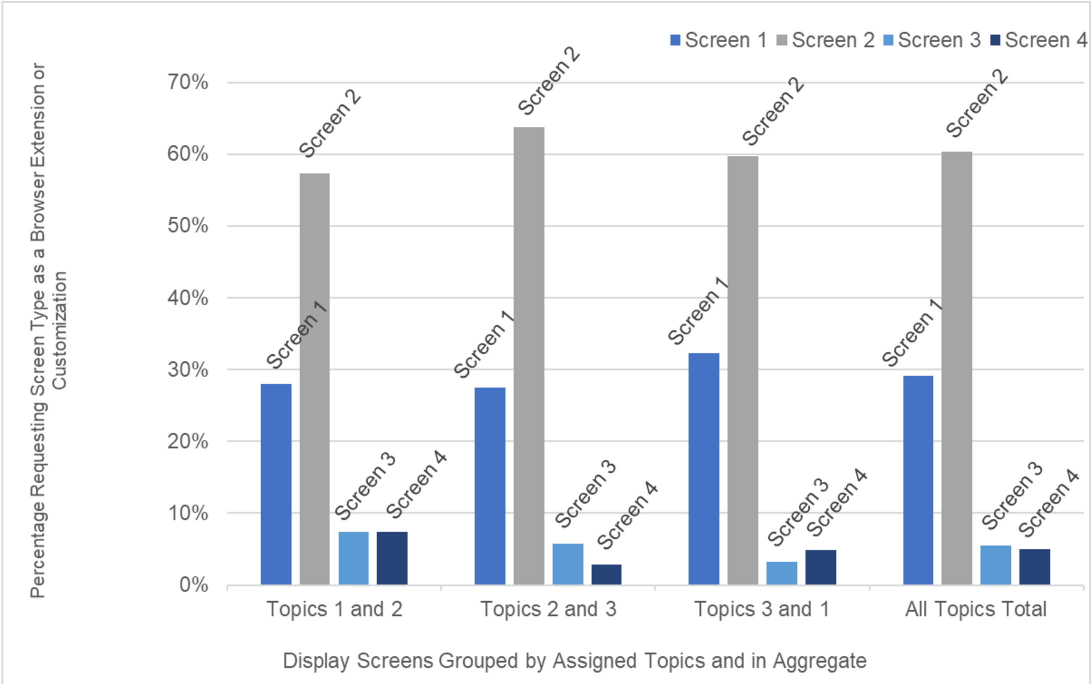


Figure B20. Relative frequency distribution of participant requests for using search engine display screen as a browser extension or as a customization by health topics assigned and by all health topics combined. Topics 1 and 2 Participants n=68. Topics 2 and 3 Participants n=69. Topics 3 and 1 Participants n=62. Total Participants n=199.

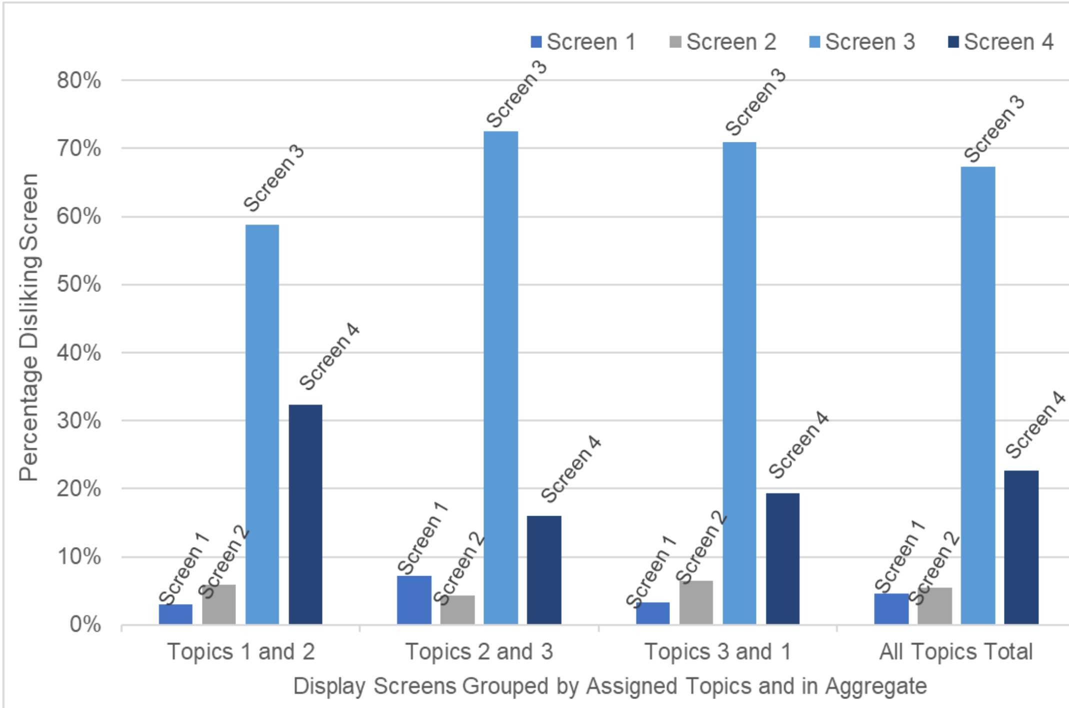


Figure B21. Relative frequency distribution of participant dislikes of search engine results screens by health topics assigned and by all health topics combined. Topics 1 and 2 Participants n=68. Topics 2 and 3 Participants n=69. Topics 3 and 1 Participants n=62. Total Participants n=199.

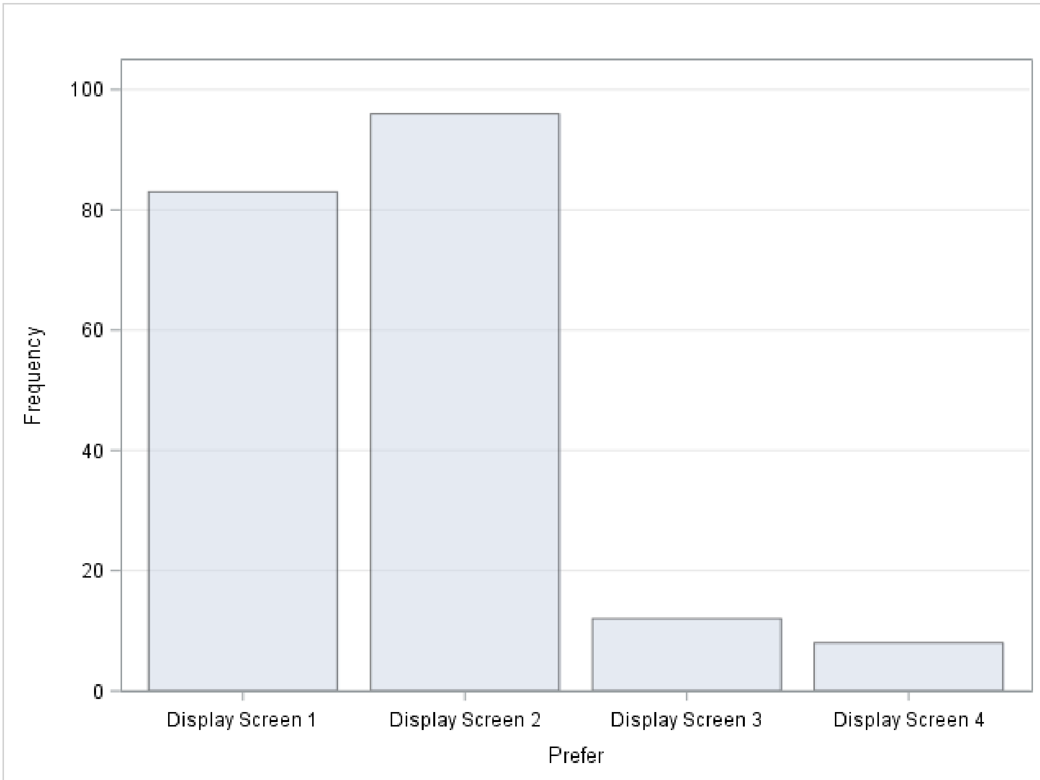


Figure B22. Frequency distribution for prefer variable. n=199 participants.

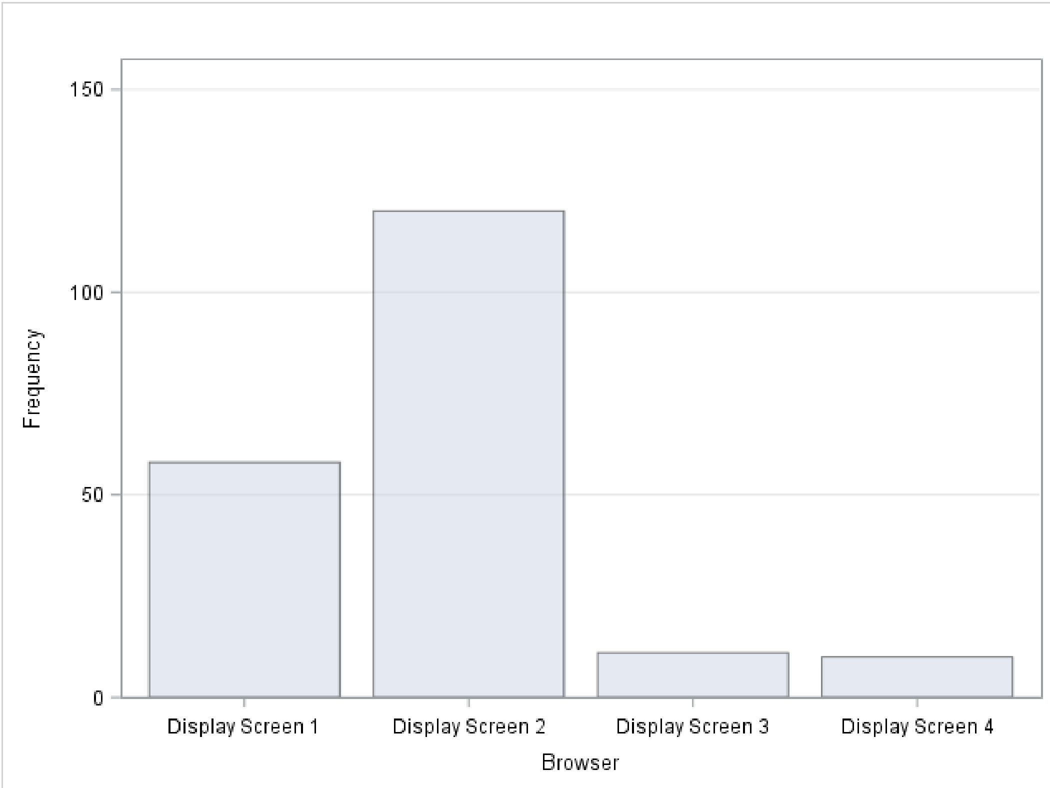


Figure B23. Frequency distribution for browser extension or customization variable. n=199 participants.

