May 2014

SEARCHING AS THINKING: THE ROLE OF CUES IN QUERY REFORMULATION

Veronica Maidel
Syracuse University

Abstract

Given the growing volume of information that surrounds us, search, and particularly web search, is now a fundamental part of how people perceive and experience the world. Understanding how searchers interact with search engines is thus an important topic both for designers of information retrieval systems and educators working in the area of digital literacy. Reaching such understanding, however, with the more established, system-centric, approaches in information retrieval (IR) is limited. While inherently iterative nature of the search process is generally acknowledged in the field of IR, research on query reformulation is typically limited to dealing with the "what" or the "how" of the query reformulation process. Drawing a complete picture of searchers' behavior is thus incomplete without addressing the "why" of query reformulation, including what pieces of information, or cues, trigger the reformulation process. Unpacking that aspect of the searchers' behavior requires a more user-centric approach.

The overall goal of this study is to advance understanding of the reformulation process and the cues that influence it. It was driven by two broad questions about the use of cues (on the search engine result pages or the full web pages) in the searchers' decisions regarding query reformulation and the effects of that use on search effectiveness. The study draws on data collected in a lab setting from a sample of students who performed a series of search tasks and then went through a process of stimulated recall focused on their query reformulations. Both, query reformulations recorded during the search tasks and cues elicited during the stimulated recall exercise, were coded and then modeled using the mixed effects method. The final models capture the relationships between cues and query reformulation strategies as well as cues and search effectiveness; in both cases some relationships are moderated by search expertise and domain knowledge.

The results demonstrate that searchers systematically elicit and use cues with regard to query reformulation. Some of these relationships are independent from search expertise and domain knowledge, while others manifest themselves differently at different levels of search expertise and domain knowledge. Similarly, due to the fact that the majority of the reformulations in this study indicated a failure of the preceding query, mixed results were achieved with identifying relationships between the use of cues and search effectiveness. As a whole, this work offers two contributions to the field of user-centered information retrieval. First, it reaffirms some of the earlier conceptual work about the role of cues in search behavior, and then expands on it by proposing specific relationships between cues and reformulations. Second, it highlights potential design considerations in creating search engine results pages and query term suggestions, as well as and training suggestion for educators working on digital literacy.

SEARCHING AS THINKING: THE ROLE OF CUES IN QUERY REFORMULATION

by

Veronica Maidel

B.S., Tel-Aviv University, 2001
M.S., Ben-Gurion University, 2008

DISSERTATION

Submitted to the Graduate School at Syracuse University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Information Science and Technology

SYRACUSE UNIVERSITY
Syracuse, New York
May, 2014

Acknowledgements

This research would not be possible without the help and support of many people. First, I would like to thank my advisors Howard and Liz for being so patient, supportive, responsive, thoughtful, and for pushing me forward. The two of you always kept my best interests in mind and I am sincerely grateful for that. I also want to express gratitude to the rest of my committee, Barbara Kwasnik and Nick Belkin for their input and advice. Thanks to Diane Kelly and Yang Wang, my readers, for their feedback, which made the final manuscript more rigorous and complete. I would also like to thank Jeff Stanton, who I got to work with during my initial years at the iSchool and who taught me a lot about how to approach research.

I am very grateful for the Jeffrey Katzer Doctoral Award and the Katzer Doctoral Research Fund provided by the iSchool. Those were crucial for timely completion of the data collection and analysis.

I would like to thank my friends and family for their moral support and encouragement. Thank you Mom, Dad, Jenya, Elik, Matan, and my extended family for cheering me along the away and for being there for me. A special thank you goes to all my friends locally and worldwide for helping me stay sane. I am particularly grateful to Julia Pearlman and Avital Sabo, who encouraged and entertained me from afar on a daily basis. I am thankful for your loyal friendship in spite of the geographical distance. Thanks to my cohort, especially Mary Grace Flaherty and Angela Ramnarine-Rieks, for making this journey more social.

Finally, last but not least, the light of my life, my husband Dima and daughter Shira. You brighten up my days, even the roughest ones. Dima, I wouldn't be able to complete this quest without your endless help reading my drafts, exchanging ideas, and pushing me forward when I was truly discouraged.

Thank you!
Veronica

Table of Contents

List of Figures

List of Tables

1 Problem Statement

1.1 Motivation for the study

The rapidly growing volumes of information and its ever increasing importance in contemporary society, made search a ubiquitous aspect of how people experience the world. As such, understanding how people interact with search engines is an important topic for educators, designers, and users of information retrieval systems. Reflecting the change in the importance and ubiquity of search, there is a growing interest in the user focused research in the field of information retrieval (IR); that is in contrast to the system focused research that characterized the field in its early days when search activity was bound primarily to professional settings. Drawing on the established notion of search as an inherently iterative process that includes numerous query reformulation attempts, we see a growing interest in situating this behavior in context of people's daily routines and motivations (e.g. M. J Bates, 1989; Kelly, 2009; Saracevic, 1996a; Spink, 1997; Xie, 2000).

Query reformulation behavior on the web, seems to be a straight forward process. Online search typically involves formulation of a query consisting of query term(s) with output presented in the form of retrieved web pages arranged in a particular way on the search results page (SERP[1]). Based on these search results, the searcher can decide to (a) terminate the search and end the search session, (b) terminate that particular search and engage in another search on a different topic, or (c) reformulate the initial query in some manner with the intent to improve the retrieved outcome (Hembrooke, Granka, Gay, & Liddy, 2005). Another option would be to follow up on

[1] The following terminology regarding the results returned by the search engine will be used in this document: *SERPs* are the search engine result pages returned by the search engine after an execution of a query. SERPs are comprised of multiple *snippets*. Each snippet includes a *title*, *summary*, and a *URL*. Clicking on the snippet's title will lead to a *full web page*.

the results of the previous query. In this type of strategy, an initial query is used to get results which give an overview of what documents exist on the topic, or how the topic is linguistically treated, followed by a reformulated or new query on the same topic, but in light of the initial result. This whole decision-making process is largely transparent to the searchers, particularly on frequently used and well established search engines, such as Google, where users have developed habits and routines for interacting with the system. Similarly to experienced drivers, who just drive, without consciously thinking about each interaction with their car's controls, online searchers just search, without consciously evaluating each interaction with the search engine.

Such seamless integration of search-related decision-making is possible largely due to search engines becoming better from the system point view. The searchers do not feel the need to consciously learn (or guess) the internal logic of the search engine in order to (re)formulate their queries. Yet, no matter how good and refined an IR system will become, there will always be gaps between its intended uses and qualities and its actual use and adaptation by searchers in particular contexts. After all, the system designers are left with inferences about the cognitive processes of their searchers based on their behavior alone, and searchers do change and adjust their behavior when interacting with the search engine through an iterative process of (re) formulation of their queries and evaluation of the search results. As such, unpacking the tacit thinking of the users that goes into query (re)formulation is the next frontier of IR research with implication for design of web search systems, as well as for the educational efforts aimed at training better searchers.

Earlier research about search decision-making dealt primarily with the evaluation of relevance of search results (dating back to Saracevic, 1969, 1975). The overall goal of this study is to start

unpacking the decision-making process behind query reformulation. What elements of the SERP, or the full web pages, do searchers pay attention to? How do they interpret those elements – which I refer to as cues – and build on that interpretation in subsequent queries? Are those relationships between cues and reformulations random or do they form generalizable patterns? Do experienced and inexperienced searchers differ in their use of cues? Can we link the use of any of the cues to search effectiveness or lack thereof? And what role does domain knowledge play in the relationship between cues and reformulations? Exploring those questions will both advance our understanding of how people search and affect future (or validate current) search engine design.

The task of unpacking the decision-making process behind query reformulation has three areas of potential contribution. First, for classic IR research, better understanding of what cues searchers rely on in query reformulation and how exactly they utilize those cues, will add to the ability of modeling user behavior. This ability, in turn, can also help in search engine personalization. For example, knowing the difference between the behaviors of experienced and inexperienced searchers can suggest information elements to be included in the retrieved results. Similar distinctions may apply to designing IR systems for domain experts as opposed to novice searchers. Second, from an HCI perspective, mapping out the relationships between cues and query reformulations can influence the presentation of information on the SERP. Knowing that some elements are more useful in guiding productive search behavior can inform decisions to highlight those elements; conversely, knowing that some elements facilitate counterproductive search behavior can inform decisions to downplay or eliminate them from the SERP.

Finally, having formal models of user behavior in the process of query reformulation and potentially linking those models to search effectiveness, will offer an invaluable tool for educators in the area of digital literacy.

1.2 Problem definition

Query reformulation is an old problem in the field of IR. Traditionally, it has been the main interface to search behavior and has been studied as a proxy for the decision-making process in search. Lancaster (1969), for example, evaluated the performance of MEDLARS (Medical Literature Analysis and Retrieval Systems) and investigated reasons for recall and precision failures due to variations in exhaustivety and specificity of formulations (queries). Bates (1979), on the other hand, addressed bibliographic search. She developed definitions of four possible types of search tactics. One of them is *search formulation tactics*, which aid in the process of designing or redesigning the search formulation. Another category is *term tactics,* which aid in the selection and revision of specific terms within the search formulation. This older research investigated queries as a way to understand search behavior and the systems' response given this behavior.

Contemporary studies of the query reformulation process fall roughly in one of the two streams: classification of reformulation types or identification of patterns in interpretation and evaluation of SERPs by the searcher, after a query has been executed. The first stream houses important work that uses query reformulation as a proxy for various search strategies employed by searchers. Research in this stream typically relies on log analysis, focusing explicitly on the content of the queries and classifying them in terms of the perceived goals of the searcher (e.g. narrowing or generalizing a query) (Jansen, Booth, & Spink, 2009a, 2009b). This research,

however, does not explicitly focus on the reasoning of the searcher or her decision-making process.

The second stream of research on query reformulation focuses on searchers' evaluation of search results. Here, query reformulation is used as a proxy for searcher's decision-making process while deciding when to accept search results, alter the query, or terminate the search session (Aula, Majaranta, & Räihä, 2005; Cutrell & Guan, 2007; Guan & Cutrell, 2007; Lorigo et al., 2008). The tools and methodologies used in such studies vary. Some rely on click-through analysis, which, similarly to log analysis, is more system-centric even though it does represent behavioral data (Agichtein, Brill, Dumais, & Ragno, 2006; Agichtein, Brill, & Dumais, 2006; Joachims et al., 2007). Others employ eye tracking which detects the eye movements of searchers during their interaction with the search engine (Lorigo et al., 2008). The latter represents a more user-centric approach, as it adds a layer of physiological data about participants' behavior, which may be a closer proxy to cognitive processes, compared to click-through data. At the end of the day, however, even with eye-tracking data, the researchers are left to infer about the decision-making that underlies the behavior they observe. Moreover, evaluation of search results is usually studied separately from query reformulation and lacks the fine granularity that is needed to understand the cues and triggers for reformulation.

The two-stream picturing of the state of the field may appear somewhat simplistic and Chapter 2 goes into greater depth reviewing relevant literature in the field. The main point here is that there is currently little research on what drives search behaviors, what decision-making processes trigger query reformulation, or how those decision-making process play out in particular types of reformulations. In other words, the research so far has been more focused on *what* the searchers do and *how* they interact with the search engine, but not so much on *why* they perform these

actions; i.e. what is the reason behind, or what predicts, various search behaviors, such as reformulations. This project aims to address this gap by studying the interpretation and use of cues on the SERPs and on the full web pages. Understanding the mechanism that informs searchers' decision-making about query reformulation, is an important building stone in answering the *why* question.

1.3 Research objectives and questions

In order to answer the *why* question and better understand how searchers interpret search results for the purpose of query reformulation, I will focus on the analysis of cues used in this interpretation and evaluation process. As I described above and will discuss in greater detail in Chapter 2, the system-centric approach, which has dominated IR research on query reformulation, offers an important, yet limited insight into the drivers of searchers' behavior online. This limitation stems from the fact that researchers are required to infer about the reasons for behavior, based on behavioral data alone. A more human-centric approach is needed in order to unpack what drives searchers' behavior around reformulation.

In this study, I focus on investigating how searchers incorporate cues from the SERP or the full web pages in their decision-making that leads to query reformulation. The overall goal of this study is to advance understanding of the query reformulation process and the cues that may influence it. The overarching research question that guides this work is: *What cues influence the subsequent search behavior?* More specifically, I want to establish whether cues are identifiable, whether their use is systematic, and if so, whether it can be modeled. Situating this work in prior research, I want to pay particular attention to the roles of domain knowledge and search expertise

as factors known to have moderating effect on searchers' behaviors. This aspiration translates into two specific research questions[2]:

> **RQ1:** How do cues assist the searchers' decisions regarding query reformulation, in terms of which cue leads to which reformulation category? Do search expertise and domain knowledge serve as moderating factors for cue usage in reformulation?
>
> **RQ2:** What cues contribute to high/low level of effectiveness of the search? Do search expertise and domain knowledge serve as moderating factors for cue contribution to effectiveness of the search?

A few clarifications may be in order as I translate these research questions into actionable research objectives. First, the term *cue* can take on various meanings. In this study, I am particularly interested in the elements of information that drive query reformulation, whether those elements are presented on the SERP (e.g. snippet of a result or its summary, title, URL, etc.) or on the full web pages that are potential targets of the search. In other words, cues are the elements that the searchers rely on when assessing the differences between their anticipated and the actual search results . My first objective, thus, is to systematically map out the widest possible range of cues used by participants in the study. This will be largely an exploratory part of the study, as currently there is no known prior work exists on identifying and classifying cues used in manipulation.

Second, the concept of *query reformulation* will require additional operationalization in this study. Literature in the field typically refers to query reformulations as the transition between

[2] These two questions are the same questions as the ones which initially were proposed in the proposal stage, but a second part was added to the original formulation based on the third question from the proposal ("How does search expertise and domain knowledge influence the cues that searchers use to reformulate queries?"). Domain knowledge and search expertise were incorporated into these questions in order to test not only how they influence cues, but more how they influence the relationships between cues and query reformulations.

one query to the next. Those transitions can be divided into several categories such as generalization, specialization, substitution, new query, and more. Thus, my second objective is to check how well data collected for this study maps on to existing classifications (Huang & Efthimiadis, 2009; Liu, Gwizdka, Liu, Xu, & Belkin, 2010) and whether they call for establishment of additional categories. Since earlier research is focused on behavior alone, and this work tackles the reasoning, the later scenario is possible.