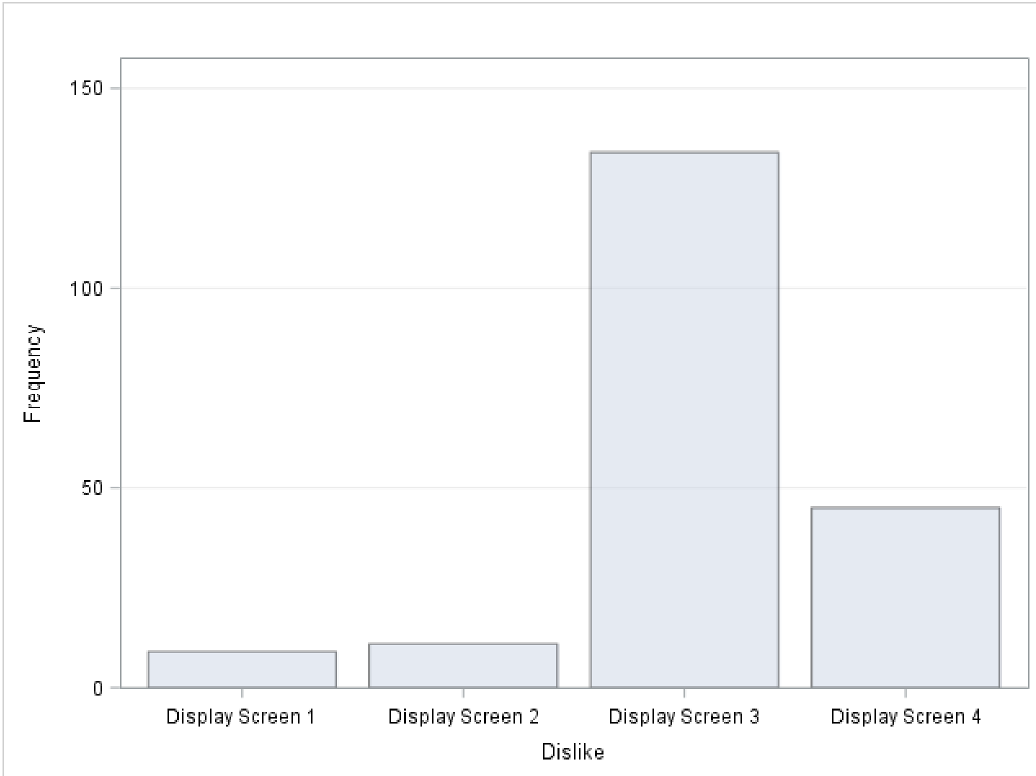


Figure B24. Frequency distribution for dislike variable. n=199 participants.

Appendix C

Recruiting Email Distributed Through Mass Email

RE: [INFORMATIONAL] Survey about health information and search engines.

Gift Card Drawing!

From: Diane Rodden, MSIS Candidate at UNC School of Information and Library Science
IRB #: Non-Biomedical 18-2487 approved on 10/24/2018 by UNC IRB

You are invited to participate in an online research study that only requires answering some simple survey questions. This research is being done for my master's paper.

In order to participate in this study, you must:

- Be 18 years of age or older
- Be able to read, write and understand English
- Have searched for health or medical information online (for yourself or for others)

The purpose of the study is to learn about user preferences for how search engines display search results for health information searches.

It is sometimes difficult to find quality health information on the Internet. By studying different layouts of search engine results for health information, better ways of presenting health information can be developed. Your input is important!

Research is designed to benefit society by contributing to new knowledge about a subject. It is important that research includes a diverse sample of individuals so that the results are representative and beneficial to society. I am hoping to get a wide variety of individuals as participants so that I can gather viewpoints representing people from all parts of society.

This survey will take you about 20 minutes. Please note that this survey will be best displayed on a laptop or desktop computer.

After completing the entire survey, you will have the option to provide your name and email address if you want to enter the drawing for one of the two $50.00 Visa® gift cards.

Your responses will be confidential. No identifying information will be collected unless you want to enter the drawing. If you want to enter the drawing, the only identifying information that will be collected is your name and UNC email address. Your name and UNC email address will be collected in a different Qualtrics data file from your survey responses.

Your participation is voluntary.

Follow this link to the survey.

Or copy and paste the URL below into your internet browser:

https://unc.az1.qualtrics.com/jfe/form/SV_9yTo79JKXo7Dc6p

If you have any questions for the researcher, Diane Rodden, you can contact her at rodden@live.unc.edu. If you have any questions or concerns regarding your rights as a research subject, you may contact the Institutional Review Board via email at IRB_subjects@unc.edu, or at (919) 966-3113 if you would like to contact the IRB anonymously. This survey has been reviewed by the UNC-CH Non-Biomedical Institutional Review Board (IRB) as part of application # 18-2487: "What Presentation of Search Engine Results Do Health Information Searchers Prefer?" approved on 10/24/2018.

Appendix D

Consent Form

Consent Form Included in Qualtrics Main Survey

IRB: Non-Biomedical 18-2487

Principal Investigator: Diane Rodden (rodden@live.unc.edu)

Title of Study: What Presentation of Search Engine Results Do Health Information Searchers Prefer?

IRB: Non-Biomedical 18-2487

Principal Investigator: Diane Rodden (rodden@live.unc.edu), a graduate student at UNC-Chapel Hill School of Information and Library Science

Faculty Advisor: Dr. Bob Losee (losee@unc.edu), Professor, UNC-Chapel Hill School of Information and Library Science

Purpose:

The purpose of the study is to learn about user preferences for how search engines display search results for health information searches.

Study Participant Criteria:

In order to participate in this study you must:

- Be 18 years of age or older
- Be able to read, write and understand English
- Have searched for health or medical information online (for yourself or for others)

Information about Research Studies: You are being asked to take part in a research study. Joining the study is voluntary. You may choose not to participate, or you may withdraw your consent to be in the study, for any reason, without penalty. Research is designed to benefit society by contributing to new knowledge about a subject. It is important that research includes a diverse sample of individuals so that the results are representative and beneficial to society.

What Will Happen During the Study:

This study will be conducted entirely within the Qualtrics survey application. In this study, you will be shown some screen shots of sample search engine results. You will be asked some questions about the various sample search engine results. You will also be asked a few demographic questions. This survey will take you about 20 minutes.

After completing the entire survey, you will have the option to provide your name and email address if you want to enter the drawing for one of the two $50.00 Visa® gift cards.

Please note that this survey will be best displayed on a laptop or desktop computer. Some features may be less compatible for use on a mobile device.

Risks:

There are minimal risks to participants in this survey. You will be asked for your feedback on how the health information search results are displayed. There are no right or wrong answers, so there should be minimal risk of embarrassment or emotional distress to you.

If you decide to participate in this study, you have the right to withdraw and discontinue answering of questions at any time for any reason. You can also skip a specific question if you feel uncomfortable answering it. If you skip a question, you will be asked by the Qualtrics survey to confirm that you do, in fact, prefer to skip the question.

Your responses will be confidential. No identifying information will be collected unless you want to enter the drawing. If you want to enter the drawing, the only identifying information that will be collected is your name and UNC email address. Your name and UNC email address will be collected in a different Qualtrics data file from your survey responses. The names and email addresses of participants will be deleted after the drawing for gift cards has been conducted and the gift cards have been awarded. Of course, there are always unforeseen risks where a third party could obtain access to your name, email address and/or survey responses despite precautions in place to guard your privacy.

Benefits:

There is no direct benefit to you for participating in this study.

Drawing:

After completing the entire survey, you will have the opportunity to participate in a drawing for one of two $50.00 Visa® gift cards.

Institutional Review Board Approval:

All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns about your rights as a research subject you may contact, anonymously if you wish, the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu. If you contact the IRB, please refer to study number 18-2487 (Non-Biomedical).

Consent:

By clicking below to consent, you confirm and acknowledge all of the following:

- you are 18 years of age or older
- you are able to read, write and understand English
- you want to participate in this study
- your participation in the study is voluntary
- you are aware that you may choose to terminate your participation in the study at any time and for any reason

○ I consent, begin the study

○ I do not consent, I do not wish to participate

Appendix E

Qualtrics Survey Questions

Qualtrics Survey Questions

IRB: Non-Biomedical 18-2487

Principal Investigator: Diane Rodden (rodden@live.unc.edu)

Qualtrics Main Survey Screening Questions (Before Informed Consent)

Are you age 18 or older?

- ○ Yes
- ○ No
- ○ I prefer not to answer

Have you ever searched for health or medical information online (either for yourself or for someone else)?

- ○ Yes
- ○ No
- ○ I prefer not to answer

Qualtrics Main Survey Introduction

This survey will consist of 3 sections:

- Section 1 is the longest.

  You will see a health question or concern for which a user might use a search engine to locate information. There will be 4 different types of displays of search engine results, each accompanied by questions about the given display.

  Then, for an additional health information question or concern, you will again see the 4 types of displays of the search engine results and answer some questions about each of the displays.

  There will be a total of 8 pages in Section 1, because there are 4 types of displays for the first health question or concern and then there are the 4 types of displays again for the second health question or concern.

- Section 2 will consist of 3 questions about the search engine results displays. It will be 1 page long.

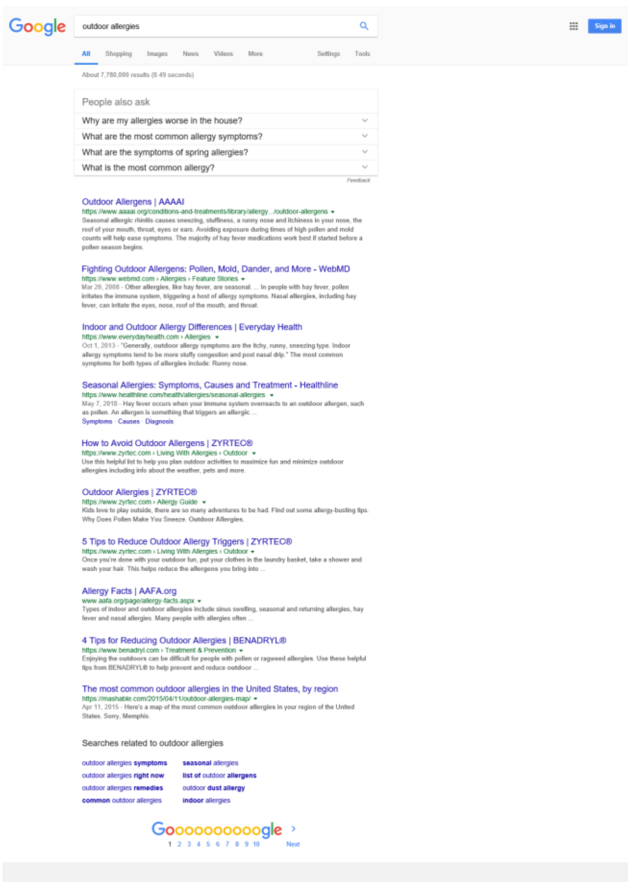
- Section 3 will contain some demographic questions about your background. It will be 1 page long.

Qualtrics Main Survey Questions for Health Topic #1 (Participants will see 2 topics out of 3)

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: What substances are common causes of outdoor allergies?



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

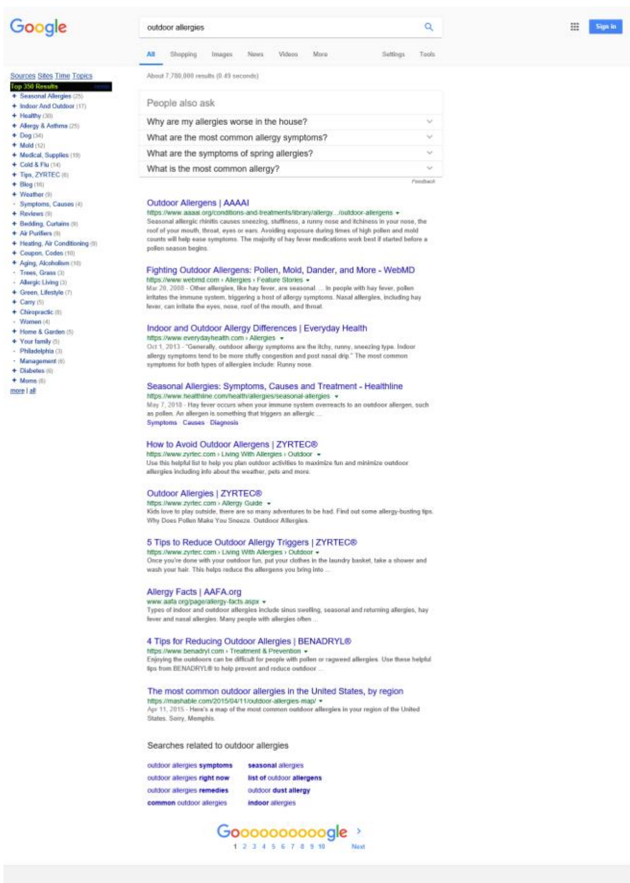
Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **your opinion** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot. **Health question: What substances are common causes of outdoor allergies?**



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by using the folders in the left sidebar.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **your opinion** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: What substances are common causes of outdoor allergies?

Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud below each listing in the search results.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

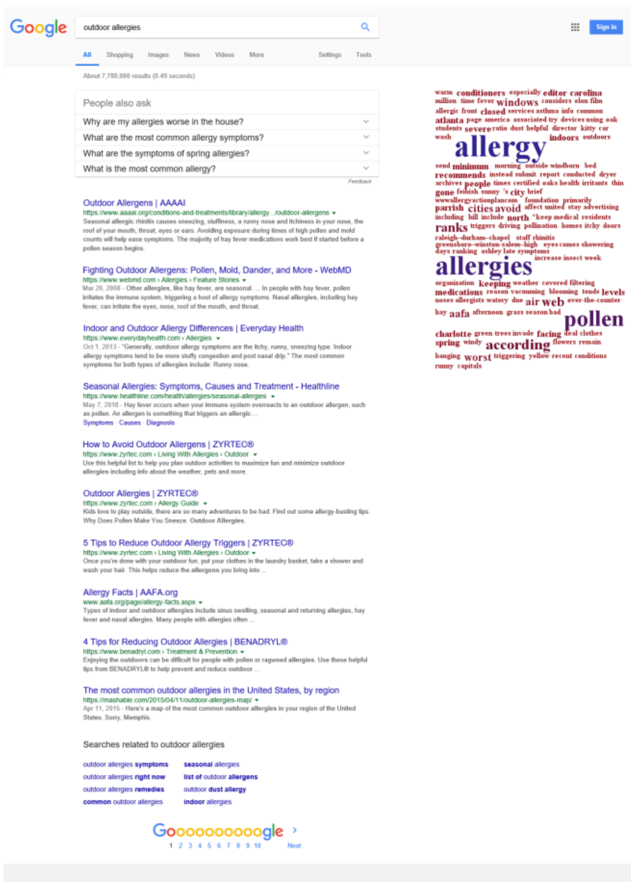
Overall, what is **your opinion** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: What substances are common causes of outdoor allergies?



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud in the right sidebar.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **your opinion** of this search engine results display?

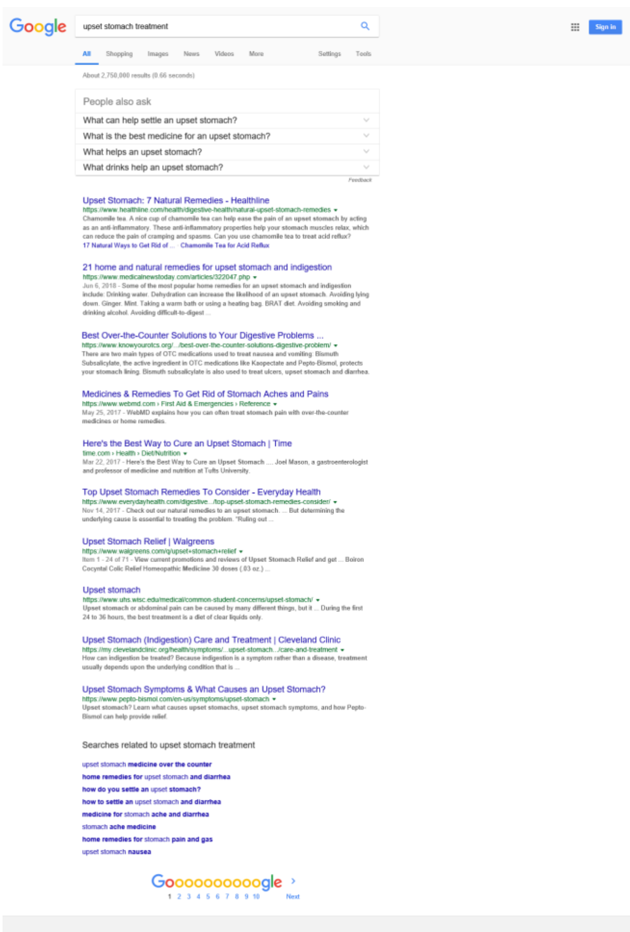
Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

Qualtrics Main Survey Questions for Health Topic #2 (Participants will see 2 topics out of 3)

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: When should I see a medical professional for an upset stomach?



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

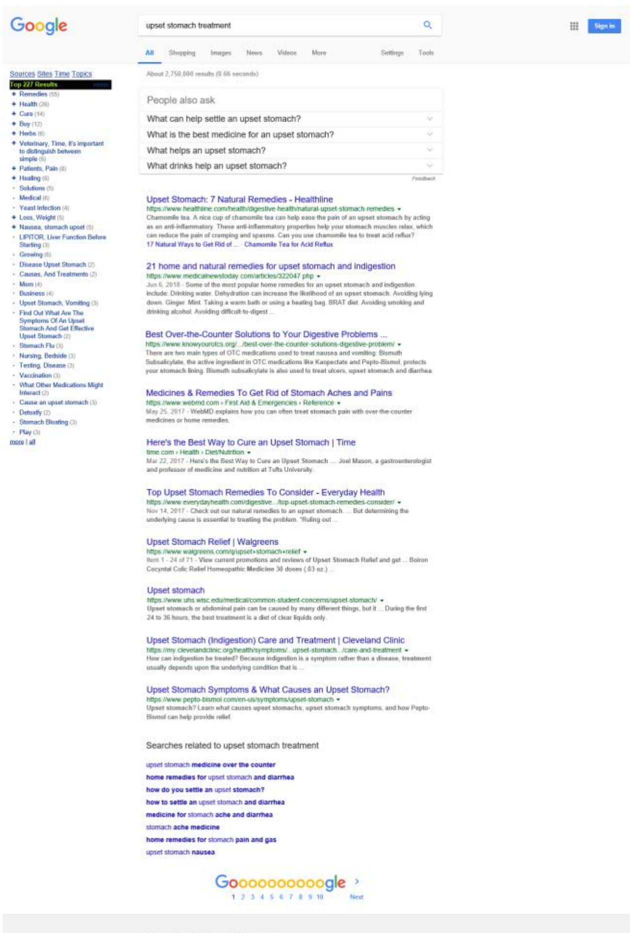
Overall, what is **your opinion** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: When should I see a medical professional for an upset stomach?



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by using the folders in the left sidebar.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **your opinion** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: When should I see a medical professional for an upset stomach?

Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **<u>refine your query</u>**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud below each listing in the search results.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **<u>visual appeal</u>**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

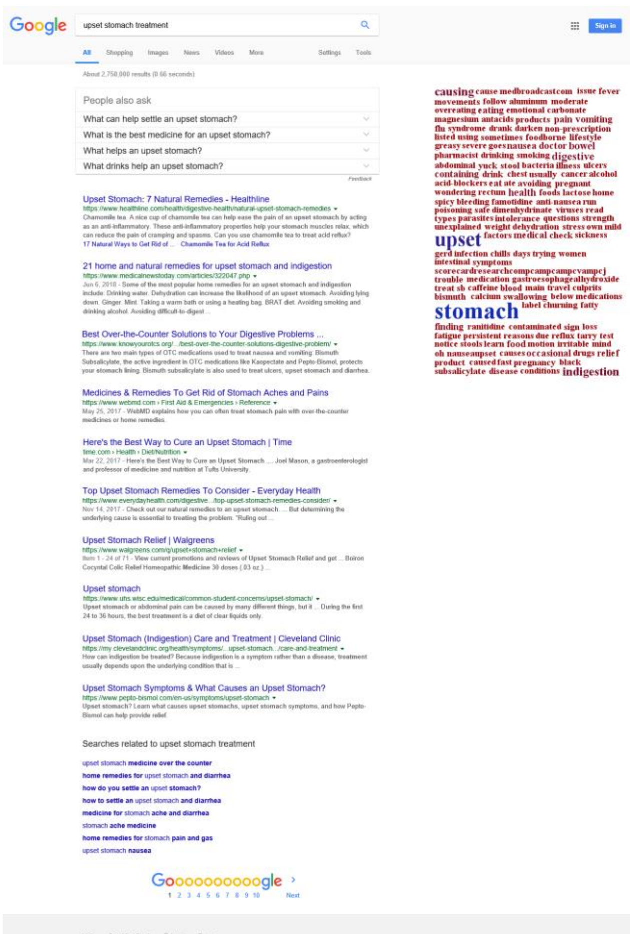
Overall, what is **<u>your opinion</u>** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: When should I see a medical professional for an upset stomach?



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud in the right sidebar.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **your opinion** of this search engine results display?