The type of task assigned to the participants will also require some investigation, since in order to trigger multiple query reformulations, the tasks need to be fairly difficult, but have a high level of clarity as to what the participant is expected to find. Meaning, that if the tasks are in question form, finding the answer should not be a trivial task, that is, the answer should not be retrieved on the first page by simply submitting the question verbatim. It should require several reformulations until the answer is found. As part of this study, I will explore which type of task is the most suitable for the purpose of cue discovery. Should they be in a form of simulated work tasks or in question form? Having tasks in question form, will make the evaluation of their correctness easier, because they will have a definite answer that can be evaluated objectively. Also, they will enable the participant to be more focused, as it will be more clear to them what the answer should look like and what they are looking for. On the other hand, simulated work tasks are more similar to real life situations and therefore are less artificial.

This study includes an exploration of the entire search process in each task, starting with formulation of the initial query, going through multiple query reformulations, and ending when the searcher decides to stop the search. I will refer to this process for the rest of this document as the *search sequence*. It starts with formulation of the initial query, continues through multiple

query reformulations, and ends when the searcher decides to stop the search[3]. My final objective for this project is to model the decision-making process in this sequence. One model will predict query reformulation as a function of the use of a repertoire of cues for participants with different levels of search expertise and domain knowledge (RQ1). Another model will predict potential contribution of the use of cues to search performance effectiveness for participants with different levels of search expertise and domain knowledge (RQ2).

[3] As I discuss later in the document, in this study a search sequence ended when the searcher decided to stop or when she was stopped due to time limitations, whichever came first.

2 Literature Review

The following review of the literature will cover topics and studies that have dealt separately or combined with query reformulation, interaction with search results, and evaluation of search results, specifically the notion of relevance. The review will start with an overview of studies that deal with information-seeking behavior in context, which focuses on humans, their information needs, and information behaviors. The next part of the review will focus specifically on studies that investigate the interaction of searchers with SERPs and describe the various features of that interaction. Studies that performed analysis of transaction logs for the study of query reformulations will be presented next followed by investigations of possible categories of query reformulation. Reformulation categories are of significant importance for this study, because part of the model I've developed includes the various types of reformulations that cues may influence. Finally, since the process of evaluation of results and decision-making regarding further query reformulation is tightly knit with the notion of relevance, research on this topic and an overall literature review conclusion will end this chapter.

2.1 Studies on information-seeking behavior in context[4]

In order to develop a deeper understanding of the decision-making process that a searcher goes through, there is a need to be able to address the *why* question. In the 80s, Belkin, Oddy, & Brooks (1982) developed the Anomalous State of Knowledge (ASK) model which aspires to explain *why* information need situations occur. Based on their framework, the anomaly in the user's state of knowledge which the user is unable to describe precisely is what causes the information need to rise. Trying to understand *why* searchers perform certain actions when

[4] The name of this section is borrowed from a typology by (Kelly, 2009) in which she describes a continuum of types of IIR research ranging from highly system-focused research to a highly human-focused research. Information-seeking behavior in context is at the very end of the scale, on the human focused side.

interacting with search engines in order to resolve their information need situations, can be achieved through methods that are usually applied in studies on information-seeking behavior in context. This type of research is described in Kelly's (2009) research continuum for conceptualizing interactive information retrieval research. According to this continuum, as opposed to system-focused information retrieval research, studies on information-seeking behavior place the largest emphasis on the human aspects of information retrieval. They focus on documenting search behavior in people's natural settings, while paying attention to the cognitive processes the users are going through as they search. In these studies, researchers may embed themselves into a setting as an observer and rely on mixed techniques, such as surveys, structured-observations, and interviews.

Another type of research that may be able to deal with the *why* question as well, is the "information-seeking behavior with IR systems" type of research. In these studies, often there are no experimental systems employed, but instead researchers prefer observing and documenting searchers' natural search behaviors and interactions with the search engine. This may include studies of searchers' search tactics, studies of how users make relevance assessments, or studies of how users re-find information on the web (Kelly, 2009).

Studies conducted with approaches that involve structured-observations and interviews often operate with small N (as opposed to studies using transaction logs). This type of data gathering and analysis aims to answer a different set of questions in the puzzle of searcher behavior, compared to studies that rely on transaction logs. More specifically, studies of information-seeking behavior in electronic environments and in context try to establish a link between the real information needs of the searchers and their subsequent information seeking behavior (Kelly, 2009), which ventures beyond the realm of pure search and sometimes involves

intermediaries (e.g. power searchers) and alternative ways of acquiring information (e.g. word of mouth).

One example of such a study that employed the interview technique is by O'Day & Jeffries (1993), who studied how library users dealt with the information they got back from human intermediaries. They conducted semi-structured interviews in the offices of the users, where the users had to recall how they expressed requests to the intermediary, the formats in which the results were delivered, and what they did to interpret and use the information when it arrived. O'Day and Jeffries focused on how the results of a search are digested and used to solve the problem that triggered the search. The findings showed that people explore by conducting a series of interconnected but diverse searches on a single, problem-based theme, rather than one extended search session per task. The authors characterized the information seeking process by presenting "triggers" and "stop conditions" that guide people's search behaviors. These findings are consistent with Bates' (1989) approach which stated that searchers seek information piece by piece rather than in one retrieved set, as well as that the searcher's queries evolved in the process.

Teevan et al. (2004) is a good example of a study focused on people's search behavior in their natural settings. Following the method of diary studies, the authors conducted semi-structured interviews in which participants reported their most recent search activity. The researchers interviewed each participant twice daily on five consecutive days, interrupting them in their offices at unspecified times. The researchers asked the participants to describe what they had most recently "looked at" and what they had most recently "looked for" in their email, their files, and on the web. Teevan et al. observed that instead of jumping directly to their information target using keywords, the participants navigated to their target with small, incremental steps (for

example, navigating to a certain website and once there, browsing through the website's pages until the desired information is found), using their contextual knowledge as a guide, even when they knew exactly what they were looking for in advance.

Hargittai (2002) used a similarly contextually-rich, yet more controlled approach to conduct structured observation of how people from the general population find information online. She conducted in-person observations of searchers on an internet-connected computer provided by the researchers and configured so as not to bias initial actions of the participants. The browser was configured to no preset homepage, there was no search engine field in the navigation toolbar of the browser, and the cache of the browser was cleared between sessions, so each participant could start with a clean slate. Focusing explicitly on the participants' ability to find information and navigate the web, the study revealed significant lack of relevant skills, including difficulty to use the browser's navigational features (e.g. "back" button), extensive reliance on the default settings of their browser or their internet service providers, limited use and understanding of search engines, and constrained ability to enter valid search terms including the common occurrence of spelling mistakes. Specifically with regard to search engines, the findings showed that knowing some of the workings of how to use a search engine can be extremely valuable (e.g. use of Boolean operators). People who realize the value of typing in more than one search term have a much easier time finding sites that address their information needs. Moreover, understanding how search engines rank pages and being able to understand search results (including the URLs of the results) can be quite valuable.

While Hargittai's 2002 study was geared towards evaluating the information skills of the general population, in a 2010 study Hargittai et al. focused explicitly on the notion of "trust" online. Using a similar setup to the 2002 study (i.e. computers configured to be information-

environment-neutral), this study targeted the entire process of information seeking, from search engine selection, through the evaluation of search results, all the way to the final destination. The searchers were presented with a series of tasks ranging from simpler questions with limited consequences to more complex assignments with potentially greater repercussions (e.g. health). The researchers who administered the tasks asked the participants to talk throughout their online activities in order to collect information about the searchers' thoughts and perceptions regarding web navigation. This study sheds additional light on the contextually-sensitive nature of the search processes. Hargittai et al. found that "the process of information-seeking is often as important as verifying the results when it comes to assessing the credibility of online content" (p.479), specifically she discovered that the participants tend to rely extensively on the search engine rankings. Moreover, the researchers suggest that the lower levels of information literacy are associated with higher trust in the search engines. Another finding of the study suggests that searchers have information seeking routines and that those routines are built primarily around brands, such as specific search engines or information repositories (e.g. Wikipedia).

The above mentioned studies of information-seeking behavior in electronic environments and in context bring the scholarly work closer to answering the question of *why* people make particular search choices, yet the quest for a more comprehensive answer continues. This situation is partially due to the fact that these studies do not aspire to answer the question of *why* searchers perform the actions that they do. For example, while Teevan et al. (2004) aimed to explore the range of orienteering behaviors, Hargittai (2002) explored the question of internet skills, and Hargitai et al. (2010) focused on evaluation of content presented in search engine results. Another limiting aspect of the existing inquiry lies in methodology. While studies like Teevan et al. pay a lot of attention to the actual information environment of their participants and the real-

life information seeking tasks, they rely primarily on user-reported behavior and recalling of their actions after they have been performed. At the same time, studies like Hargittai et al. do collect both user-reported and actual behavior of their participants, but this comes at the cost of context – first, the participants are forced to operate in an alien information environment and second, as much as the tasks are close to real-life situations, they are still artificial for the individual participants.

2.2 Interaction with Search Engine Result Pages (SERPs)

Gathering cues to help guide further interaction with the search engine can be performed by the searcher through evaluation of the search results such as looking at the snippets, the title, the URL, clicking on the target, and more. This section will explore the various ways in which previous research has studied the manner in which a searcher interacts with the results. Query reformulation is one of the ways the searcher can interact with the search engine in general. As such, it offers a "window" to study the search-related decision-making processes as well as the information needs of the searcher. Yet, the searcher's interaction with the search engine is not limited to query reformulation. Once the search results are presented, she can also scan them and click on links she perceives potentially relevant or deems as a good source for further investigation. These actions can be collected, measured, and analyzed according to automatically identifiable features (e.g. the order of the clicking or the scanning, the time it took to click on a result after query execution). This section does not deal with research that uses query content to identify reformulation patterns, but rather with work that uses non- content aspects of searcher's interactions with search engine results.

Typically, research on the interaction of searchers with web SERPs strives to identify interactional behavior patterns. This research usually explores how various types of searchers

employ different types of systems and exhibit different levels of performance. Studies in this vein do not have a declared or explicit goal of analyzing query reformulation, but sometimes they do employ query reformulation and its features as measures of interaction. In some cases, the purpose of these studies is to elicit implicit feedback regarding the usefulness of the results. In others, the goal is to discover the differences in searcher behavior in response to the SERP, under different independent variables. Research has been conducted to improve our understanding of how searchers evaluate the results (e.g. in what order) and how independent variables affect aspects such as query length or query execution rate. Examples of some of the independent variables employed in these types of studies include: searching expertise (Saito & Miwa, 2001; White & Morris, 2007), task type (Aula, Khan, & Guan, 2010; Liu et al., 2010), and system performance (Smith & Kantor, 2008; Smith, 2009).

Query-related interactions with the SERP are usually captured through the characteristics of the query and its execution. For instance, White and Morris (2007) explored the differences in search styles between advanced and non-advanced users. Among other factors, the authors also measured query related features, such as the number of queries per second (average number of queries per second between initial query and end-of-session), query repeat rate (fraction of queries that are repeats), and query word length (also employed in query reformulation studies). In other words, the authors utilized query-related features in order to characterize the searcher's interaction with the system. Smith and Kantor (2008) employed similar measures of searcher behavior, such as query rate (number of queries entered per minute of elapsed topic search time) and query length, in order to capture how searchers respond to SERPs from degraded retrieval systems. They found that when facing a system with poor performance, the searchers increased their query rate and were less likely to resubmit previously entered queries during the same task.

For the purpose of query reformulation research, query-related features can serve as independent variables that contribute to a more accurate prediction of query reformulation categories in an interactive setting. Previous research suggests that such independent variables may in fact have an effect on the likelihood of one query reformulation category or the other. For example, C. Liu et al. (2010) showed that searcher's query reformulation behavior can be explained with independent variables such as task type (in this study, the task types were: simple, hierarchical, and parallel). Their results demonstrated that specialization was most frequently used in simple tasks and word substitution was most frequently used in parallel tasks.

Additional studies utilized different query-related features related to interaction with the SERP. The rest of this section will review a number of examples of those various features. Reviewing these examples, it is important to keep in mind that none of the studies below was developed explicitly and exclusively for analysis of query reformulation. Thus, query reformulation features are not always treated as the dependent variable.

2.2.1 Time-related features

Capturing time as one of the interaction features has been performed in different ways. Time as an independent variable will also be mentioned in other sections, but this section focuses on time being a dependent variable. Aula and Nordhausen (2006) chose to model search success with Task Completion Speed (TCS). The researchers measured the speed of query reformulation and the speed of evaluating result documents in addition to the length of the queries and proportion of precise queries. They found that the speed of composing queries, the average number of query terms per query, the proportion of precise queries, and the participants' own evaluation of their search skills had significant influence on the TCS. The researchers also found that web experience and self-evaluated search skills had effects on the TCS. Similar time-related features

were employed by Lorigo et al. (2006) with the help of eye-tracking equipment. The authors looked at average time to complete the task, average time spent on web documents per question, and the percentage of time spent on Google result pages per question. They discovered that the overall time was influenced by whether the search task was informational or navigational. The percentage of time spent on result pages was lower for informational compared to navigational. Informational searches took more effort and time on average, but no task influence was found with respect to success.

Brand-Gruwel et al. (2005) and Saito and Miwa (2001) employed time-related features as their dependent variables. In both studies, the researchers examined the length of the search process as a function of the searchers' expertise. While Brand-Gruwel et al., found that experts spend more time on search tasks, Saito and Miwa, discovered that there were no significant differences between the experts and the novice searchers both for the general and the specific tasks. This difference in findings between the two studies could be due to differences in the types of tasks used by each study. Since Brand-Gruwel's et al. study put more emphasis on problem solving their tasks may have been more difficult. Although none of these studies addressed the question of query reformulation directly, the features analyzed in these works and the relationships found can inform research on query reformulation. For example, one can ask whether time spent on query reformulation can be used as a proxy for searchers' expertise.

2.2.2 Click-through features

Click-through data is another way to capture the interaction between the searcher and the retrieved search results. It enables exploring which results were examined by the searcher, the sequence of the clicks, the result evaluation strategy, and more. For example, in their investigation of searchers' interaction with SERPs, White and Morris (2007) looked for

differences in styles between advanced and novice users. Some of their measures related to the SERP itself and included: click position (average rank of clicked results), seconds to click (average search to result click interval), number of steps (average number of steps from the page following the results page to the end of the trail), and others. The searcher's expertise was determined by the extent of query operators, or syntax, utilization during the search (e.g. quotes, plus, minus). The findings demonstrated that those, who used syntax (the expert searchers), were more likely to explore search results by visiting the target pages, took less time to click on results, and overall were more successful in their searching. As before, for the query reformulation research, this example offers not only a number of additional ways to measure search-related behavior, but it also suggests a number of detailed relationships that can inform query reformulation related hypotheses (such as a relationship between the number of steps and the next reformulation category).