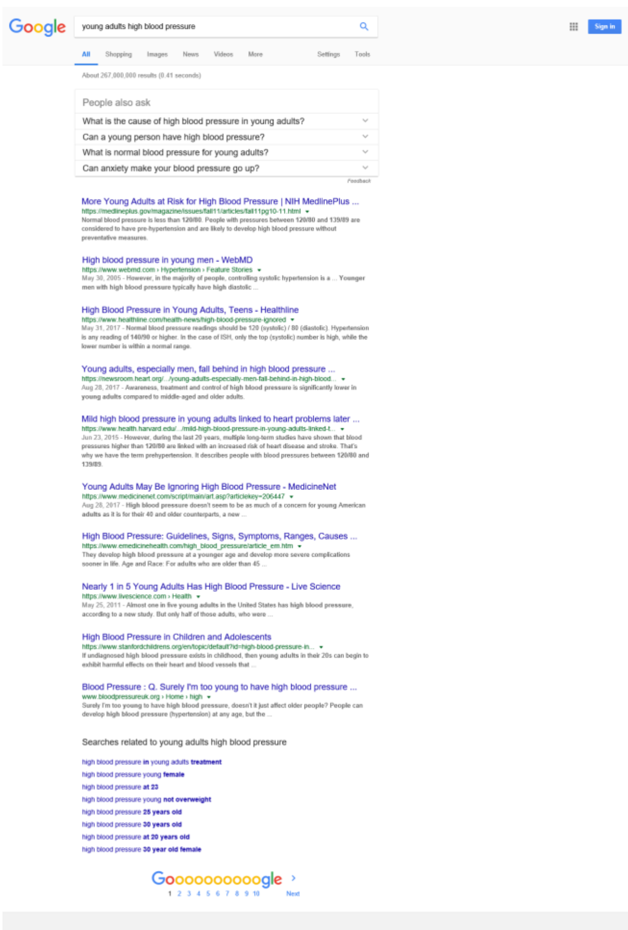
Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

Qualtrics Main Survey Questions for Health Topic #3 (Participants will see 2 topics out of 3)

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: Can college students get high blood pressure?



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

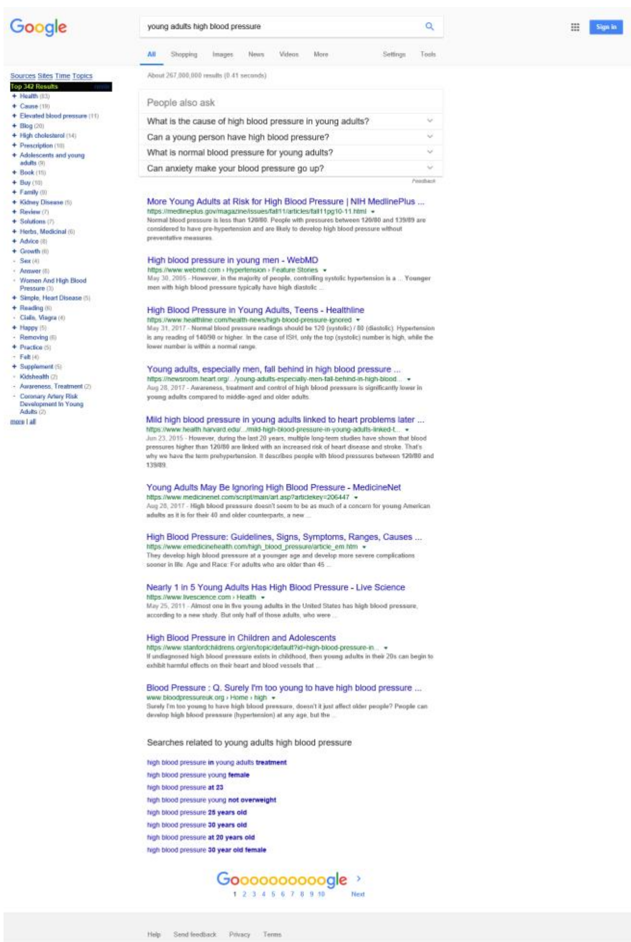
Overall, what is **your opinion** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: Can college students get high blood pressure?



Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **<u>refine your query</u>**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by using the folders in the left sidebar.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **<u>visual appeal</u>**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **<u>your opinion</u>** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: Can college students get high blood pressure?

Please rate this search engine results display on how helpful the display would be in helping you choose **<u>relevant</u>** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **<u>credible</u>** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **<u>quickly find what you need</u>**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **<u>refine your query</u>**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud below each listing in the search results.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **<u>visual appeal</u>**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **<u>your opinion</u>** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

The screen shot below shows one possible way to display search engine results for a health question.

Please review the display and then answer the questions below the screen shot.

Health question: Can college students get high blood pressure?

Please rate this search engine results display on how helpful the display would be in helping you choose **relevant** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how helpful the display would be in helping you choose **credible** results.

Not helpful at all	Slightly helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

Please rate this search engine results display on how useful it would be in helping you **quickly find what you need**.

Extremely useless	Moderately useless	Slightly useless	Neither useful nor useless	Slightly useful	Moderately useful	Extremely useful
○	○	○	○	○	○	○

Please rate this search engine results display on how easy it would be to **refine your query**.

For all four types of displays of search engine results in this study, assume that you can refine a query by:

- Using the People also ask questions
- Using the links in the "Searches related to ____" section at the bottom of the results
- Typing a new query

In addition, this search engine results display would allow you to refine a query by clicking on any of the terms in the word cloud in the right sidebar.

Extremely difficult	Moderately difficult	Slightly difficult	Neither easy nor difficult	Slightly easy	Moderately easy	Extremely easy
○	○	○	○	○	○	○

Please rate this search engine results display on **visual appeal**.

Terrible	Poor	Average	Good	Excellent
○	○	○	○	○

Overall, what is **your opinion** of this search engine results display?

Extremely negative	Moderately negative	Slightly negative	Neither positive nor negative	Slightly positive	Moderately positive	Extremely positive
○	○	○	○	○	○	○

Qualtrics Main Survey Informational Screen

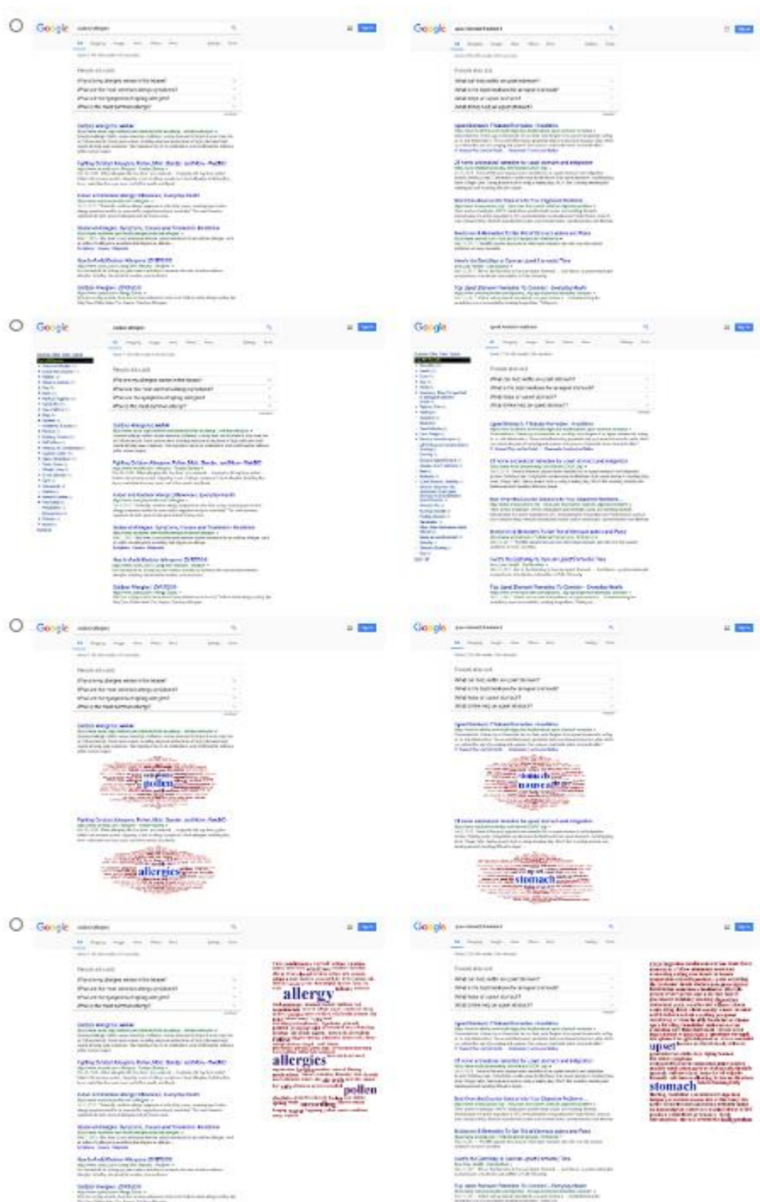
Good work! You made it through Section 1. Congratulations!

Sections 2 and 3 are much shorter.

Qualtrics Main Survey Concluding Questions

(For participants who viewed the screen shots for outdoor allergies and upset stomach. All questions and display types were presented randomly.)

Overall, which type of display did you **prefer** the most for viewing search engine results from health information searches?

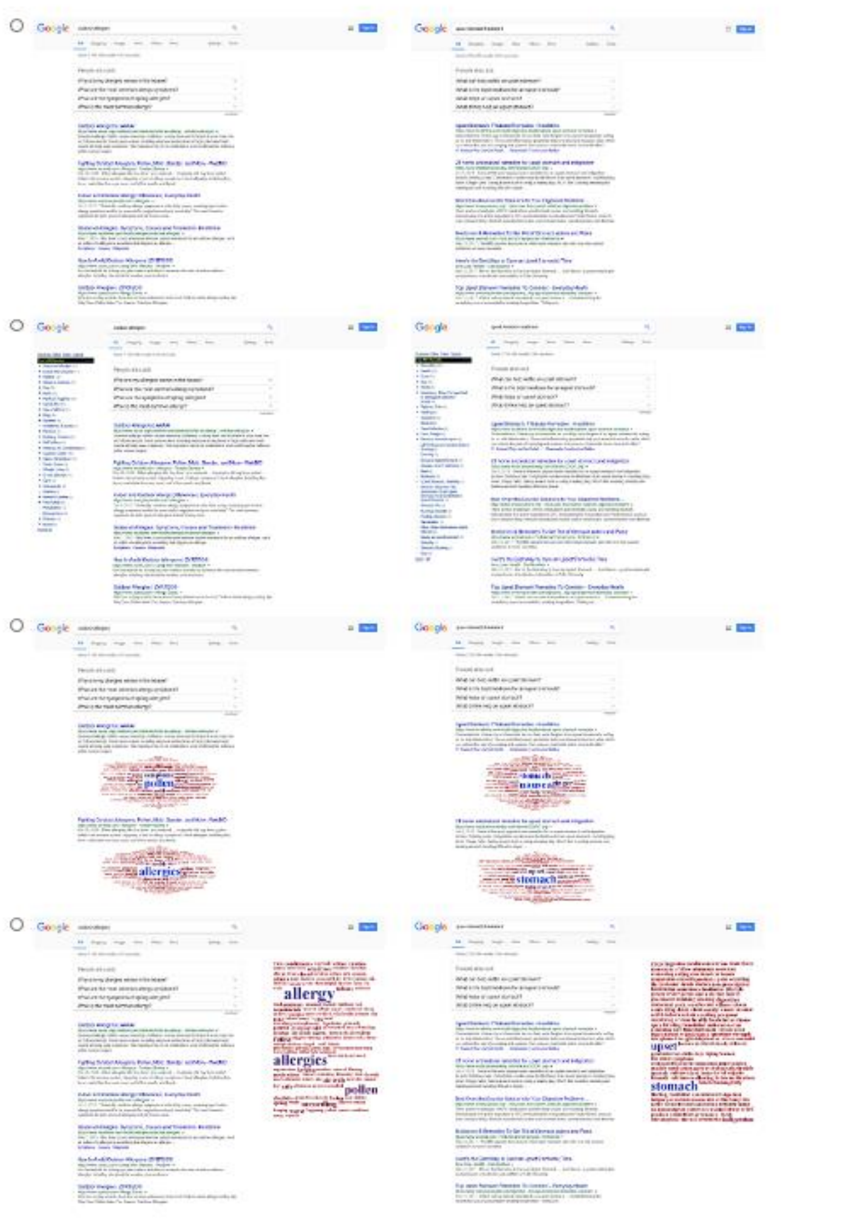


(For participants who viewed the screen shots for outdoor allergies and upset stomach.)