Click-through data has also been used together with data on query characteristics to infer the relevance of web pages (e.g. Joachims et al., 2007). Other than implicit feedback, interaction with search results has also been used to understand the intent of each individual query, but not specific reformulations or categories of reformulations. Some researchers have built models based on click-through and other interaction features in order to predict user preferences or infer search results relevance. For example, Agichtein, Brill, Dumais, and Ragano (2006) incorporated query-text features (e.g. average fraction of words shared with the next query), browsing features (number of hops to reach page from query), and click-through features (e.g. deviation from expected click frequency) to predict user preferences for search results. Agichtein's (2006) research was one of the few who used query contents in a quantitative manner, by using features like average fraction of words shared with the next query. Ashkan et al. (2009) found that SERP

and ad click-through features, query features, and the content of SERPs, as a group, are effective in detecting query intent (navigational, informational, and transactional). To automatically detect the same types of intent, another study (Guo & Agichtein, 2008) explored mouse trajectories on the result page. Although not directly related to query reformulation, such studies suggest that since click-through features have been shown useful in prediction of intent, they may also be useful in predicting query reformulation behavior. One recent study that incorporated both click-through features and query reformulations was performed by Kim and Can (2012). The authors employed indicators to characterize users' intent. They used characteristics such the number of query iterations, the number of identical queries used, the number of failed queries, the number of click-through queries, the number of saved queries, and the query interval. The researchers found that users' intent class might be related to those indicators.

2.2.3 Eye tracking to investigate interaction with SERPs

Eye tracking is one popular method for evaluating searchers' interaction with the SERPs. Eye tracking uses the ability to capture eye movements of searchers during their interaction with the search engine in order to depict their behavior in a more granular fashion (e.g. Lorigo et al., 2008). Aula et al. (2005) used eye tracking to study evaluation styles of search results. They found that searchers may be classified as economic or exhaustive evaluators. Economic evaluators made a decision regarding their next action (query reformulation, following a link) faster and utilized less information compared to exhaustive evaluators. Aula et al. results demonstrated that the exhaustive evaluators, possibly due to their lack of expertise, carefully evaluate the results before following a link or re-formulating a query. Compared to the economic evaluators, the exhaustive evaluators also depended more on snippets returned by the search engine. Findings showed that the economic evaluation style was particularly beneficial when

most of the results in the result page were relevant. In these cases, the task times were significantly shorter for economic than for exhaustive evaluators.

Other eye-tracking studies, such as those described in Lorigo et al. (2008) focused more on what snippets the users looked at, in what order, and for how long. The order of results was also explored by Guan and Cutrell (2007). They examined how search behaviors may change when target results are displayed at various positions, not necessarily the first one. The authors discovered that when targets were placed lower on the first page, searchers spent more time searching and were less successful in finding the target. The same authors conducted another study (Cutrell & Guan, 2007), in which they used eye tracking to explore the effects of changes in the way search results are presented. They found that adding information to the search results snippet significantly improved performance for informational tasks but degraded performance for navigational tasks.

The popularity of eye tracking studies has grown significantly in recent years, because it allows exploring interaction with the SERP, in order to gain a deeper understanding of where people invest attention and in what order, before they make a selection on the SERP (Hornof & Halverson, 2003). This type of inquiry offers an intriguing avenue for query reformulation research. For example, if we assume that the last viewed element on the SERP is the trigger for query reformulation, identifying patterns in such behavior can add a layer of reason-categories to the existing categories of query-reformulation behavior. However, eye tracking is limited only to data gathered from the searcher's gaze, and it requires numerous assumptions in order to be interpreted as reason-categories. The fact that a searcher looked at a certain element on the page does not necessarily mean it was the trigger for a subsequent reformulation or that it at all influenced the decision making process.

Another way to use eye tracking is to verify that the cues discovered in the interviewing process were in fact looked at during the observation. Although this may be a robust way to validate my analysis, the costs of equipment, the logistical complexity of its set-up, and additional data analysis on top of the analysis required to discovered the cues and their usage in conjunction with query reformulations rendered this option as impractical at this exploratory stage of my research. At a later stage eye tracking will be an important addition to my research program of exploring query reformulation cues.

2.2.4 Content features

A small body of studies explicitly examines searchers' interaction with SERPs together with the content of the queries. These studies offer concrete examples of how factors related to the interaction with the SERP considered together with query content can offer new perspectives on the query reformulation process. For example, Aula et al. (2010) found that in unsuccessful tasks (as opposed to successful tasks), searchers formulated more question queries (starting with why, what, when, etc.), used advanced operators more often (such as quotes in Google), and formulated the longest query in the middle of the session (instead of at the end). Along with query related measures the authors also utilized time as an indicator for unsuccessful tasks. Their results showed that in unsuccessful tasks, the searchers spent more time on the SERP both in terms of absolute time and as a proportion of the overall task time.

Hölscher and Strube (2000) conducted a qualitative study of the query reformulation process in response to SERPs. They found that novice searchers often make only small and ineffective changes to their queries, which in turn forces them to reformulate repeatedly. These studies offer concrete relationships between the features of interaction with the SERPs and query reformulation. In addition, another set of studies which will be discussed later, in the section on

query reformulation categories, uses query content analysis to arrive at these categories, but rarely makes any inferences between interaction and categories. A possible explanation why there hasn't been a lot of research that makes such inference is because it requires input from the searcher regarding her reasoning and therefore is much more labor intensive compared to analysis of data collected by the system.

2.3 Analysis of transaction logs for the study of query reformulations

Since the advent of search engines, one common resource used to study query reformulation is the transaction log. Transaction logs contain a record of the searcher's actions including click-through, query content, and time stamps. Transaction log analysis has been very popular because the data is readily accessible to researchers working for or with a search engine provider, contains ample transactions, and requires no need to deal with human subjects. One of the first studies to include logs was conducted by Meister & Sullivan (1967) who evaluated users' response to a system for retrieving document citations by analysis of transaction logs. Penniman (1975) was also one of the pioneers to provide research that dealt with user behavior based on transaction logs. As web-searching emerged, researchers increasingly began employing transaction logs for the purpose of understanding the user. In their book, Spink and Jansen (2004) offer an elaborate summary of research performed on web-search through analysis of transaction logs (Jansen, 2006).

The idea behind research that utilizes transaction logs, is to explore *what* decisions people make when faced with a list of search results or about the information needs that underlie their search behavior and how to predict them (Agichtein, Brill, Dumais, et al., 2006; Agichtein, Brill, & Dumais, 2006; Downey, Dumais, Liebling, & Horvitz, 2008; Jansen, Booth, & Spink, 2008; Jansen & Spink, 2006; Jansen, 2006; Rieh & Xie, 2006; Silverstein, Marais, Henzinger, &

Moricz, 1999). The focus is mainly on the path the searchers take through the search results until the moment they decide to click on a particular link or to reformulate.

Transaction log analysis studies are focused primarily on the system, yet they aspire to shed light on the behavioral aspects of the search. Studies that incorporate transaction logs analysis usually examine the characteristics of search episodes in order to isolate trends and identify typical interactions between the searchers and the system. In this context, interaction has several meanings and addresses a variety of transactions including query submission, query modification, viewing of search results, and use of information objects (e.g., web page, pdf file, video, etc.). Different types of analysis may be applied to transaction log data. The analysis may be on a query term level (Silverstein et al., 1999) measuring factors like term occurrence (frequency of that term), total terms in the dataset, unique terms, high usage terms, and term co-occurrence (the occurrence of term pairs). Another type of analysis may focus on a query level (Jansen et al., 2009a) examining query modification, query repetition, query complexity (query syntax, including the use of advanced searching techniques such as Boolean and other query operators), and failure rate (a measure of the deviation from the published rules of the search engine) (Jansen, 2006).

Transaction log analysis studies require numerous assumptions about the searchers' intentions and their decision-making process, yet the volume of the analyzed data allows drawing statistically significant conclusions about behavioral patterns of the users. For example, Rieh and Xie (2006) used Excite web search engine log analysis to discover the types of query reformulations performed by the users. They found three facets of query reformulation (content, format, and resource) as well as nine sub-facets of modifications that were derived from the data such as specified and generalized modifications. Silverstein et al. (1999) also performed analysis

on the query level and presented their analysis of individual queries, query duplication, and query sessions from the Alta Vista search engine log. They also explored correlations on the term level in order to study the interaction of terms within queries. The results showed that web users type in short queries, mostly look at the first 10 results only, and seldom modify the query. Jansen and Spink (2006) looked at multiple search engines and examined characteristics and changes in web searching from nine studies of five different web search engines. Their goal was to identify trends and differences in the number of one-query sessions, number of one-term queries, number of results pages viewed, and differences in search topics. The analysis helped the authors identify trends regarding query length and reveal that it's not increasing significantly over time, the same as single-term queries and use of query operators, which were found to be steady. This is in contrast to their previous research (Jansen, Spink, & Pedersen, 2005) on AltaVista only, which showed that query length moved slowly upwards[5]. In terms of viewing of SERP, Jansen and Spink (2006) did notice that the viewing of only the first page of the results is extremely high and it significantly increased over time on the Excite web search engine.

As mentioned above, the main strength of transaction log analysis is the large N. This method allows analyzing large datasets based on large numbers of users and search instances. In addition, this method is unobtrusive, inexpensive, and is collected during search in the searcher's natural settings, as opposed to other approaches that require the searcher to act in the artificial conditions of a lab. On the other hand, this method is limited in terms of types of data that appear on the log. For example, logs are usually missing information such as users' identities. In cases when there is an identifier present so that the search instances can be separated, there is still no

[5] The differences between the conclusions of the two studies could be due to a different method of measuring query length or possibly be attributed to differences between AltaVista when taken separately and when analyzed together as a group with multiple search engines.

record of the demographic data of the users[6]. More importantly, a transaction log does not record the users' motivation or intention and does not track their cognitive processes, such as reasons for the search, the decision-making process when it comes to selecting a search result, and more. This is why many times logs are used in conjunction with other types of data sources (Jansen, 2006).

As mentioned earlier, transaction log analysis enables researchers to gather information on *what* the searchers did, in which order, and the timing of their actions. However, this data is not sufficient for understanding the motivation and the reasoning behind the actions of the searchers. Therefore, transaction logs analysis, although an important method of understanding searcher's behavior is not well suited for investigating the decision-making process of query reformulation and interaction with the SERP.

2.4 Query reformulation categories

The study of the query reformulation process typically addresses classification of the logged history of query reformulation steps in order to identify search strategies employed by the searcher and predict future query reformulation categories or steps. In order to analyze patterns of query reformulation and develop predictive models, researchers often employ query logs or observations, basing their analysis mainly on the content of the query (e.g. Fidel, 1985; Jansen et al., 2009a). Since the focus is on the transition from one query to the next, the common question pursued in this type of studies is: "What was the transition from the previous query to the current one?" or "To what general group, does the transition from the previous query to the current one

[6] Commercial search engines are capable of collecting this information from its users, but usually this data is proprietary and is accessible only to the company and not to outside researchers.

belong?" The main practical purpose of these studies is to find a way to assist the searcher in her reformulation attempts by suggesting single query terms or full queries.

Fidel (1985) was one of the first to study the query reformulation process from the aspect of content and meaning. She performed observations on searchers and confirmed previous research that query reformulations are made when retrieved sets are too large, too small, or off-target. Based on the results of her observations, Fidel divided the query steps into operational (query modifications that do not change the meaning of the query) and conceptual steps (query modifications that change the meaning of the query). Other researchers, based on log analysis and experiments, unpacked Fidel's classification into more nuanced categories such as: specification, generalization, replacement with synonyms, parallel movement, and term variations (Liu et al., 2010; Rieh & Xie, 2006).

One of the purposes of the studies mentioned above was to improve automatic classification of queries and subsequently, to be able to predict the searcher's next step. For example, Lau and Horvitz (1999) generated a list of query reformulation categories in order to develop an algorithm for automatic classification. These categories included: generalization (removing query terms), specification (adding query terms), new query, reformulation (term replacement), and interruption (query on one topic is interrupted by a search on another topic). Lau and Horvitz based their automatic classification on the change of query content and query length, through modeling the probabilistic relationships among temporal activity patterns, informational goals, and query reformulation categories. Their model was based on Bayesian networks, which incorporated inter-query intervals and adjacent actions, as well as took into consideration the searcher's informational goals. Lau and Horovitz's taxonomy was later used by He, Göker, and Harper (2002) as well as by Jansen and his colleagues (2009a, 2009b) for query reformulation

prediction purposes. Huang and Efthimiadis (2009) have further extended Lau and Horovitz's taxonomy by identifying eight more query reformulation types: remove words (same as generalization), add words (same as specialization), and word substitution (same as reformulation) they also detected other types including word reorder, stemming, abbreviation, acronym (both create and expand), and others. Categorization schemes, such as these, as well as the automatic classification algorithms are the main outcomes of studies that investigate query reformulation.

Since this study aims to understand the link between the cues searchers pay attention to and their query reformulations, there is a need to be able to map the actual reformulation into categories. Some of the reformulation categories presented in this section were elicited for prediction purposes, which are equally important, but they can also be used as part of the model to be developed in this study. These categories are useful as a starting point in content analysis of query reformulation types from the data collected in the study.

2.5 Relevance in query reformulation

Relevance has been acknowledged as one of the key notions in information science in general and specifically in the field of information retrieval (Saracevic, 2007b). This explains the high volume of work about the definition of relevance and its different aspects. Saracevic has a series of articles ranging from systematizing the debate on relevance (1975), through reconsideration of some of the relevance-related issues (1996b), to a more recent review and synthesis of the literature dedicated to understanding relevance and its manifestations (Saracevic, 2007a, 2007b). In this study, relevance is an important topic because it is directly tied to query reformulation. Relevance can serve either as a trigger for performing reformulation or it can be inferred based on the query reformulations. Since this study will be focusing on the cues that searchers pay

attention to when reformulating their queries, exploring prior literature on relevance is potentially useful because the cues that affect the searcher's judgment regarding relevance may be similar to the cues that influence query reformulation.

The ultimate goal of defining and understanding relevance is to measure and improve retrieval effectiveness and to improve our understanding of the retrieval process. More specifically, relevance feedback provided by the users has been employed to perform query expansion, term disambiguation, user profiling, filtering and personalization (Kelly, 2005). In this section I am particularly interested in discussing the role of relevance as a means for query improvement by the searcher. In this context, the notion of relevance intersects with the process of query reformulation when assessment of relevance of one set of retrieved results influences the subsequent query. In IR literature, this technique is referred to as a form of 'relevance feedback' (RF).

In the context of query reformulation, RF aims at improving the query by adding (or up-weighting) terms from documents that have been retrieved and assessed as relevant by searchers (manual relevance assessments) or by an algorithm (automatic/pseudo RF) (Saracevic, 2007b). The pseudo RF algorithm retrieves the documents automatically and assumes that the top-ranked documents are the most relevant ones and uses those to extract terms to be added to the query or to be up-weighted. Negative relevance feedback can also be used in a similar way to identify terms that would be assigned a lower weight when the query is executed. The only role that the searcher has in this interaction is to indicate relevance or non-relevance of a retrieved document. The query reformulation process takes place internally in the system, and usually it is opaque to the searcher, as her only knowledge of that action is through the list of documents retrieved as a result of the reformulated query (Belkin, 2000). Typically, studies that involve RF compare the

performance of a system which uses RF to improve queries and one that does not (Koenemann & Belkin, 1996; Ruthven, Lalmas, & Van Rijsbergen, 2003). Even though the process that uses RF is a bit different than query reformulation performed by a searcher, because the revised query is usually not visible to the searcher, there is still some similarity, since RF influences the way a query is to be changed.

Spink and Saracevic (1998) conducted a study that addressed user query reformulation as a result of relevance evaluation more explicitly. The study investigated mediated online searching, where academic users with real information problems provided a question to search online. Their searches were performed by professional search intermediaries. The authors investigated user's feedback and classified it into different types of categories. The 'content relevance feedback' category represented a situation in which the retrieved items were judged by the user for relevance and followed by a new query on a new topic (if the feedback was positive) or reformulation (if the feedback was negative). The second category, 'term relevance feedback', represented the cases in which the subsequent query included new search terms from the retrieved output. Another category, 'magnitude feedback', encompassed cases where the following query was affected by assessment of the size of the output from the current query. Finally, the 'tactical review feedback' category represented a situation in which the user decided to display the search strategy history (provided by the functionality of the system) which influenced her decision regarding the subsequent query.

The categories elicited by Spink and Saracevic (1998) provided an insight into the decision-making process, showing how it relies on relevance and how it affects the reformulation of queries. One of the things one may learn from this is that when the searcher finds the retrieved documents not relevant, the result would typically be query reformulation (this may seem

obvious, but even obvious things need to be confirmed by empirical research). In addition, the second category shows that searchers use terms from retrieved documents in subsequent queries. Even though this is a step towards a better understanding of the decision making that goes on when the searcher performs query reformulations based on previous system output, the categories elicited in this study are too general and do not represent the reasoning behind them. For example, if 'term relevance feedback' is identified, this means that the retrieved results influenced the searcher's choice of terms in the subsequent query, but there is no way to know what led to the choice of these specific terms as candidates for the next query.

The influence of relevance on query reformulation wasn't in the scope of the 'relevance clues' concept as it was defined by Saracevic (2007b). The author synthesized sixteen studies that explored relevance clues and, according to his definition, he focused on the research that "aims to uncover and classify attributes or criteria that users concentrate on while making relevance inferences. The focus is on criteria users employ while contemplating what is or is not relevant, and to what degree it may be relevant (p. 2127)." In my opinion, the criteria (relevance clues) that people use when making a decision regarding relevance, also have a role in influencing their decision on how to reformulate the query.

Saracevic (2007b) found that the synthesized studies observed a similar set of relevance clues. However, the searchers assigned different importance to the given clues depending on the task, progress in task over time, and class of users (for instance, children don't attribute any importance to authority, while faculty members do). The author arrived at a set of classes of relevance clues based on the generalized results of the synthesized studies. The classes of relevance clues included: content (topic, quality, depth, scope, currency, treatment, and clarity), object (characteristics of the document, such as representation, availability, costs, etc.), validity

(accuracy, authority, verifiability), situational match (appropriateness to situation, usability, urgency), cognitive match (novelty, mental effort), affective match (emotional responses, frustration), and belief match (e.g. confidence). It should be noted that some of these clues overlap with the manifestations of relevance mentioned above (e.g. affective match is equivalent to affective relevance).

Barry (1998) also examined clues for relevance and examined *why* certain document representations enable searchers to predict document relevance better. Her goal was to identify the extent to which various document representations contain clues that allow users to determine the presence or absence of traits that establish the relevance of the document to the user's information need. She interviewed users who discussed their reasons for pursuing or not pursuing documents based on information contained within representations of those documents (i.e. titles, abstracts, indexing terms, source traits, etc.). Barry concluded that the utility of the clues contained within document representations may depend less on the document's representation itself, but rather on the user's context, both in terms of the user's previous knowledge and the specific qualities the user seeks. This means that a user seeking a specific quality of information, may interpret any document representation in terms of that quality, regardless of whether the representation makes any explicit statements about that quality or not. The author speculated that any information about a document may prove useful to some user in some situations. Yet, Barry's study also showed that it is probable that some clues are inherently tied to certain types of document representations. In other words, a user's ability to determine the presence of certain qualities does depend on the document representations that are made available. The author noted that the next step would be to identify very specific clues that are utilized by users to predict some aspects of documents. In a way, the research described in this

document was set out to explore these specific clues, only in the context of query reformulation and not explicitly relevance.

Another study that performed a similar kind of research on relevance is by Park (1993). He presented a model that reflects the nature of the thought processes of searchers who are evaluating bibliographic citations produced by a document retrieval system. Park investigated how the searcher perceives the value of each element of a citation in relation to her need. The author looked at elements such as title, style of the title, author name, journal name and document type, and the abstract. Park found that relevance is not fixed but is a temporal and fluid concept that is influenced by factors such as: perceptions about search quality, search goal, and the anticipated end product of the research for which the search is being performed. Variables related to the information itself, such as scarcity, availability, timeliness, and scope, were also found to be affecting factors. If these are the properties of relevance, it would be interesting to see if then the cues investigated in this research also exhibit similar properties related to relevance.

Quiroga and Mostafa (2002) attempted to understand the various elements of the rationale of searchers providing relevance feedback on a document. They identified the characteristics of the searcher (demographic, domain expertise, lifestyle) and characteristics of the documents (orientation facets, specificity level, combining topics, credibility, novelty, format, and availability) that influenced what drew the searcher's attention. These elements are not as fine grained as the cues that the research described in this proposal will explore, but they can provide a foundation when analyzing the data and looking for the elements that contribute to the process of evaluating search results.

There is one more point where query reformulation and relevance overlap. Introducing the 'human factor' into the process of RF creates additional challenges because it requires the searchers to provide explicit feedback to the system either by specifying keywords, selecting or evaluating documents, or answering questions about their interests (Kelly, 2005). In tackling this problem, a number of studies looked for ways for implicitly inferring relevance, based on the searcher's behavior and her interaction with the search results. In this user-centered type of research, relevance evaluation is addressed as "user preference" (Agichtein, Brill, Dumais, et al., 2006) or as "usefulness" of the documents (Kelly & Belkin, 2004). This approach is similar to Saracevic's (1996a, 1996b) affective relevance, which uses satisfaction, success, and accomplishment as criteria for assessment of the searcher's motivational relevance. In fact, one can find references to the complexity in assessing relevance from a practical perspective in the earlier writing of Saracevic (1975), when he describes the pragmatic view of relevance, which considers the relation between the searcher's immediate problem and the characteristics of a document, including concepts such as utility and preference as the basis for interference. However, describing what constitutes relevance is not always sufficient and often there is a need to infer based on searcher's behavior, whether or not the searcher deems a certain document as relevant. Access to transaction logs from web-search engines allows this kind of inference.

There are numerous indicators that can be used to infer relevance. Kelly (2005) reviewed several studies that explored how behavior can be used as implicit relevance feedback. Among the various indicators, actions such as viewing, listening, scrolling, finding, querying, selecting (click-through), and browsing were explored. Some have proved to not be very indicative of relevance. For example, under the view indicator, Kelly and Belkin (2004) found that there was no direct relationship between display time and usefulness. The authors discovered that display

times differed significantly across various tasks and across different searchers. Similarly, selection (or as it's also called, click-through) was also found to be a weaker form of implicit feedback than reading time or scrolling, since it does not distinguish between useful and non-useful web pages requested by the user. Kelly (2005) also pointed out that other implicit indicators such as saving, printing, and book marking may be useful, but are more difficult to collect; such information must be gathered from the client machine or extensive work must be performed on the server side in order to collect the data.

Other than being able to distinguish between useful and non-useful web pages, there is also a need to overcome the "trust" bias, which is the searchers' inclination to click on higher ranked results compared to lower ranked results, regardless of the actual relevance of the document. Joachims et al. (2007) encountered this bias when examining the reliability of implicit feedback generated from click-through data and query reformulations in web search. The authors used eye-tracking in order to understand the extent to which clicks are a result of an informed decision. They also compared the implicit feedback from click-through to manual relevance judgments by the users. They found that clicks are informative but biased. In addition to the "trust" bias, the study revealed that the clicking decisions were influenced not only by the relevance of the particular link, but also by the overall quality of the other snippets on the results page. The main conclusion of this study highlighted that clicks have to be interpreted in relation to the order of presentation and in relation to the other snippets.

Agichtein, Brill, Dumais et al. (2006) also used click-through data in their study as indication of relevance, but in addition they also employed query and browsing features. The purpose of their technique was to automatically predict relevance preferences for web search results based on various features elicited from the interaction with retrieved search results. The authors used three

types of features: query text features (such as fraction of shared words between query and summary, fraction of words shared with next query), browsing features (such as average dwell time on a page, number of hops to reach page from query), and click-through features (e.g. position of URL in current ranking). The model Agichtein et al. built employed query reformulation in order to learn about the searcher's interaction with the results and from this, implicitly infer relevance preferences, but not the other way around. In other words, the objective is not to understand how relevance is used for reformulation, but rather how relevance could be predicted from query reformulation and other features.

The clues gathered by Saracevic (2007b) and mentioned earlier are driving the searchers' relevance assessment, but it is possible that these clues (whether in this format or more specific ones), may also serve as clues for the purpose of query reformulation. For instance, if the document is not relevant due to the existence of the "topic clue" category (from the list of classes of relevance clues Saracevic arrived at in his research), then the searcher would probably reformulate the query to adjust the topic. Another example would be if the document is not relevant because it appears on an untrustworthy website (validity clue), the searcher could reformulate the query to exclude results from this website. These clues could also be broken down to clues with more detail in order to determine what specific clues the searchers used that resulted in a decision regarding the more general cues and eventually resulted in reformulation of the query. For example, what specific clues from the content of the document the searcher used in order to conclude that the topic of the document was not relevant? In other words, what elements in the document, with the combination of the query, did the searcher pay attention to in order to decide that the document was on a topic that does not overlap with the topic of the query? It is difficult to assess at this point whether the clues investigated by Saracevic work for

reformulation in the same way they work for relevance judgment, but this study is intended to shed some light on this question.

2.6 The role of tasks in user-centered information retrieval studies

There is a long-standing consensus that the attributes of a task utilized in a user study are a major determinant of human decision making (Einhorn & Hogarth, 1981) and can have a significant impact on the results of the study. More than twenty five years ago, when discussing the methodological issues in experimental information systems research, Jarvenpaa et al. (1985) stated that the use of diverse and often unrelated and incomparable tasks makes the integration of findings across studies difficult, because differences in participants' performance result more from the task than from the use of the system.

The choice of experimental tasks is directly related to the ability to ensure internal validity of the experiments (Jarvenpaa et al., 1985). If the effects of the tasks are not isolated, then there is no way of knowing whether what is being measured is indeed a result of the manipulation of the independent variable or it is endogenous to the task. In much the same way that being aware of the effects of tasks for information systems evaluation is important (Jarvenpaa et al., 1985), it is essential to be familiar with the effects of search tasks on searchers' behavior. Other than making sure that the effects of search tasks are being controlled for, being aware of these effects can help manage the drawbacks that may arise under different conditions. This sub-section deals with the effects of various types of tasks, with regard to the research questions at hand.

2.6.1 Understanding tasks

Search task attributes and their influences have been studied both with respect to realistic, self generated tasks, and also in relation to tasks that are assigned by the researchers in a user study.

While naturalistic studies are meant to observe searchers performing genuine tasks in their natural settings, experimental studies aim to isolate particular effects on user behaviors and provide an opportunity to compare across subjects and across studies that use the same tasks. Therefore, assigning search tasks by researchers represents an attempt to control for variables related to search tasks in experimental studies. This type of user study is usually guided by the search tasks, either in order to control for the effect of the task by assigning the same tasks to all the subjects or in order to be able to manipulate it as an independent variable. Whether the studies have used control or manipulation of search task, in either case, researchers are hindered by the lack of understanding of how the search task influences study findings (Wildemuth & Freund, 2009).

Since search tasks can vary along many dimensions, the main disadvantage of assigned search tasks lies in generalizability (external validity). In other words, findings of a user study with a set of assigned tasks may be valid for these particular tasks and context, but it is difficult to determine whether the same effect would be achieved in different experimental or natural settings (Wildemuth & Freund, 2009). On the other hand, when the effects of the different search task types are known, controlling for the search task allows isolating specific relationships that the researcher is interested in studying. Depending on the research question, the researchers would often want to maintain some desirable effects throughout the experiment, so that they can be studied. Knowing that a particular type of assigned task may create specific effects and being able to predict how the tasks would influence the dependent variable, is highly important. For example, if multiple query reformulations during the course of the experiment are to be studied and previous research shows that a difficult task may lead to multiple reformulations, then a

researcher might choose to assign difficult tasks to ensure that multiple query reformulations occur.

Designing tasks that are both realistic and that allow a certain level of control by the researcher is the subject of on-going research. Borlund and her colleagues have been investigating for over a decade what constitutes a well-designed search task. Borlund and Ingwersen (1997) came up with the term "simulated work task situation" as an equivalent for assigned search tasks and provided guidelines that could assist in creating a well-designed task. Simulated work task situations are scenarios in which participants are asked to envision they have a particular task from which to derive their own information needs. In their recent study, Borlund and Schneider (2010) claim that the major challenge of this process lies in designing both authentic and applicable simulation work task situations, which would be relevant and realistic to the study of searchers who would apply these tasks in experiments.

Borlund has shown in previous research (2000a, 2000b) that a well-designed task will possess certain attributes such as: being tailored to the group of participants, one that the participants can relate to, in which they can identify themselves, and also find it topically interesting. Borlund and Schneider (2010) have also recommended employing a combination of simulated work task situations with the participants' genuine information needs and by doing so enjoying the benefits of both worlds – those of assigned and those of searcher generated tasks. At the same time, one universal disadvantage of an assigned task (that does not depend on the research question) may stem from the searchers' lack of familiarity with the topic of the assigned task. For example, Wen et al. (2006), who looked at the effect of topic familiarity on the assessment behavior of online searchers found that for unfamiliar tasks, relevancy criteria such as depth/scope, or accuracy could not be easily employed by the searchers. In this case, the researcher needs to be

able to control for familiarity and either record it (their level of domain knowledge) for each participant or recruit participants who are familiar with the topic.