Which type of display did you **dislike** the most for viewing search engine results from health information searches?

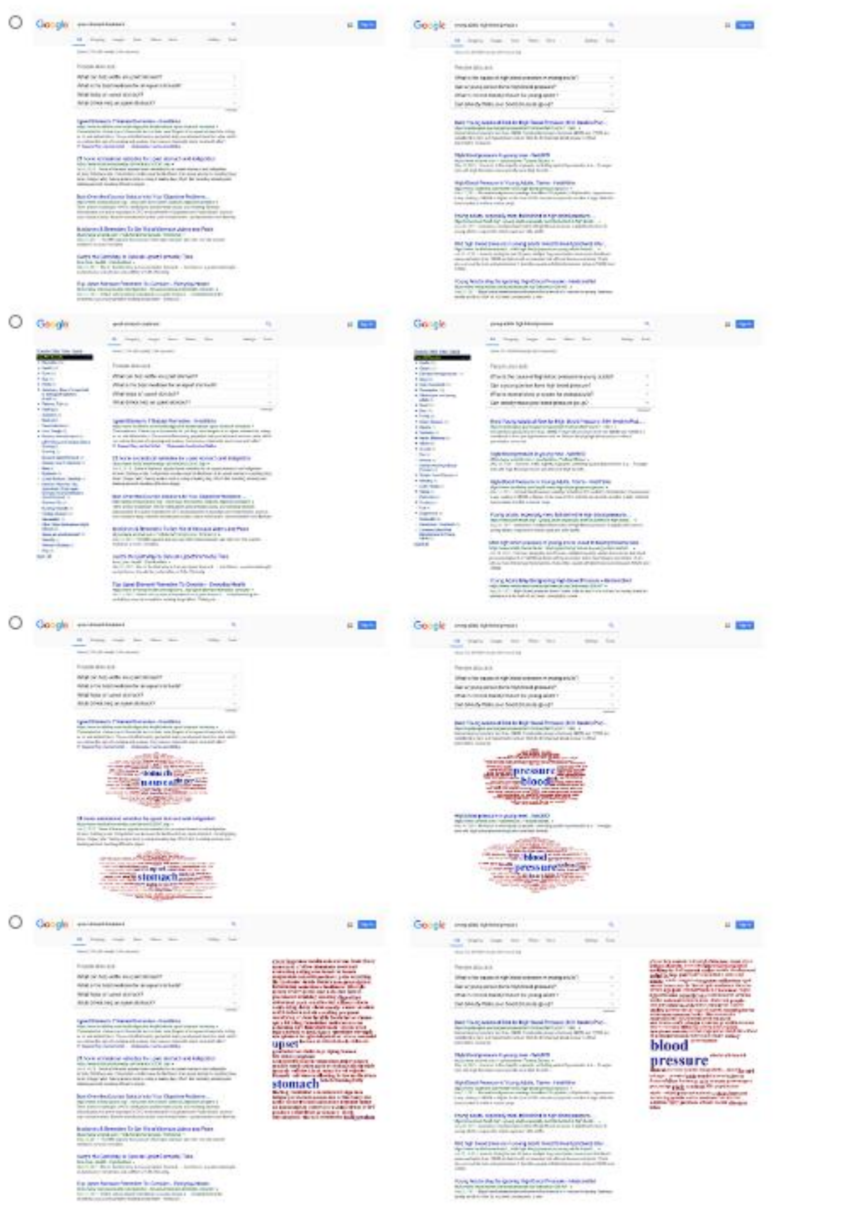
(For participants who viewed the screen shots for outdoor allergies and upset stomach.)

If a **browser extension or other customization** was available to ensure that your search engine results from health information searches would appear like any of these displays, which option **would you pick**?



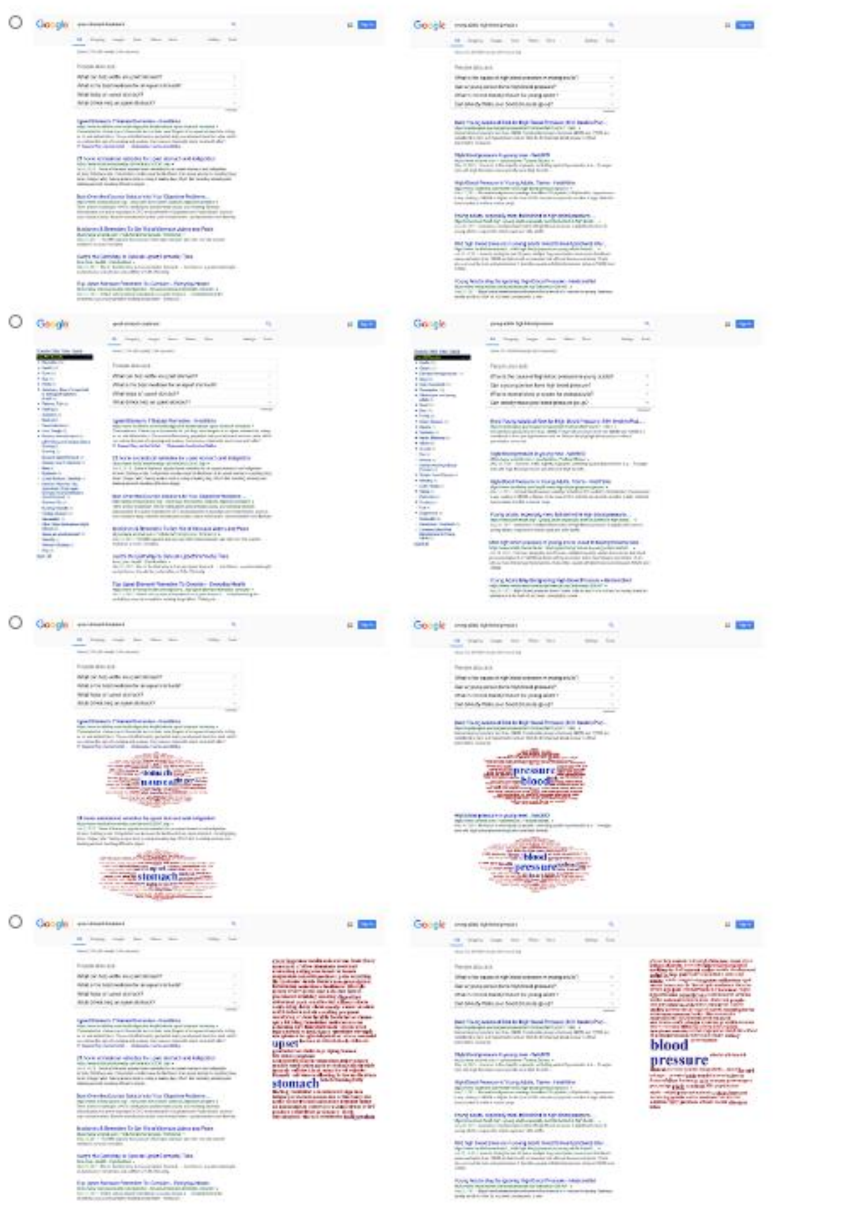
(For participants who viewed the screen shots for upset stomach and high blood pressure in young adults. All questions and display types were presented randomly.)

Overall, which type of display did you **prefer** the most for viewing search engine results from health information queries?



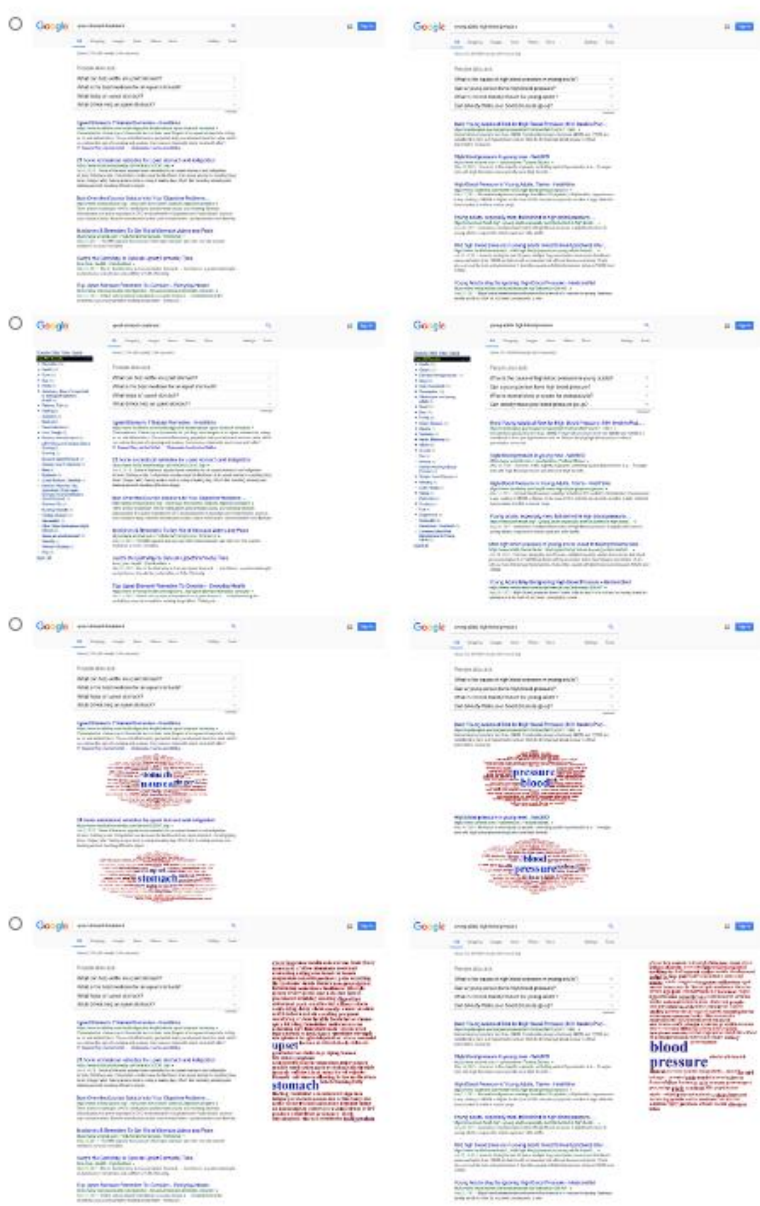
(For participants who viewed the screen shots for upset stomach and high blood pressure in young adults.)