2.6.2 Task attributes

It has been shown that attributes of a search task can have a wide range of impacts on search performance and behavior. The way different attributes of a task such as complexity, difficulty, or breadth affect search behavior, which is manifested through usefulness, relevance feedback, eye movements, query reformulations, and more, have been studied extensively (e.g. Byström & Järvelin, 1995; Cole et al., 2010; Kelly & Belkin, 2004; Liu et al., 2010; Smith, 2009; White, Ruthven, & Jose, 2005). For instance, Saracevic and Kantor (1988) found that broad search tasks led to higher precision of retrieval as opposed to specific tasks which led to lower precision. A more recent study (White et al., 2005), revealed that when the search task was more complex (varied by the number of potential information sources and types of information required, to complete a task), users rarely found results they regarded as completely relevant and struggled to find relevant information. As a result, the users were unable to communicate relevance feedback to the search system. Judging by these results, while a struggling user could be an interesting subject of inquiry for one type of study, this situation may be rather useless for a study which requires relevance feedback from the user. Therefore, search tasks should be chosen with care and their attributes should be taken into consideration.

Based on previous research, some studies have made attempts to create typologies or classifications of task attributes. For example, Wildemuth & Freund (2009) used previous literature to introduce a typology, which consists of the following pairs: complex vs. simple, specific vs. general, exploratory vs. lookup, and navigational vs. informational. Bilal (2002) also introduced a taxonomy of tasks which includes three levels: task type (e.g. open ended and

closed), task nature (e.g. complex and simple), and task administration (e.g. fully self generated, semi assigned, and fully assigned). Toms et al. (2008) had a different characterization of task type, which was split into fact finding, information gathering, and decision making. Another attribute that they took into consideration was task structure (parallel and hierarchical). Toms et al. (2008) had also defined principles which guided the creation of their tasks, such as that no search could be answered in a single page and that the tasks should have semantic content that requires interpretation. Li and Belkin (2008) have developed a faceted classification scheme, which is aimed at identifying different aspects of a task (work task and search task) by defining categories, facets, sub-facets, and values. The scheme is comprised of a generic facet of a task (source of task, task doer, time, product, process) and common attributes of a task, which are sub-divided to task characteristics (objective complexity, interdependence) and user's perception of the task (salience of a task, urgency, difficulty, subjective task complexity, knowledge of task topic, knowledge of task procedure). At the same time, Xie's (2009) classification is comprised of dimensions of a work task (nature, stages, and time frame) and a search task (origination, types, and flexibility). There is apparent agreement in the field, that task definition matters and task attributes have inherent impacts on search behavior. At the same time, there is no agreed or dominant classification of task attributes that would allow far reaching claims about uniform task influences across the board. This places further burden on the external validity claims in user-centered information retrieval research.

In addition to the lack of consistency across the attempts to classify task attributes and their impact, another layer of complexity is introduced through inconsistency in the definitions of the attributes themselves. Even when the titles of the attributes in various typologies match, their descriptions differ or some concepts still require detailed explanations regarding the meaning

behind each task definition. For example, as Wildemuth and Freund (2009) pointed out, complexity, which is among the more common attributes of search tasks, has been operationalized and defined in many different ways. In various studies, this attribute tends to be defined as a collection of one or more of the following task dimensions: structure (Lazonder, Biemans, & Wopereis, 2000; Sharit, Hernández, Czaja, & Pirolli, 2008), the number of search concepts involved (Saracevic, Kantor, Chamis, & Trivison, 1988b), the number of paths involved while engaging in the task (Li & Belkin, 2008), certainty or *a priori* determinability (Bell & Ruthven, 2004; Browne, Pitts, & Wetherbe, 2007; Byström, 2002; Campbell, 1988), number of facets (Bilal, 2001; Capra, Marchionini, Oh, Stutzman, & Zhang, 2007), length of the search path (Capra et al., 2007; Cole et al., 2010; Gwizdka & Spence, 2006), cognitive effort (Bilal, 2001; Capra et al., 2007), and topic familiarity (Bilal, 2001; Browne et al., 2007).

Other task attributes are more straight-forward and some are used across studies. An example of such an attribute is informational vs. navigational. Broder (2002) was the first to make this distinction stating that while the purpose of an informational task is to "acquire some information assumed to be present on one or more web pages" (p.5), the purpose of a navigational task is "to reach a particular site" (p. 5). Another popular attribute is specific vs. general task. Across various studies where "specific" tasks usually have more clearly defined goals than "general" tasks and this definition has been applied across studies. Specific tasks may be equated with known-item search tasks, factual tasks, or simple lookup tasks (Wildemuth & Freund, 2009). Also Marchionini's (1989) definition of "open" (has multiple answers) and "closed" (has one answer) can also be categorized under the specific and general category where open is the more general one and closed is more specific. In terms of exploratory vs. lookup tasks, exploratory searching is defined as searching that supports learning and investigating, while lookup tasks, are

geared toward finding particular facts or answering specific questions (Wildemuth & Freund, 2009).

2.6.3 Varying utility of tasks

How to take the effects of a certain task attribute into consideration depends on the research question - what the researchers are looking to investigate in a particular study. For example, if a researcher investigates the query reformulation process and could benefit from numerous reformulations, then a task that generates many reformulations will be more suitable for her. A specific example could be observed in Toms et al. (2008) who found that users formulated fewer queries for hierarchical tasks. In terms of task type, the study showed that decision making and fact finding tasks contained more queries than information gathering tasks. With this knowledge, a researcher investigating query reformulations can assign parallel tasks and prefer decision making and fact finding tasks over information gathering tasks.

One characteristic of assigned tasks to take into account is its influence on the searchers' choice of query terms. If a researcher is interested in ensuring that the queries are user-formulated and contain the least terms copied from the task, some of Toms el al. (2008) findings may be of use. For example, they found that in the decision-making tasks, users employed the least amount of self-formulated terms and relied on terms provided in task definition. At the same time, in the information gathering tasks the participants used more self-formulated terms and relied less on the language of the task definition provided to them. Therefore, a researcher who wishes to study user-formulated queries, which include more users' terms and fewer terms borrowed from the task, could apply these findings when creating search tasks to be used in their study.

Additional utility of a task also depends on its effect on search behavior. A set of studies of this nature was conducted by a group of researchers from Rutgers who've conducted research on how

task attributes affect search behavior. Each study incorporated different types of tasks (or tasks with different attributes). For instance, Cole et al. (2010) investigated how transitions between scanning and reading behavior in eye movement patterns are an implicit indicator of the current search task that the user is dealing with. The authors found that for copyediting tasks most participants made the decision to switch from scanning to reading less frequently. When the users were reading they more frequently decided to switch back to scanning. Opposite behavior was observed for advanced obituary task and interview preparation task - both had similar impact on the reading models. On the general level Cole et al. hypothesized that "the product and level task characteristics had more influence on the tendency to decide to switch from scanning to reading and to switch to a new reading sequence rather than scan after the end of a reading sequence." This could be useful for eye tracking research of post-search behavior, in which the researchers want to know in advance what kind of eye activity will be taking place.

In another paper by the same group of authors, C. Liu et al. (2010) used the same three types of tasks and categorized query reformulations into groups such as specialization, generalization, substitution, and generation of a new query. Their results showed that specialization was more frequently used for simple and hierarchical tasks than in parallel tasks, and that word substitution was more frequently used in parallel tasks than in simple and hierarchical tasks. Future studies may utilize these findings with respect to their research goals. If researchers are interested in studying word substitutions (for example), they could benefit from assigning parallel tasks, because this is the type of tasks that would generate more word substitutions.

Gwizdka & Spence (2006) showed that perceived task difficulty may also influence users' behavior when faced with a task that is subjectively difficult. The authors studied subjectively perceived post-task difficulty and objective task complexity in factual information-seeking tasks

and administered search tasks of varying complexity. This study's original goal was to find a way to predict user-generated tasks based on search behavior, but findings from this study could also be applied when deciding on the difficulty of assigned tasks as they are perceived by users. Subjective task difficulty was found to be correlated with many measures that characterize the user's actions. Higher search effort, lower navigational speed and lower search efficiency were found to be good predictors of subjective post-task difficulty assessment. The study showed that task complexity was found to affect the relative importance of the predictors of subjective task difficulty. These findings can shed light on any process that includes taking into consideration lower navigational speed of users when faced with a task that they perceive as a difficult one. This study also emphasizes the importance of pilot studies or at least some kind of evaluation of tasks by users similar to the population of the study.

As can be seen for the studies described above, while there are many different task attributes, the purpose of task selection is usually the same and that is to be able to artificially control the searchers' information needs in order to isolate task effects.

2.6.4 Other considerations

On some occasions, task type will interact with other factors and being aware of these interactions could be helpful for the purpose of controlling these factors. For example, Ford et al. (2005) found that the influence of task complexity depends on individual differences. They demonstrated that as task complexity increases, searchers who are characterized by navigational disorientation (feeling that they tend to get lost) and a somewhat unplanned approach (adopt a strategy of being willing to sift through irrelevant material) migrate to Boolean search as a result. On the other hand, they also found that searchers displaying the exact opposite of these attributes migrate to best-match strategy. As for the least complex task, they may choose either strategy—

with equal success in terms of finding useful information. This finding is similar to earlier findings of K.S. Kim (2001) who reached the conclusion that online search experience (novice vs. experienced searchers) interacted with task type (known-item vs. subject search tasks) to influence navigational style and the number of results visited. Another study (K. S. Kim & Allen, 2002) on the interaction of task types with problem solving style revealed that users who possess an inefficient problem solving style would need help, especially when carrying out general or ill-structured search tasks that require a higher level of planning and problem-solving than specific tasks. This shows that sometimes controlling the task only is not enough and other variables (such as novice vs. experienced searchers) need to be controlled or taken into account as well.

Eye tracking studies have also been employed for the purpose of determining the effects of different task types. For example, Lorigo et al. (2006) found that for informational searches, the exploration of the retrieved web page is a critical part of the search process, and the results page is only an intermediary for that type of search. On the other hand, their results also showed that a greater proportion of time is spent on Google result pages for navigational tasks since navigational questions do not require much additional inspection of web documents outside of Google. At the same time the results implied that the process of searching within the query results abstracts may yield similar levels of cognitive arousal for the two tasks (informational and navigational). Given these results, a researcher can choose the type of task according to the purpose of the study. If the study seeks to investigate mainly the interaction of the user with the search results page and not necessarily the exploration of the contents of retrieved web pages, then the attributes of navigational tasks mentioned above would constitute an advantage for this type of inquiry.

In their eye tracking study, Guan and Cutrell (2007) manipulated the position of the results in order to see how users interact with the results under these conditions. They found that when targets were placed relatively low in the first page of search results, users spent more time searching and were less successful in finding the target, especially for informational tasks. Their analysis of eye movements showed that the decrease in search performance was partially due to the fact that users rarely looked at results which were ranked lower. In contrast to navigational tasks, where the target is more obvious from information presented in the title and snippets, in informational tasks, users try the top ranked results even if these results are perceived as less relevant for the task. From this, one can learn that for informational tasks searchers, who cannot get enough information from the snippet and the title, would often go into the retrieved websites.

In summary, there is an apparent consensus that search tasks influence searcher behavior. There are various factors to take into consideration when choosing search tasks. Tasks can introduce external validity threats, influence the searchers' choice of query terms, or enact a dissonance between the task and the searcher's task-related knowledge. In addition, there is no agreed classification of task attributes that would allow far reaching claims about uniform task influences across the board. Instead, each set of task types has its own influences, which would be relevant or not, depending on the research question at hand. For example, if my research focuses on query reformulation and the searcher's ability to evaluate the results, it would be essential that the tasks that I choose yield multiple queries and that the searcher has at least a minimal familiarity with the topic of the tasks. Therefore, since there is no "magic" formula for picking the most suitable tasks, they should be selected according to the conditions set by the goals of the study and if possible, reused from other studies that had similar goals. Even if the

tasks have been used in previous studies, this does not cancel the necessity of performing a pilot and testing them out.

2.7 Conclusion from the literature review

The literature reviewed above addressed different aspects of searcher's interaction with a search engine as they may be relevant for this study which deals with query reformulation and what triggers it. These issues have been researched from different angles, often employing the data or methodology that would be the most efficient and that can be automatically manipulated. However, while utilizing such data or methodology can be convenient, it is more suitable for research questions that involve the questions of *how* or *what* actions were performed by the user as part of the interaction and not *why* it was performed. The "*why"* question has been rarely addressed in the literature presented in this section, especially with regard to query reformulation and interaction with SERPs, and this is a gap that needs to be filled with more in-depth observation. While eye tracking may be useful for collecting implicit feedback regarding searcher's evaluation of the results, this kind of research is still lacking, because it cannot indicate what exactly triggered the reformulation or be precise enough to actually identify the cues without questioning the searcher. In addition, literature shows the importance of a careful choice of tasks which are suitable for the research questions and the goal of the research.

The study described in this document attempts to tackle the *why* question by exploring the cues associated with different types of query reformulation, something which to my knowledge hasn't been addressed in previous research.

3 Theoretical Framework

This section attempts to demonstrate that both in pre-web and web search engines, searchers have been utilizing some form of cues when deciding how to reformulate their query. It also shows how domain knowledge or search expertise influences query (re)formulation and therefore might also affect cue elicitation and usage. It is difficult to predict the exact influence these two parameters will have on cue elicitation and usage, but as long as some effect exists, domain knowledge and search expertise, should be taken into consideration in this study.

3.1 The existence of cues

This section will discuss current models or frameworks that describe search behavior that may incorporate cues. In early search literature, one of the models closely related to iterative cue utilization is Oddy's dialogue structure. Oddy (1977) modeled his system, THOMAS, based on a dialogue structure in reference retrieval. In the system, this type of dialogue between the system and the searcher formed an image of the searcher's interest, displayed references, associated authors, and subject terms according to the state of the image. Then the system modified the image, based on the searcher's responses to the displayed elements: references, authors, and terms. In THOMAS, the searcher reacted to the displayed elements in the result by indicating yes and/or no to the elements (reference as well as the terms and authors displayed). The rejected reference (the ones that got a 'no'), authors, or terms were then removed from the image. The accepted reference and the authors or terms were then added to the image.