Which type of display did you **dislike** the most for viewing search engine results from health information queries?



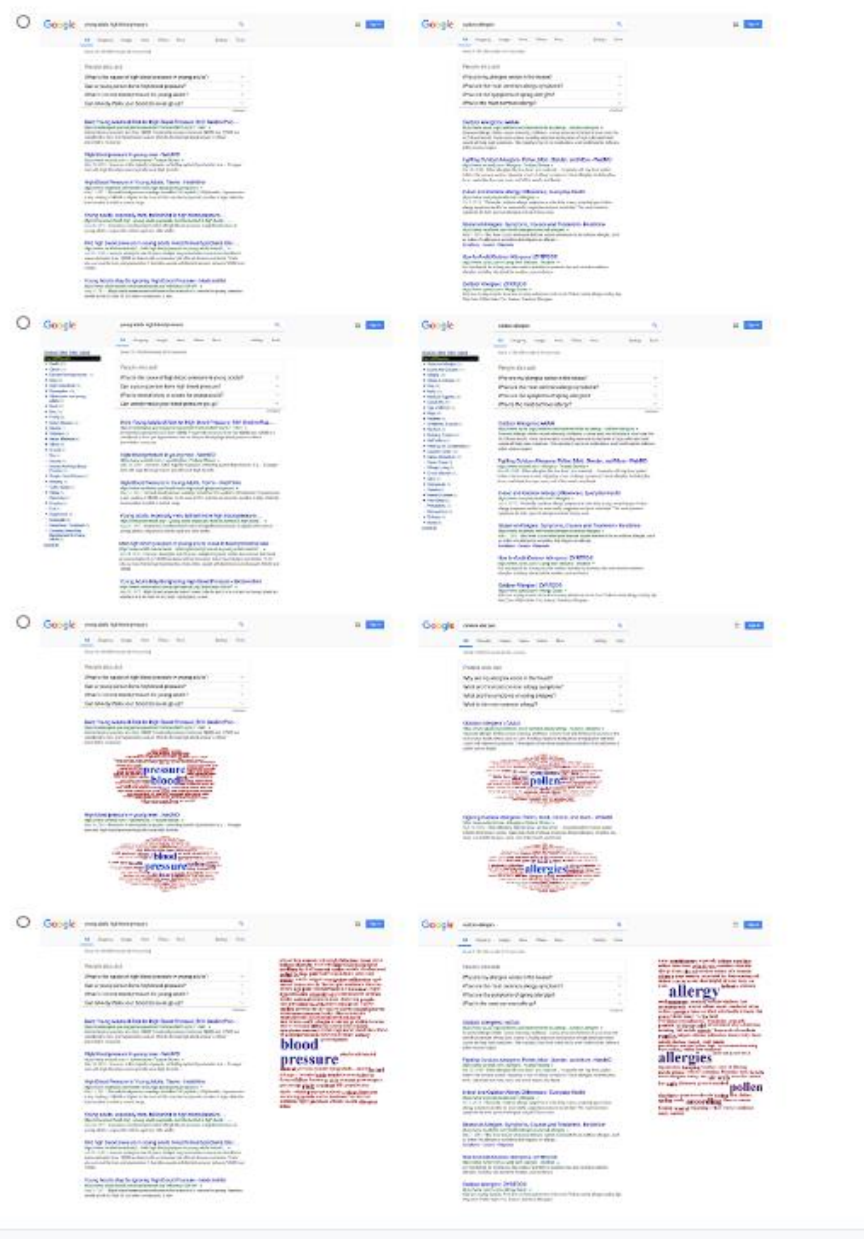
(For participants who viewed the screen shots for upset stomach and high blood pressure in young adults.)

If a **browser extension or other customization** was available to ensure that your search engine results from health information searches would appear like any of these displays, which option **would you pick**?



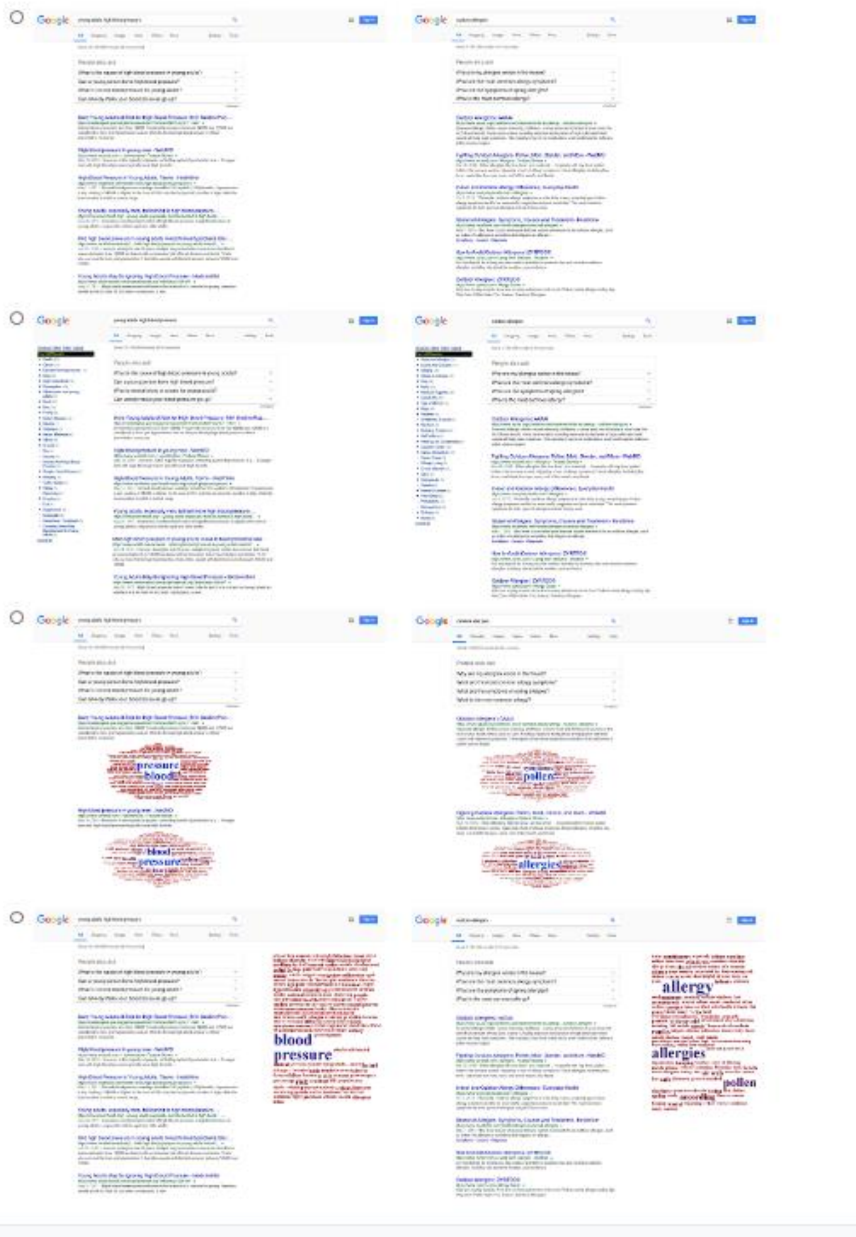
(For participants who viewed the screen shots for outdoor allergies and high blood pressure in young adults. All questions and display types were presented randomly.)

Overall, which type of display did you **<u>prefer</u>** the most for viewing search engine results from health information queries?



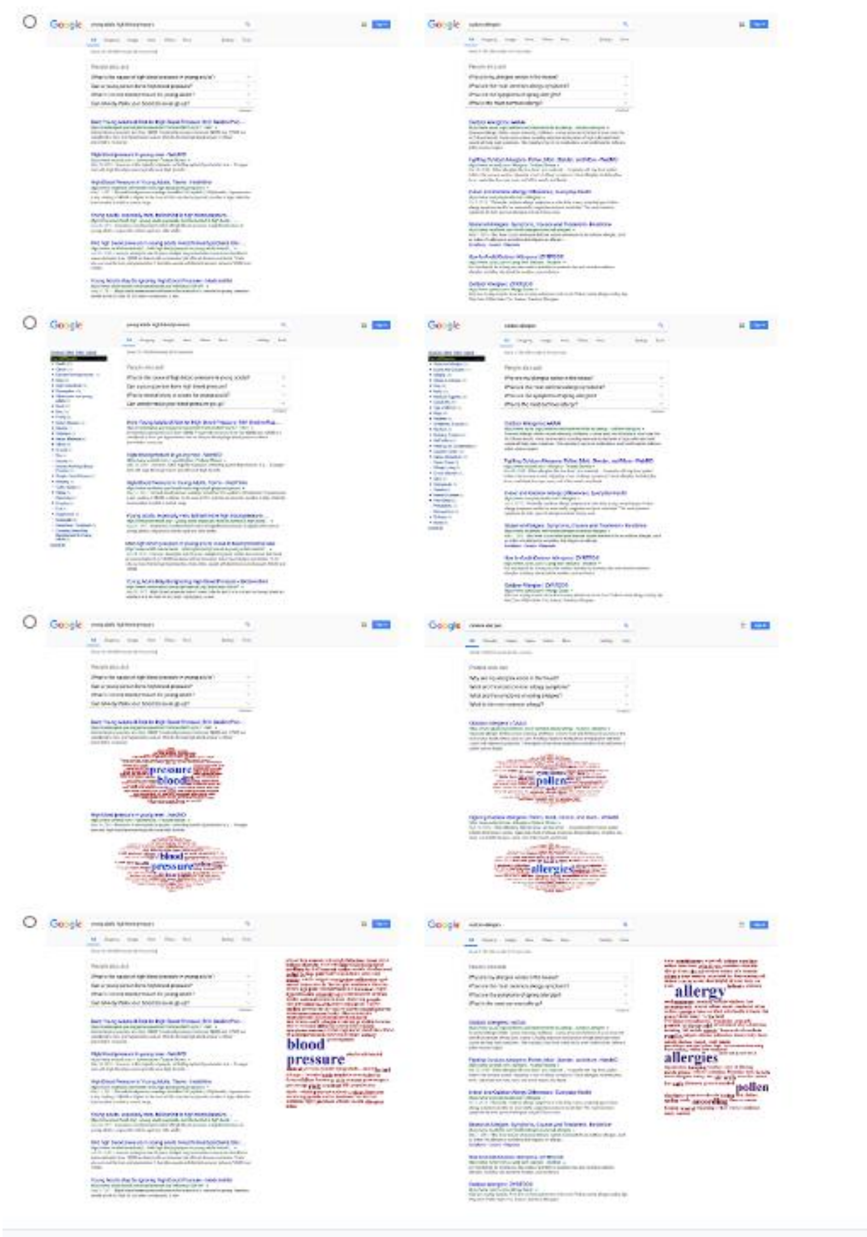
(For participants who viewed the screen shots for outdoor allergies and high blood pressure in young adults.)

Which type of display did you **dislike** the most for viewing search engine results from health information queries?



(For participants who viewed the screen shots for outdoor allergies and high blood pressure in young adults.)

If a **browser extension or other customization** was available to ensure that your search engine results from health information searches would appear like any of these displays, which option **would you pick**?



Qualtrics Main Survey Demographic Questions

What is your age?

- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65 - 74
- 75 or older

What is your ethnicity (check all that apply)?

- American Indian or Alaska Native
- Asian
- Black or African American
- Caucasian or White
- Hispanic or Latino/Latina
- Native Hawaiian or Pacific Islander
- Other

What is your gender identification?

- Female
- Male
- Transgender, Non-binary or Intersex
- Other
- Prefer not to say

What is your university affiliation (check all that apply)?

- Undergraduate Student
- Graduate, Postdoc or Professional Program Student
- Visiting Student or Other Type of Student
- Faculty
- Staff
- Hospital
- Retiree
- Other

Qualtrics Main Survey Ready to Submit Your Survey Responses Screen

You have reached the end of the questions.

Please use the next button to submit your answers in order to complete the survey. You will also be given the opportunity to register for the drawing for gift cards.

If you need to change an answer before submitting, please use the back button now.

Qualtrics Main Survey Question About Entering the Drawing

Thank you for taking the time to complete the survey. Your responses have been submitted. Do you want to provide your name and UNC email address in order to enter the drawing?

- ○ Yes
- ○ No

Qualtrics 2nd Survey Used to Enter the Drawing for One of Two Gift Cards

Please enter the below information in order to be entered into the drawing for one of two $50.00 gift cards. Your contact information will be stored in a different file than your responses to the survey.

What is your name?

What is your UNC email address?

If you do not have a valid UNC email address because you are a visiting scholar, adjunct faculty member, remote lab employee, offsite faculty member, research collaborator from another institution or a student at another university, please enter your email address here:

Submit

Qualtrics End of 2nd Survey Confirming Participant Has Been Entered into Drawing

You have been entered into the drawing for the gift cards.

Thank you.

Appendix F

Survey Logic for Main Survey in Qualtrics (extracted from Qualtrics)

Group: 18orOlder Screening Group

Standard: 18orOlder (1 Question)

Branch: New Branch
If
If Are you age 18 or older? No Is Selected
Or Are you age 18 or older? I prefer not to answer Is Selected
EmbeddedData
18orOlder = 0
EndSurvey: Screened Out
Branch: New Branch
If
If Are you age 18 or older? Yes Is Selected
EmbeddedData
18orOlder = 1

Group: Health Searcher Screening Group

Standard: Health searches? (1 Question)

Branch: New Branch
If
If Have you ever searched for health or medical information online... No Is Selected
Or Have you ever searched for health or medical information online... I prefer not to answer Is Selected
EmbeddedData
HealthSearcher = 0
EndSurvey: Screened Out
Branch: New Branch
If
If Have you ever searched for health or medical information online… Yes Is Selected
EmbeddedData
HealthSearcher = 1

Informed Consent Form

Standard: Informed Consent (1 Question)

Branch: New Branch
If
If Consent I do not consent, I do not wish to participate Is Selected
EndSurvey: I do not consent, I do not wish to participate

Standard: Discussion of 3 sections (1 Question)

Part 1 of Main Survey

BlockRandomizer: Evenly Present 2 of 3 Elements

Group: Health Topic 1 Group

EmbeddedData

Topic1Used = 1

BlockRandomizer: 4 - Evenly Present Elements

Block: Topic 1 Screen 1 (Ratings Questions)

Block: Topic 1 Screen 2 (Ratings Questions)

Block: Topic 1 Screen 3 (Ratings Questions)

Block: Topic 1 Screen 4 (Ratings Questions)

Group: Health Topic 2 Group

EmbeddedData

Topic2Used = 1

BlockRandomizer: 4 - Evenly Present Elements

Block: Topic 2 Screen 1 (Ratings Questions)

Block: Topic 2 Screen 2 (Ratings Questions)

Block: Topic 2 Screen 3 (Ratings Questions)

Block: Topic 2 Screen 4 (Ratings Questions)

Group: Health Topic 3 Group

EmbeddedData

Topic3Used = 1

BlockRandomizer: 4 - Evenly Present Elements

Block: Topic 3 Screen 1 (Ratings Questions)

Block: Topic 3 Screen 2 (Ratings Questions)

Block: Topic 3 Screen 3 (Ratings Questions)

Block: Topic 3 Screen 4 (Ratings Questions)

Standard: Done with Section 1

Part 2 of Main Survey

Branch: New Branch
If
If Topic1Used Is Equal to 1 And Topic2Used Is Equal to 1
Block: Topic1plus2 block (3 Concluding Questions)
EmbeddedData
Topic1and2Block = 1
Branch: New Branch
If
If Topic2Used Is Equal to 1 And Topic3Used Is Equal to 1
Block: Topic2plus3 block (3 Concluding Questions)
EmbeddedData
Topic2and3Block = 1
Branch: New Branch
If
If Topic3Used Is Equal to 1 And Topic1Used Is Equal to 1
Block: Topic3plus1 block (3 Concluding Questions)
EmbeddedData
Topic3and1Block = 1

Part 3 of Main Survey

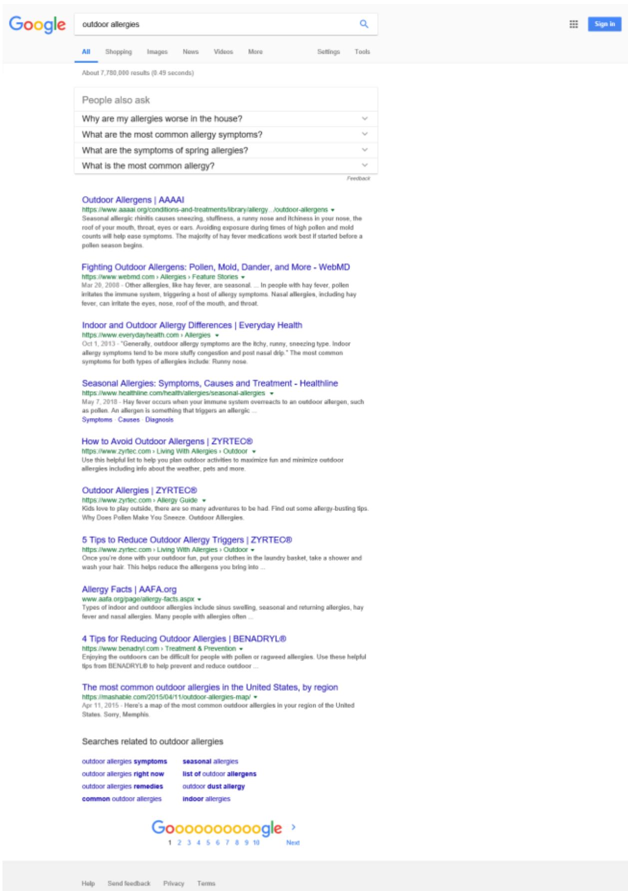
Standard: Demographic Questions (4 Demographic Questions)
Standard: BeforeSubmit (2 Questions)

Branch: New Branch
If
If Register for drawing? Yes Is Selected
EndSurvey: Survey is submitted and participant is redirected to a new URL for 2nd survey to register for the drawing
Branch: New Branch
If
If Register for drawing? Yes Is Not Selected
EndSurvey: Survey is submitted and participant is not redirected

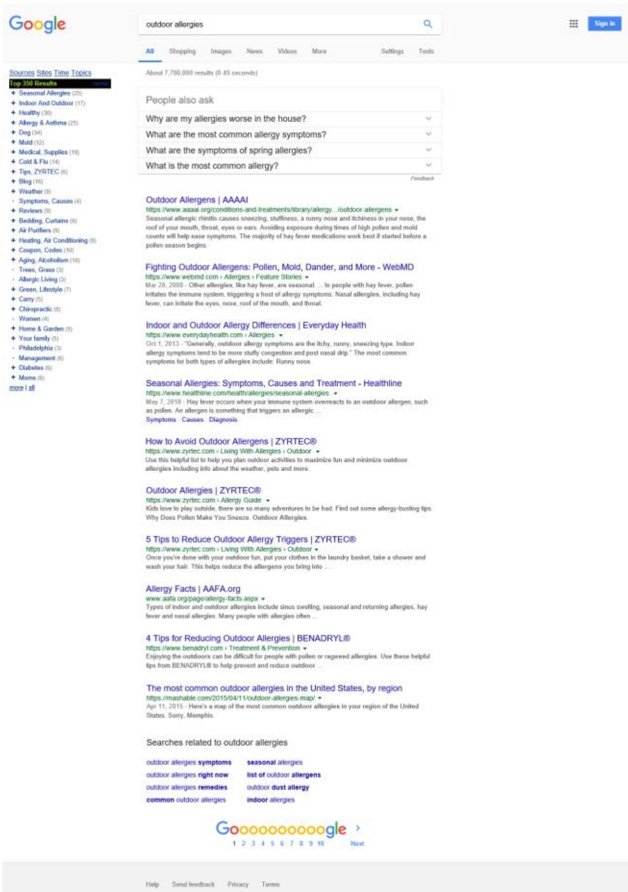
Appendix G

Examples of the Search Engine Results Displays

Display 1: Standard Google (Used as the Control)