This dialogue represents an iterative process of the searcher collecting bits and pieces from the displayed references, authors, and terms and the system using these pieces to construct an image of the searcher's interest in relation to the presented results. The searcher's rejection or

acceptance represents the cues gathered from each iteration in the dialogue, as new pieces of information. Same as acceptance or rejection in THOMAS shapes the image, the cues help shape the searchers' understanding of the results and affect the subsequent queries. In other words, the building of the image in THOMAS is a mental process which is similar to the process of cue gathering for the purpose of query reformulation, because it forces the searcher to interpret the results in terms of the wanted and unwanted elements in it and use this information to modify the query. In the current study, I chose to focus on these new pieces of information which affect the image and are comparable to cues the searchers gather from elements presented on the SERP (e.g. snippet of a result, its title, terms.) or from the target pages (full web pages that the search results lead to).

The main challenge in constructing a model of the searchers' decision making process when presented with a list of search results is to describe the nuanced relationships between the elements of the presentation and the eventual decision. Spink and Saracevic (1998) have provided one of the most useful foundations for mapping the types of cues that may affect subsequent query reformulation. The authors offer five situations or conditions in which the query may be influenced and identify them according to the type of interactive feedback they generate between the searcher and the system. Four of these conditions are relevant to web search[7]:

1) *Term relevance* represents a situation in which the subsequent query includes new search terms from the retrieved results;

[7] The original study identified another type of situation, "term review", which is user input followed by a strategy related judgment to display terms in the inverted file influencing the subsequent query. It is not presented here because it is irrelevant to the current web search engines.

2) *Magnitude* represents cases when the size of the output from a query affects the next query reformulation;
3) *Tactical review*[8] represents a situation in which the user reviews her search strategy history for the purpose of subsequent query reformulation;
4) *Content relevance*[9] represents a condition where the user judges the output for relevance, which is then followed by a query on a new topic (if the output was judged relevant) or reformulation (if the output was judged as irrelevant).

Spink and Saracevic's typology suggests that the searcher's response to system output is expressed in subsequent queries. These conditions are comparable to cues, since they are reviewed by the user and then followed by a subsequent query reformulation.

Another useful study which dealt with classification of cues is O'Day and Jeffries' (1993) inquiry into how the search results are digested and used to solve the problem underlying the initial motivation to search. The authors identified "triggers" and "stop conditions" that guided users' search behaviors. In their study, the triggers are the reasons for query reformulation. They were classified into the following four categories:

- There is a plan for what queries will be executed and the next step is part of this plan;
- Something interesting arose and prompted exploration;
- There was some change to be explained (for example a document with revenue increases was found and the next query was formulated to search for information that would help explain them);
- There was something missing from the data.

[8] This is similar to web-search in a way that searchers look at their query history by clicking on the "back" button in the browser or by using the search engine's option to view search history (where available).
[9] Theoretically, the outcome of this category should also contain the option of ending of search, but the authors did not include it in the description of the category.

Even though the idea of triggers was developed as part of a study on library users and intermediaries, and the users were mostly searching for business related information, they offer a useful construct and further support the logic of potential utility of cues in search-related decision-making. Placed against the background of the iterative nature of interaction with the search results (i.e. Oddy's model) the studies presented above identify categories of information that the searchers or users gather as they reformulate queries. These categories contain cues. The cues are usually part of the content of the retrieved documents, but can sometimes also be the characteristics of the retrieved result set (such as zero retrieved results) or visual (such as bolded query terms in the text). These are the cues that this research aims to identify and develop into a model. Previous research has not made a very fine grained distinction between the different types of cues that may be present in the retrieved text.

More recent studies specifically examine web-search engines in an attempt to understand the influence of search result features such as snippets, title, and URL (all together called 'caption') on web search behavior. However, this was performed without interviewing actual users but rather with automatic means and transaction logs. An example of such a study is Clarke et al. (2007) who explored what features of caption pairs, if any, lead users to prefer one caption over another. The authors used implicit feedback based on click-through data and extracted a number of features characterizing snippets such as: snippet missing in caption A and present in caption B, short snippet in caption A with long snippet in caption B, title of caption A contains matches to fewer query terms than the title & snippet in caption B, title & snippet &URL of caption A contains matches to fewer query terms than caption B, caption A URL is longer than caption B URL, and more. Their findings suggested that simple caption features such as the presence of all query terms, the readability of the snippet (readability can sometimes be negatively affected by

its query-dependent nature), and the length of the URL in the results (lengthy and complex URLs may have negative impact) can affect searchers' web search behavior.

Another study that was employed on the same commercial web search engine (Agichtein, Brill, & Dumais, 2006) incorporated user behavior data to improve ranking of top results in real web search setting. The data is collected from very detailed transaction logs. The authors built a model that includes features which represent user interactions with web search results that are divided into three groups: click-through features (such as: position of the URL in current ranking, probability of a click for each query and URL, whether or not the next or the previous position was clicked), browsing features (such as: if a link was followed, average time on page per query, number of hops to reach page from query), and query-text features (words shared between query and title, words shared between query and snippet, fraction of words shared with next query). This model uses data from overall user population to re-rank search results. The results of this study showed significant improvement in ranking (measured by Precision at top K results, Normalized Discounted Cumulative Gain, and Mean Average Precision) over methods that do not consider implicit feedback in ranking or re-ranking. They also found that this type of implicit feedback was particularly valuable for queries with poor original ranking of results.

The features demonstrated in the above listed studies may or may not be consciously considered by searchers when browsing through the results in order to make an informed decision regarding query reformulation. The features, or as they are referred to in this document, the cues, that searchers actually gather and pay attention to for reformulation purposes will be explored in this study through thorough questioning of the searchers regarding their decision-making process. The main set of cues will mostly be on the snippet level, but some will be on the whole page level as well.

The purpose of this section was to show that the relevant literature provides evidence that cues may exist, but how cues are represented and incorporated in searchers' decision-making and reasoning as well as how they influence query reformulation, is still to be discovered in this study.

3.2 Domain knowledge and cue elicitation and usage

Domain knowledge expertise has been explored with regard to different search parameters, such as retrieval success, query contents, query reformulation, and more. As Wildemuth (2004) noted in her literature review, based on previous research using pre-web retrieval systems, there had been no conclusive evidence regarding the effects of domain knowledge on retrieval success. The inconclusive nature of the results may be a consequence of how the research was conducted, on the nature of system studied, and whether or not the system is specialized or is intended for the general audience (like web search engines). More recent publications, such as White et al. (2009) have found that domain experts were more successful than non-experts when searching within their domain of expertise. In any case, even if it is unclear how domain knowledge affects retrieval success, if there is evidence suggesting some kind of relationship between domain knowledge and search success, there may also exist a relationship between domain knowledge and cue usage. Since this study also looks at the effectiveness of search, accounting for domain knowledge may be an opportunity to see how it manifests itself in cue elicitation and usage while the cues contribute to the different levels of effectiveness of search.

Additional research in the area has shown that domain knowledge affects other aspects of search behavior, such as searcher evaluation of the results and query reformulations. Since cues are gathered by the searcher from search engine results, the way searchers evaluate the results is an important aspect that may influence cue elicitation. Since Jenkins et al. (2003) found that domain

and web novices did little to no evaluation of the search results, this argues for including domain knowledge as one of the variables in this study. In addition, domain knowledge has also been shown to influence the reformulation technique (less effective for low domain knowledge and more elaborate for high domain knowledge) (Hembrooke et al., 2005), the vocabulary and term selection in the query (Allen, 1991; Hsieh-Yee, 1993; Vakkari, Pennanen, & Serola, 2003; White et al., 2009), as well as the number and length of queries (White et al., 2009; Wildemuth, 2004). These search behaviors, which are affected by domain knowledge level, may be tied to elicitation and employment of cues. If searchers indeed employ cues for the purpose of query reformulation, this means that actions such as selecting new terms for the modified query are driven by cues and mediated by their level of domain knowledge. In other words, since this study deals with how cues lead to query reformulation, anything that affects query reformulation and term selection, should be taken into consideration.

In addition, since this study is exploring how people evaluate search results in order to come up with terms for subsequent queries and what strategy they employ when reformulating the query, domain knowledge is an important factor to take into account. As Hembrooke et al. (2005) pointed out:

> since subject knowledge will impact how well a user is able to articulate their information need, their initial search term query may well be limited in terms of complexity, appropriateness, and their ability to relate the important semantic relation between multiple search terms. The documents retrieved then are apt to be less relevant. Here the novice is doubly disadvantaged: The novices' lack of conceptual sophistication to begin with presents the additional challenge of how well they can be expected to assess the relevancy of an already compromised "information patch". (Hembrooke et al., 2005, p. 868)

This conclusion emphasizes that domain knowledge affects the searcher's term selection as well as her ability to evaluate the results properly and therefore is an important factor that may affect cue elicitation and usage when reformulating a query.

In conclusion, since domain knowledge has been shown to affect query reformulation behavior and search results evaluation, it may also affect the way cues which appear on search engines result pages or the target pages are used by the searcher when reformulating a query. Therefore, this study took this factor into account and explored how domain knowledge mediates the cues employed by the searcher.

3.3 Search expertise and cue elicitation and usage

During the 80s, researchers explored various searcher characteristics related to search expertise. This included characteristics such as search experience, training, cognitive characteristics, and intelligence and personality traits as they correlate with outcome measures and process variables for the older pre-web systems (Hsieh-Yee, 1993). In these studies, the searchers were usually trained professionals and the characteristic of experience was based on parameters such as how long they've been searchers and how many searches executed per month (e.g. Howard, 1982). The findings of these studies were rather counter-intuitive. As Hsieh-Yee points out, "contrary to what common sense would lead one to expect, search experience, training, and cognitive styles were found to have little association with search outcome or search process" (Hsieh-Yee, 1993, p. 162). In order to explain the surprising results, Hsieh-Yee referred to Fidel's (Fidel, 1987) insight, according which the reason was experiments which were not adequate for identifying the relationships between search experience and the dependent variables. As part of her explanation, Fidel claimed that "the number of search terms used" did not describe the process and measures, but rather only how many terms were keyed in by a searcher. Also, she stated that there was

more than one way of measuring searching experience, which could be one of the reasons for obtaining results that contradict common sense. Hsieh-Yee also added that within-group variability and small sample sizes could also be a possible reason for these findings.

While the association between search experience and search outcome hasn't been proven to be very strong, there are other elements, which search experience has been shown to have effect on. Some of these elements were reviewed in the Spink and Saracevic' (1997) study on selection and effectiveness of search terms. For example, Saracevic et al. (1988a) compared the search terms chosen by experienced searchers dealing with the same information problem. They found that different professional searchers selected significantly different search terms. On average, the overlap of terms selected for the same questions was only 27%. In her study, Hsieh-Yee (1993) reached a similar conclusion. She found that given the same information problem, novice and experienced searchers selected and manipulated search terms differently and also varied in their use of synonyms. More specifically, novices relied on non-thesaurus search terms and employed fewer sources than the experienced searchers. These studies suggest that search experience is a factor to consider because it influences term selection. There is, however, a need to keep in mind that when compared to the modern web search engines, systems described in the literature mentioned above had a different functionality, especially pertaining to term selection. In addition, since the users of web search engines are generally not professional searchers, their experience cannot be measured in the same way.

With the advent of web search engine research some studies used search expertise and web expertise/experience interchangeably. In addition, both when talking about the web and about search, authors referred to experience and expertise interchangeably as well. It means that they didn't always differentiate between experience (in terms of years of experience, time working on

the web or with search engines) and expertise (in terms of competence and knowledge skills). These two types of mix-up could be observed in several studies (Hölscher & Strube, 2000; Jenkins et al., 2003; Lazonder et al., 2000). These studies were exploring the effects of web expertise and experience on search performance and regarded web expertise/experience as a variable that represents search expertise. Some authors also incorporated into their models both the variable of web expertise/experience and domain knowledge as two independent variables. Web experience or expertise may have been useful parameters back in the web's early days. Nowadays, however, there may be web searchers, especially those raised using Web search, who have the same experience/expertise on the web, but may have a different set of skills and competence when it comes to search engine expertise. In this case web expertise and search expertise probably do not represent the same concept. As Hargittai and Hsieh (2012) have shown, familiarity with advanced search may contribute to the users' internet skills. This, however, may not necessarily work the other way around, meaning that the user's internet skills may not indicate how good her searching skills are. In addition, experience and expertise also have different meanings and therefore, when measuring expertise, there is a need to find a measure that represents skills and competence and not only the amount of time spent with the search engine or a number of queries executed.

Along with the above mentioned representations of search expertise, throughout the years another indicator has emerged. It is based on the mental model that a user has of the system. A mental model is an internal conceptualization of the interaction between the person and the system that helps the person master the system (Norman, 1983). One of the studies which employed the mental model theory with search rather early is Borgman's (1986) research. It was based on the premise that people can be trained to develop a "mental model" or a qualitative

simulation of a system. This model is supposed to assist the user to produce methods for interacting with the system and keeping track of their position in the system. The model-based training of the users did not affect their performance on simple tasks, but the users with model-based training did perform better on complex tasks that required the users to build on the basic operations of the system (Borgman, 1986).

A more recent study (Holman, 2011) which explored millennial students' mental models of search discovered their tendency to formulate simple keyword or phrases searches with common misspellings and incorrect logic. Holman found that none of the students in her study had strong mental models of search mechanisms, but the ones that had stronger models managed to construct more complex searches than the students with weak models. When trying to elicit the students' mental models of the search systems they used, Holman conducted post test interviews with her participants, during which she asked various questions intended to elicit the students' mental model. Holman's interview included questions such as "How does a search engine know what you're looking for?" and "With as much detail as possible, explain how your search tool works. In other words, what does the system "DO" with your search terms?" These results were consistent with past research mentioned by Holman, which showed performance differences for users with different mental models. For example Dimitroff's (1992) findings showed that college students with more robust mental models were searching in an online library catalog more effectively than those without some mental image of the system , who were not as competent. Mental models seemed also to have an effect on the speed of search. In a study by Kerr (1990) faster searchers had more developed conceptualizations of the system than slower searchers did.

As prior studies show, the searchers' mental model seems to influence the way searchers perform and how they formulate and reformulate queries. As described above, other ways of representing searcher's expertise, such as web expertise/experience and years of experience with search engines, may not be suitable for this study, because these measures were designed for much older systems and/or for searchers who are different than the typical web searcher. Therefore, in this study, mental models are used as a measure that represents search expertise. Eliciting the searcher's mental model should also enable isolating the skills and competence of search engine usage from measures that represent search experience instead of expertise. With the help of questions similar to the ones used by Holman (2011), extracting the searcher's mental model of the web search engine should be an adequate way of representing the searcher's search expertise as it may affect cue elicitation and usage.