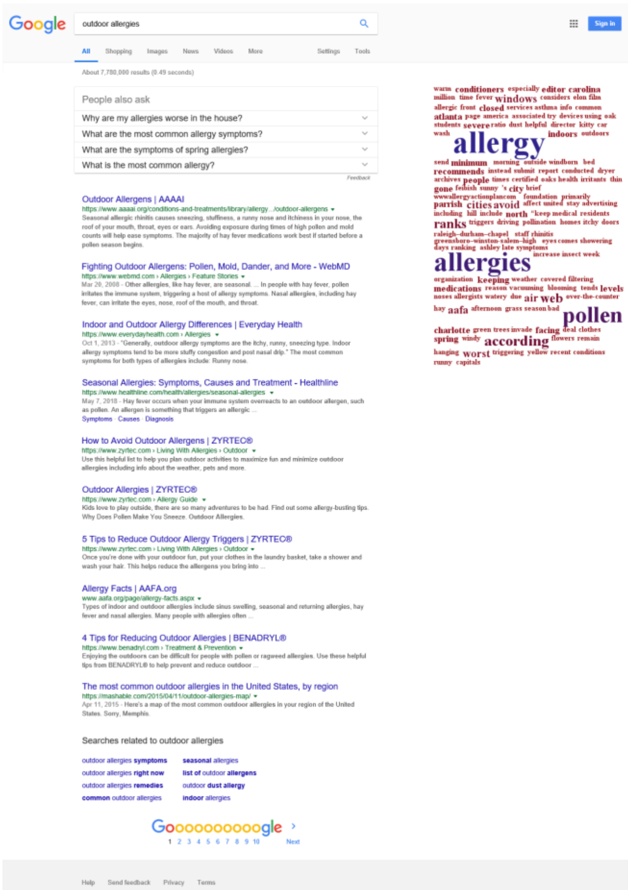


Display 2: Google Enhanced with Faceted Browsable Categories



Display 3: Google Enhanced with Word Cloud for Each Search Result

Display 4: Google Enhanced with Overview Word Cloud for Collection of Search Results



Appendix H

Very Preliminary Correlation Analysis

Part 1 Variables

Table H1

Cronbach Coefficient Alpha with Deleted Variable for Part 1 Variables

	Raw Variables		Standardized Variables	
Deleted Variable	Correlation with Total	Alpha	Correlation with Total	Alpha
Relevant	0.83	0.91	0.83	0.92
Credible	0.72	0.92	0.72	0.93
Quickly Find	0.86	0.90	0.86	0.91
Refine	0.76	0.91	0.75	0.93
Visual	0.77	0.91	0.76	0.92
Opinion	0.87	0.90	0.87	0.91

Table H2

Pearson Correlation Coefficients for Part 1 Variables

Row Labels for Results	Relevant	Credible	Quickly Find	Refine	Visual	Opinion
Relevant: Pearson Correlation Coefficients	1.00	0.74	0.78	0.67	0.64	0.76
Relevant: Prob > |r| under H0: Rho=0		<.0001	<.0001	<.0001	<.0001	<.0001
Relevant: Number of Observations	1592	1591	1592	1591	1592	1592
Credible: Pearson Correlation Coefficients	0.74	1.00	0.67	0.56	0.57	0.66
Credible: Prob > |r| under H0: Rho=0	<.0001		<.0001	<.0001	<.0001	<.0001
Credible: Relevant: Prob > |r| under H0: Rho=0	1591	1591	1591	1590	1591	1591
Quickly Find: Pearson Correlation Coefficients	0.78	0.67	1.00	0.75	0.69	0.81
Quickly Find: Prob > |r| under H0: Rho=0	<.0001	<.0001		<.0001	<.0001	<.0001
Quickly Find: Credible: Prob > |r| under H0: Rho=0	1592	1591	1592	1591	1592	1592
Refine: Pearson Correlation Coefficients	0.67	0.56	0.75	1.00	0.62	0.72
Refine: Prob > |r| under H0: Rho=0	<.0001	<.0001	<.0001		<.0001	<.0001
Refine: Quickly Find: Prob > |r| under H0: Rho=0	1591	1590	1591	1591	1591	1591
Visual: Pearson Correlation Coefficients	0.64	0.57	0.69	0.62	1.00	0.80
Visual: Prob > |r| under H0: Rho=0	<.0001	<.0001	<.0001	<.0001		<.0001
Visual: Refine: Prob > |r| under H0: Rho=0	1592	1591	1592	1591	1592	1592
Opinion: Pearson Correlation Coefficients	0.76	0.66	0.81	0.72	0.80	1.00
Opinion: Prob > |r| under H0: Rho=0	<.0001	<.0001	<.0001	<.0001	<.0001	
Opinion: Visual: Prob > |r| under H0: Rho=0	1592	1591	1592	1591	1592	1592

Note. 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings. One rating missing from credible and one rating missing from refine due to one participant not completing the ratings. Combined credible with refine is missing 2 ratings.

Table H3

Spearman Correlation Coefficients for Part 1 Variables

Row Labels for Results	Relevant	Credible	Quickly Find	Refine	Visual	Opinion
Relevant: Spearman Correlation Coefficients	1.00	0.73	0.79	0.67	0.64	0.77
Relevant: Prob > |r| under H0: Rho=0		<.0001	<.0001	<.0001	<.0001	<.0001
Relevant: Number of Observations	1592	1591	1592	1591	1592	1592
Credible: Spearman Correlation Coefficients	0.73	1.00	0.67	0.55	0.55	0.65
Credible: Prob > |r| under H0: Rho=0	<.0001		<.0001	<.0001	<.0001	<.0001
Credible: Number of Observations	1591	1591	1591	1590	1591	1591
Quickly Find: Spearman Correlation Coefficients	0.79	0.67	1.00	0.75	0.69	0.82
Quickly Find: Prob > |r| under H0: Rho=0	<.0001	<.0001		<.0001	<.0001	<.0001
Quickly Find: Number of Observations	1592	1591	1592	1591	1592	1592
Refine: Spearman Correlation Coefficients	0.67	0.55	0.75	1.00	0.61	0.72
Refine: Prob > |r| under H0: Rho=0	<.0001	<.0001	<.0001		<.0001	<.0001
Refine: Number of Observations	1591	1590	1591	1591	1591	1591
Visual: Spearman Correlation Coefficients	0.64	0.55	0.69	0.61	1.00	0.80
Visual: Prob > |r| under H0: Rho=0	<.0001	<.0001	<.0001	<.0001		<.0001
Visual: Number of Observations	1592	1591	1592	1591	1592	1592
Opinion: Spearman Correlation Coefficients	0.77	0.65	0.82	0.72	0.80	1.00
Opinion: Prob > |r| under H0: Rho=0	<.0001	<.0001	<.0001	<.0001	<.0001	
Opinion: Number of Observations	1592	1591	1592	1591	1592	1592

Note. 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings. One rating missing from credible and one rating missing from refine due to one participant not completing the ratings. Combined credible with refine is missing 2 ratings.

Table H4

Kendall Tau b Correlation Coefficients for Part 1 Variables

Row Labels for Results	Relevant	Credible	Quickly Find	Refine	Visual	Opinion
Relevant: Kendall Tau b Correlation	1.00	0.66	0.70	0.58	0.57	0.67
Relevant: Prob > |tau| under H0: Tau=0		<.0001	<.0001	<.0001	<.0001	<.0001
Relevant: Number of Observations	1592	1591	1592	1591	1592	1592
Credible: Kendall Tau b Correlation	0.66	1.00	0.58	0.47	0.48	0.55
Credible: Prob > |tau| under H0: Tau=0	<.0001		<.0001	<.0001	<.0001	<.0001
Credible: Number of Observations	1591	1591	1591	1590	1591	1591
Quickly Find: Kendall Tau b Correlation	0.70	0.58	1.00	0.66	0.59	0.72
Quickly Find: Prob > |tau| under H0: Tau=0	<.0001	<.0001		<.0001	<.0001	<.0001
Quickly Find: Number of	1592	1591	1592	1591	1592	1592
Refine: Kendall Tau b Correlation	0.58	0.47	0.66	1.00	0.52	0.62
Refine: Prob > |tau| under H0: Tau=0	<.0001	<.0001	<.0001		<.0001	<.0001
Refine: Number of Observations	1591	1590	1591	1591	1591	1591
Visual: Kendall Tau b Correlation	0.57	0.48	0.59	0.52	1.00	0.71
Visual: Prob > |tau| under H0: Tau=0	<.0001	<.0001	<.0001	<.0001		<.0001
Visual: Number of Observations	1592	1591	1592	1591	1592	1592
Opinion: Kendall Tau b Correlation	0.67	0.55	0.72	0.62	0.71	1.00
Opinion: Prob > |tau| under H0: Tau=0	<.0001	<.0001	<.0001	<.0001	<.0001	
Opinion: Number of Observations	1592	1591	1592	1591	1592	1592

Note. 199 participants x 2 health topics x 4 screens rated per topic = 1592 total ratings. One rating missing from credible and one rating missing from refine due to one participant not completing the ratings. Combined credible with refine is missing 2 ratings.

Part 2 Variables

Table H5

Cronbach Coefficient Alpha with Deleted Variable for Part 2 Variables

	Raw Variables		Standardized Variables	
Deleted Variable	Correlation with Total	Alpha	Correlation with Total	Alpha
Prefer	0.37	-0.91	0.37	-0.91
Dislike	-0.33	0.85	-0.33	0.85
Browser	0.36	-0.86	0.36	-0.86

Note. Data represented as binomial to make numerical data for correlation analysis.

n = 796 (4 screens x 199 participants)

Table H6

Pearson Correlation Coefficients for Part 2 Variables

Row Labels for Results	Prefer	Browser	Dislike
Prefer: Pearson Correlation Coefficients	1.00	0.74	-0.30
Prefer: Prob > |r| under H0: Rho=0		<.0001	<.0001
Browser: Pearson Correlation Coefficients	0.74	1.00	-0.31
Browser: Prob > |r| under H0: Rho=0	<.0001		<.0001
Dislike: Pearson Correlation Coefficients	-0.30	-0.31	1.00
Dislike: Prob > |r| under H0: Rho=0	<.0001	<.0001	

Note. Data represented as binomial to make numerical data for correlation analysis.

n = 796 (4 screens x 199 participants)

Table H7

Spearman Correlation Coefficients for Part 2 Variables

Row Labels for Results	Prefer	Browser	Dislike
Prefer: Spearman Correlation Coefficients	1.00	0.74	-0.30
Prefer: Prob > |r| under H0: Rho=0		<.0001	<.0001
Browser: Spearman Correlation Coefficients	0.74	1.00	-0.31
Browser: Prob > |r| under H0: Rho=0	<.0001		<.0001
Dislike: Spearman Correlation Coefficients	-0.30	-0.31	1.00
Dislike: Prob > |r| under H0: Rho=0	<.0001	<.0001	

Note. Data represented as binomial to make numerical data for correlation analysis.

n = 796 (4 screens x 199 participants)

Table H8

Kendall Tau b Correlation Coefficients for Part 2 Variables

Row Labels for Results	Prefer	Browser	Dislike
Prefer: Kendall Tau b Correlation Coefficients	1.00	0.74	-0.30
Prefer: Prob > |tau| under H0: Tau=0		<.0001	<.0001
Browser: Kendall Tau b Correlation Coefficients	0.74	1.00	-0.31
Browser: Prob > |tau| under H0: Tau=0	<.0001		<.0001
Dislike: Kendall Tau b Correlation Coefficients	-0.30	-0.31	1.00
Dislike: Prob > |tau| under H0: Tau=0	<.0001	<.0001	

Note. Data represented as binomial to make numerical data for correlation analysis.

n = 796 (4 screens x 199 participants)