4 Methodology

4.1 Overview of the methodological approach

As opposed to research on query reformulation category prediction or on utilization of reformulations to get implicit feedback, this study attempts to explore what factors influence how the searcher reformulates a query. Therefore, the study relies on methods that emphasize user's reasoning about her experience with the system. The main challenges in this regard are how to elicit from the searchers information relevant to the research questions and how to create a valid and reliable design with sufficient controls over the search process. This section describes a design which was revised based on the results of a pilot study and a pre-test, both conducted in order to test both the elicitation method and the search tasks.

Earlier studies, such as Lorigo et al. (2008), Pan et al. (2007), Granka et al. (2004) as well as Guan and Cutrell (2007), were able to identify elements on the SERP that the searchers were looking at in order to feed their decision-making process. Yet, eliciting how searchers utilize those elements as cues and which cues are actually used - remains a challenging task. The method that was assessed during the pilot study, pre-test, and in the actual study is *Stimulated Recall* (Kelly, 2009). This method is used to collect the same type of data as the *think-aloud protocol*, but is different, because data is collected both during and after search. In *Stimulated Recall,* the researcher records the screen of the computer as the participant completes a searching task. After the task is complete, the recording is played back to the participant, who is asked to articulate thinking and decision-making that took place during the completion of the search task. General instructions can be provided or the participant can be asked specific questions or to focus on specific features or processes. In the described study, the *Stimulated Recall* method was

employed by screen capturing during the search session and then playing the captured videos, while the observer asked specific questions regarding the decision-making process at certain points in the recording.

When it comes to research design, there is a need to find a suitable compromise between conducting a study in a relatively controlled environment in order to ensure internal validity and allowing the participants to perform in an environment as close to their natural settings as possible, to maintain external validity. In this regard, related earlier studies ranged from completely controlled lab environments, to observations at people's workplace performing their own tasks. Based on the available knowledge about query reformulation and given the purpose of this research, there is a need to control for a number of factors in the search settings as well as some of participants' attributes that can be collected with a questionnaire. Some attributes, which have been reported in the literature, were found to influence query (re)formulations. As discussed in sections 3.2 and 3.3, attributes such as domain knowledge and search expertise have been shown to affect query (re)formulation. Therefore, these searcher attributes are examined in this study for the purpose of exploring their effects on cue usage, during the data analysis stage.

As mentioned earlier, the general study environment is also important to the success of this research. A number of elements can be essential when attempting to control for the search settings. One such element is the search task which was already discussed in detail in section 2.6. Another element is the search environment (i.e. search engine). To control for the search environment, the same search engine was used by all the participants. Using the same search engine for all the participants provides control for some variables including visual and content cues. Google, due to its popularity and prevalence, is an appropriate choice for this research.

Finally, controlling for participants' attributes can be done through sampling of participants or by gathering data regarding the values of those attributes and controlling for these variables later in the analysis stage. As demonstrated in the literature review, previous research suggests that one option could be stratified sampling by search engine expertise since users with more experience with search engines may use cues differently than the ones who are not as skilled. Another basis for stratification could be domain knowledge of the tasks, as it has also been found as influencing query reformulation. For addressing these variables later, in the data analysis stage, the participants were asked in a post-observation questionnaire regarding their domain knowledge in each of the domains that the search tasks cover. Three questions that assessed the participants' search engine mental model, appeared on the questionnaire, in order to evaluate their search expertise. The score that reflected the participants' search expertise was measured through their search engine mental model. The mental model was assessed based on their level of understanding of how query terms are used in document retrieval of a basic search engine, as it was expressed in their answers to the questionnaire questions. When evaluating their answers, I was expecting to see explanations that mentioned query terms and the closer the explanation was to the way a basic search engine utilizes query terms to rank the results, the higher was the score of the participant.

Since the participants were recruited among college students, this poses limitations on variability of their search engine expertise and domain knowledge. This issue will be addressed in the section on threats to validity.

4.2 Assigned tasks

As mentioned before, choosing appropriate tasks for search engine research is crucial to the success of the study. Previous studies have shown that certain attributes of a search task have a

wide range of impacts on search performance and behavior. In order to develop assigned tasks that would elicit certain types of search behavior from the searchers, some studies tried to identify tasks characteristics. For example, Wildemuth and Freund (2012) identified a set of task characteristics associated with exploratory search. The effect of different attributes of a task such as complexity, difficulty, breadth, etc. on search behavior has been studied extensively (e.g. Byström & Järvelin, 1995; Cole et al., 2010; Kelly & Belkin, 2004; Liu et al., 2010; Smith, 2009; White et al., 2005). Those studies tried to find the relationships between task attributes and search behavior, which is manifested through relevance feedback, eye movements, query reformulations, and more. For instance, Saracevic and Kantor (1988) found that broad search tasks led to higher precision of retrieval as opposed to specific tasks which led to lower precision.

The tasks in this study were intended to be challenging and difficult tasks, ones that would generate multiple query reformulations due to the difficulty locating relevant answers. The question of what constitutes a difficult task is not an easy one; there is no particular consensus in the literature on this topic. Should it only be a difficult task, one that the answer to it is not easy to find or also a complex task? Previous studies examined different task attributes and their effects on searcher behavior, but as demonstrated next, the operationalization of these attributes was problematic.

Creating a task that has the appropriate level of complexity is a challenging endeavor. One reason, discussed in previous research, is the fact that even when the tasks in different studies are characterized as "complex", their descriptions differ markedly. In various studies, complexity tends to be defined as a collection of one or more of the following task dimensions: structure (Lazonder et al., 2000; Sharit et al., 2008), the number of search concepts involved (Saracevic et

al., 1988b), certainty or *a priori* determinability (Bell & Ruthven, 2004; Browne et al., 2007; Byström, 2002; Campbell, 1988), number of facets (Bilal, 2001; Capra et al., 2007), length of the search path (Capra et al., 2007; Cole et al., 2010; Gwizdka & Spence, 2006), cognitive effort (Bilal, 2001; Capra et al., 2007) and topic familiarity (Bilal, 2001; Browne et al., 2007). Smith (2010) didn't define her tasks as complex, but did wish to make every search somewhat difficult. Therefore, Smith designed the statements to require disambiguation with the help of query reformulation. The above definitions are not always helpful in generating new tasks for research because some of them can be measured only retrospectively, after the task has been used in a search (such as length of the search path, cognitive effort). Therefore, reusing or adapting a set of tasks that have been used before is helpful, because 1) it's difficult to design tasks of appropriate complexity, and 2) it allows more direct comparison with earlier work.

Another problematic issue with the complexity of a task is the ability to foresee when a task may be too complex. A study that investigated complexity (White et al., 2005), revealed that when the search task was more complex (varied by the number of potential information sources and types of information required, to complete a task), users rarely found results they regarded as completely relevant and struggled to find relevant information. As a result, the users were unable to communicate relevance feedback to the search system. This could mean when tasks are too complex the user is too focused on trying to assess relevance instead of how to reformulate the query. This means that tasks that are too complex are not suitable for this study. Therefore, the tasks chosen for this study were difficult tasks, ones that make it difficult to find answer to. Both the pilot and the pre-test (described in 4.3.1 and 4.5) tested different tasks and were designed to assess the suitability of the chosen tasks and adapt them accordingly. The final tasks and the instructions used in the actual study are as following:

Please use Google's search engine to find answers to the following questions (try to find as many sources as possible):

[copy and paste the **answer** and the **URLs that contain the answer** into the word document]

1) What minerals can float in salt water?
2) What Australian sport was adopted in Britain?
3) What is the partial pressure of oxygen at an altitude of 5000 feet?
4) When was the Jominy test invented?
5) How do you find the exact time a hard drive was last formatted on a PC?

4.3 The pilot study and its observations

The pilot was conducted mostly over the course of the spring semester in 2011, one day a week, with 1-3 participants taking part in the pilot on each of the days. In total, eight participants who were recruited from a graduate class at the iSchool took part in the pilot study and two additional fellow PhD student participants. The pilot study's protocol was quite similar to the protocol of the actual study, with an additional "cue-specific questioning" part which will be discussed later (the whole protocol can be found in Appendix 2). The purpose of the pilot was to test the protocol, the tasks, refine and clarify the instructions, and validate that the current method allows eliciting the cues that the participants used for reformulation.

4.3.1 Assigned search tasks in the pilot

The main goal of the pilot was to assess the compatibility of the tasks to this research. In order to test the search tasks in this pilot study, I collected tasks from different sources and of various types. When coming up with the tasks, I considered some of the tasks used in previous research

(such as by Toms et al. (2008) which was reused in other studies as well), but initial testing of these tasks with Google's search engine showed that they were not complex or difficult enough. Therefore, I developed the first 6 tasks (a through f) in the first set of tasks below and the rest were an adapted the rest from the TREC filtering ("TREC," 2002) and ad-hoc track ("TREC," 2001). The first set of tasks were phrased in the form of questions and the participant was asked to use Google's search engine to find answers to the following questions:

a) What is the name of an iPhone app that schedules a shutdown of other apps at a certain time in the day?

b) How much caffeine is there in maple syrup?

c) How to prevent a cat from getting asthma?

d) What are the names of dinosaurs that had a small skull, big body structure, and used to eat plants?

e) Into what languages has the book "Bad Science" by Ben Goldacre been translated?

f) Where can I find a list of congress members of all times, which includes information about their ethnic races?

g) What types of crimes are committed by people who have been previously convicted and later released or paroled from prison before committing this crime?

h) What parents are doing to prepare for spiraling cost of college tuition?

i) What actions are being taken to make the quality of children's television in the U.S better

j) What actions are being taken by U.S. airplane manufacturers to improve the safety of their passenger aircraft?

k) Why were camels domesticated?

The first four participants were each given a different set of 4 tasks from the above list. This way all 11 questions could be tested.

A second set of tasks was taken from Smith's dissertation (Smith, 2010), since she had a similar goal: crafting tasks that are fairly difficult and trigger query reformulations. One characteristic that is consistent across all tasks in this set is their ambiguous nature, which allowed more room for query reformulation. These tasks are cast as a statement rather than a question and therefore, the instructions for the assignment could also vary. For the last two participants in the pilot the format of the statements was changed into question format (Appendix 1).

l) It is difficult to produce containers that maintain the freshness of vegetables during shipping.
m) Fishermen find it difficult to earn net profit.
n) Mints and treats that look like coins are favorite holiday candies.
o) It is difficult to secure a mortgage or insurance for property directly on the bank of a river.
p) For security, conductors carry radios as they move between stations.
q) Drinking water helps you stay well.
r) It is easy to tire when driving a car.

Each of the five participants that got these tasks (4 different tasks per each participant) had a different description of how they were supposed to perform these tasks. Following are the four instruction variations, the third one being almost identical to the instructions that Smith's participants got:

1) "Please use Google's search engine to find information sources that support the following statements."
2) "Please use Google's search engine to find *websites* that support the following statements."
3) "In your job as a trainee, you support the journalists at the newspaper. Your responsibility is to find information about the journalists' article topics. Today you need to search for good sources of information about four different topics that the journalists are working on. Please find as many good information sources as possible. A good information source is a source that you could and would use to get information about the topic. Any source with information that will inform the journalist on the topic can be considered a "good" source. You need to find as many "good" information sources as you can, but it is also important to avoid choosing information sources that are not good. Please user Google's search engine to find good information sources."
4) "Please use Google's search engine to find web pages that support or refute the following statements."

Since there was a gap between the pilot sessions, this allowed changing and adjusting of the tasks and their instructions from one participant to the next. After participant #8, the tasks from the second set were modified to question format, since it appeared from earlier pilot observations that question format was more suitable because it was easier to understand and helped the participants stay focused on the outcome. One participant got a mix of tasks from both the first set of tasks and from the second set of tasks. These tasks were chosen based on the sessions with previous participants and the tasks that worked best (yielded plenty of query reformulations) in

each set, were chosen. All the tasks, their variations, and instructions which were used in the pilot can be found in Appendix 1).

The observations from using these tasks and instructions will be discussed in the upcoming sub-section dealing with observations from the pilot.

4.3.2 Pilot design

As mentioned earlier, the research design of the pilot study was almost identical to the final protocol, with an additional "cue-specific questioning" part (the two protocol versions can be found in Appendix 2). Both the search session and the interview part were recorded with screen-capturing software. The software for recording the screen in the pilot was open-source software called CamStudio. Since CamStudio does not enable reflecting what is being recorded on multiple screens and marking certain points in time, during the pilot, the observer was sitting behind the participant and recorded the time-stamps manually when each query reformulation occurred. As already noted in the research design, the software used in the actual study was different. This is due to the effects that observing from behind the participant's shoulder may have on her search behavior. Therefore, the research design of the actual study employed the Morae software, which allows flagging of certain points in time, while watching the recording in real-time on a different computer.

In the cue-specific questioning part, the participant was presented with specific cues (presented below) and asked to provide an example of each of these cues (if they existed) that precede query revisions as they appear in the recording. The purpose of this part was to discover whether some of the potential cues were actually used during the search session and whether they could be

identified by participants. The cues and prompt of additional cues that were presented on a printed page to the participants where[10]:

- **Distance** between the query terms on the results page or the full webpage.
- **Absence of** query terms or other words that you expected to appear in the results page or the full webpage.
- Words that are **immediately preceding or following** your query terms in the results page of the full webpage.
- **Parts of speech** of the query terms that appear in the results page or full webpage[11].
- **Position** of the query terms on the results page (in the body of the snippet, in the title, or the URL).
- **Order of appearance** of the query terms on the results page or the full webpage.
- Can you think of other cues that you may have used in this session when reformulating a query?

After the questioning was over, I also asked the participants if they had any comments about the way the pilot was conducted in general and about tasks in particular.

4.3.3 Observations from the pilot

Overall, the elicitation of cues using this protocol was successful and the participants found the questions to be clear enough. If the first question ("why did you make the change") was misinterpreted and the participant didn't realize that she was to talk about elements on the SERP

[10] This part was used only in the pilot and will not be employed in the proposed study.
[11] The meaning of this cue was not very clear to the participants and they needed extra explanations of what exactly it meant. This was one of the reasons why this cue-specific questioning part will be left out of the proposed study.

or the full web page, then the next one ("how did you know") or the one after that ("what exactly on that page") helped capture the cues. As a result of this questioning, most of the participants in the pilot were able to identify cues such as distance between terms, order of terms, absence of query terms in the results, and position of terms in the snippet, without being specifically prompted about them.

Some of the observations did indicate that changes needed to be made to the protocol. Several sessions showed that even though the observer and the participant went over all the reformulations, when a few minutes later, the participant was asked to provide examples of the above mentioned cues, she found it difficult to make these associations and recall the examples. This incident occurred with several participants, even though some of the cues were brought up by the participants in the general questioning (first part). This was one of the reasons for removing the questions asking for examples of cues from the protocol. Another reason for abandoning this kind of questioning is due to the fact that this is an exploratory study and the potential cues even if plausible, do not have a strong theoretical justification. Therefore, the best approach would be to discover the cues in an exploratory manner as the first part (the general questioning) allows.

Another observation which indicated that some changes needed to be made to the protocol is related to the tasks. Some of the tasks in the first set (a through k) were problematic, because even though most of them did trigger reformulations, they did not allow enough "room" for reformulation. In other words, even though the participants did wish to reformulate, there weren't many alternatives to the initial query. This was true mostly for tasks b to e. The remaining tasks in that set (f to k) yielded too many remotely related web pages or PDF files that were too general and did not answer the question directly. As a result, participants attempted to

read through the long retrieved articles as they tried to glean relevant answers and as a result neglected the evaluation of the SERP and reformulations. In their feedback, some of the participants noted that tasks a) and b) were too difficult, most probably because none of them succeeded to find an answer to task a).

The second set of tasks appeared to be at the correct level of difficulty and usually (depending on the instructions to the task) triggered query reformulations. However, it appeared that with the first set of tasks the participants were more focused on the goal of finding an answer, as opposed to trying to find information sources/web-pages that support the statement. It appeared that with the first set of tasks, since the participants were more focused and the goal was clearer to them, they paid more attention to the details of the question. This is as opposed to their response to the second set of tasks during which, they tried different angles to approach these tasks. In addition, their reaction varied significantly as a function of the instructions.

Since the second set of tasks was in statement form, I tried three different types of instructions for this set. When the journalistic style was introduced or when there was an option to refute the statements, some of the participants regarded each task in a "research-like" manner. In other words, instead of attacking the task directly and using the statement as their initial query, they looked for alternative angles to deal with the task. For example, to refute the statement "It is difficult to produce containers that maintain the freshness of vegetables during shipping", the participant looked for proof that such containers existed and hoped that this would be enough to refute the statement. The participant who got journalistic-style instructions was trying to gather evidence that would enable constructing a claim for each task. Regardless of the instructions, some of the participants ignored parts of the statements and looked for sources that supported only part of the statement (for example: looking for information on containers that maintain

freshness and ignoring the issue of difficulty of production). Another approach was to break the statement into smaller queries or to try performing the search within a website instead of using Google. These approaches made the tasks easier than they could have been had the participants formulated a query that was very similar to the description of the task.

To summarize, the observations from the pilot reveal that the tasks that were more suited for reformulations and their level of difficulties are the ones that appear in the second set (l through r). The instructions to find information sources that support the statements (simulated task) worked reasonably well, but there were instances when the pilot participants were too focused on a certain part of the statement and not on the statement as a whole. They would sometimes dwell on a certain document for too long, simply reading through it, without having a more definite goal or an idea of what the answer should look like. As mentioned above, the tasks which were framed as general questions (f to k), also led the participants reading through the long retrieved articles as they tried to glean relevant answers and as a result neglected the evaluation of the SERP in comparison to their query. Therefore after the pilot was complete, it was evident that additional testing that focused primarily on the participants' response to the search tasks needed to be performed before the actual study. For this purpose, a pre-test was conducted just before the actual study and an alternative set of tasks, which were posed as more focused questions, was examined and found more suitable.

4.4 Participants

The initial goal for the number of subjects in this study was chosen to be comparable to other studies that employed a similar method of elicitation and had a similar exploratory nature. In related studies, Barry (1998), for example, recruited 18 participants for her clues to relevance

study, while both Teevan (2004) and O'Day & Jeffries (1993) had 15 participants[12]. Based on this previous research, the goal for this study was to recruit between 20 and 30 participants, depending on the response rate. By the end of the data collection, 43 participants had taken part in the study, 10 were used for pre-testing and the remaining 33 for the actual data collection. All the participants were students from Syracuse University, Cornell University, and SUNY College of Environmental Science and Forestry (ESF). Both undergraduate and graduate students were recruited and they were offered a monetary reward to encourage them to participate[13].

During recruitment, there was only one exclusion criteria for the participants of this study. The linguistic nature of this study suggests that in order to prevent the interference of linguistic competence, native English speakers seem to be more suitable participants than non-native English speakers. Therefore, in an attempt to minimize differences in cue elicitation and usage due to insufficient English proficiency, only native speakers were recruited for this study.

Recruitment efforts were made by means of announcements in classes, flyers on billboards, and a participant recruitment system at Cornell Business School, called SONA[14]. A total of 104 students signed up for the study, but only 43 participants eventually agreed to participate and showed up for the study. The first 10 students served as pre-test participants and were offered $10 for their participation and a chance to win a $50 Amazon gift card. Given the relatively low response during the pre-test, the incentive for participation in the actual study was raised to $15

[12] Similarly to this research, these were exploratory studies, but they had very little quantitative data analysis and did not strive to achieve statistical significance in their results. In this study, however, in order to answer the specific research questions, a statistical analysis was performed. The issue of whether or not statistical significance can be achieved with data collected from a limited number of participants, such as the range of participants that is usually used for an exploratory study, depends not only on the number of participants, but more on the unit of analysis. The unit of analysis will be discussed later in the chapter, but at this point it can be noted that it won't be on a participant level, which increases the possibility of achieving statistical significance.

[13] The pilot and the actual study described in this document were approved by the Human Subject Review Board and will be amended to fit the modifications made to the protocol and the research design and filed again for approval.

[14] http://www.johnson.cornell.edu/Business-Simulation-Lab/Participate-in-a-Study.aspx

and a chance to win a $50 Amazon gift card. Out of the 33 participants that took part in the actual study, 14 were male and 19 female. Their average age at the time of the study was 22.3; the youngest being 18 and the oldest 36. Eleven of the participants were from Syracuse University, 10 from ESF, and 12 from Cornell. Table 1 contains the breakdown of participants by their home college across all three institutions.

TABLE 1: BREAKDOWN OF PARTICIPANTS BY THEIR HOME COLLEGE (ALL UNIVERSITIES COMBINED)

Home College	Count
Agriculture and Life Sciences	1
Arts and Sciences	3
Engineering	5
ESF	10
Human Ecology	1
Information Studies	7
School of Management	5
School of Public Communications	1
Grand Total	33

100% of the participants indicated that their preferred web search engine was Google. The average, maximum, and minimum self reported domain knowledge in the field of each task appears in Table 2. The scale used in these questionnaire questions were on a range of 1 to 5, five being the highest level of familiarity with that domain and 1 being the lowest level of familiarity. Domain knowledge was rather low for tasks A through D, the median being 2.0 or even lower for task C. As can be seen in Figure 1, the frequency of domain knowledge 1 and domain knowledge 2 for tasks A through D is quite high. Task E got the highest median domain knowledge of 3.0. Figure 1 also shows that the frequency of domain knowledge 3, 4, and 5 for task E was higher than for 1 and 2.

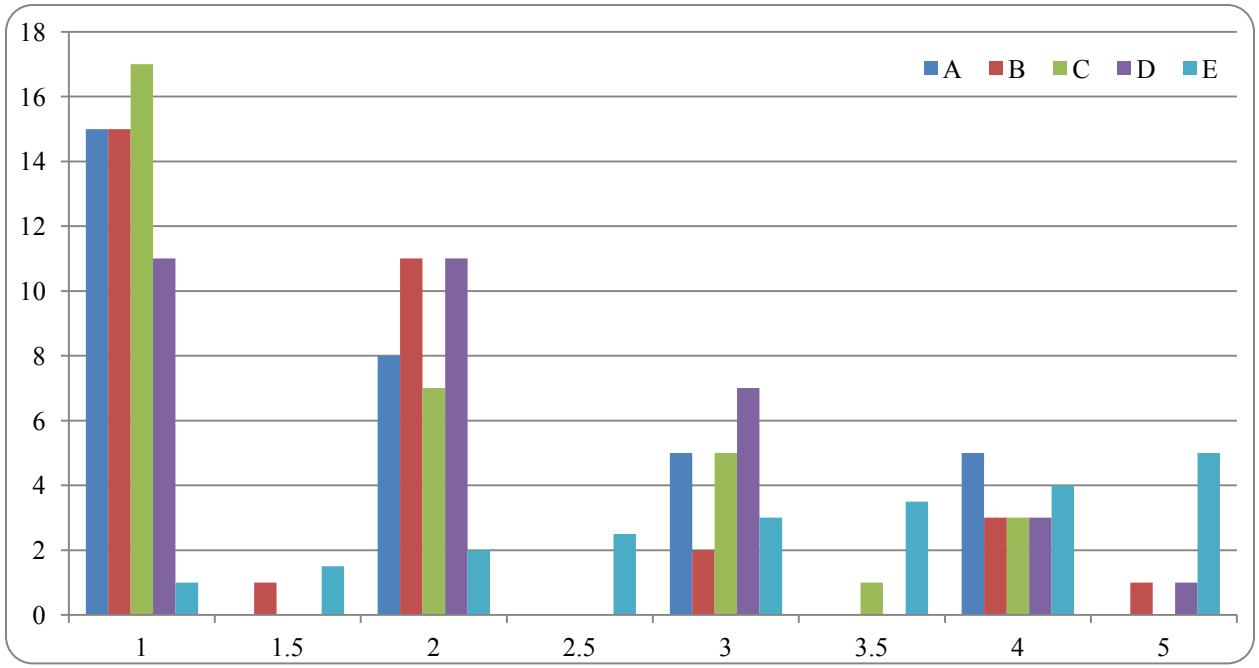

FIGURE 1: FREQUENCY OF SELF REPORTED DOMAIN KNOWLEDGE

TABLE 2: SELF REPORTED DOMAIN KNOWLEDGE

Task	Average	Standard Deviation	Max	Min
A (minerals)	2	1.1	4	1
B (Australian sports)	1.9	1.1	5	1
C (partial pressure)	1.9	1.1	4	1
D (Jominy test)	2.2	1.1	5	1
E (hard drive)	3.4	0.8	5	2

The average search expertise (evaluated by me, based on their answers to 3 search engine mental model questions) was 6.73 with a standard deviation of 1.57. Theoretically the scale was 1 to 10, but in practice, the lowest score a participant got was 3. As can be seen in Figure 2, the participants' search expertise has a bell shaped distribution, with the most frequent search expertise value being 7.

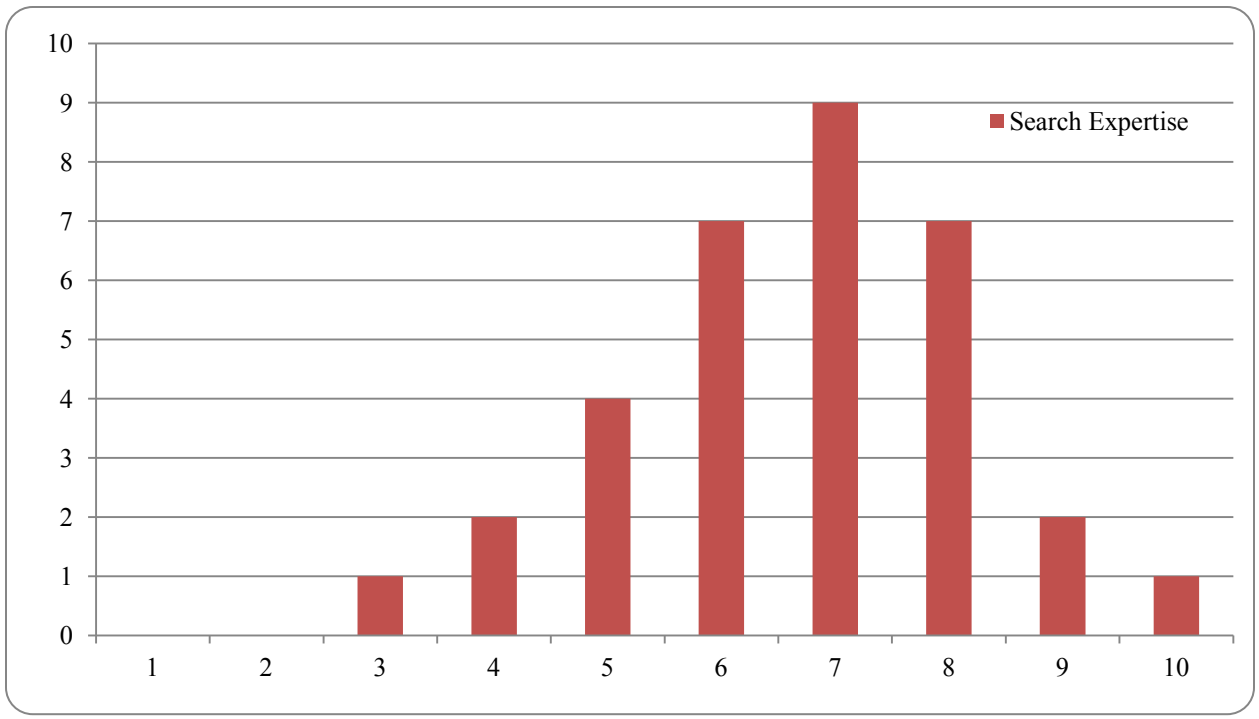

FIGURE 2: SEARCH EXPERTISE DISTRIBUTION

4.5 Research design

The design of the final study described in this document was developed based on relevant literature and adjustments were made based on the pilot and the pre-test conducted before the actual study. The main goal of the pilot and the pre-test was to examine the assigned tasks, but also to test the protocol.

The design of this research is based on the *Stimulated Recall* technique as described by Kelly (2009) and presented in section 4.1. This study employed a very similar technique along with assigned search tasks that facilitate multiple query reformulations. The regular think-aloud technique was also considered as a possible method for this study. The think aloud technique consists of instructing the participant to talk aloud while performing a task and is one of the most direct and widely employed routines to obtain information about participant's internal states (Ericsson & Simon, 1980). However, think-aloud was not very suitable for the methodological

requirements of this study, because as Ericson and Simon state: “when subjects give indications that they are working under a heavy cognitive load, they tend to stop verbalizing or they provide less complete verbalizations” (Ericsson & Simon, 1980, pp. 242–243). Since performing complex searching tasks may constitute heavy cognitive load, it may be difficult for the participants to both search and verbalize their decision-making process at the same time.

Each observation session started with the signing of the consent form. Once this was complete, the observer[15] briefed the participant regarding the purpose of the study, how many tasks will be assigned, what will be recorded, and notified her that the observer will go over the recording (together with the participant) once the search tasks are completed and will ask questions. In order to make the participant more comfortable and not anxious about the observation, the observer made it clear to the participant that this observation is not a test and that even though the goal is to be able to perform the searches successfully – the process is more important than the outcome. The briefing also included a sentence describing an example of a process, which includes query reformulation. The purpose of this part of the briefing is to inform the searcher that query revision is considered normal and does not indicate that reformulation indicates failure or that it makes them incompetent searchers. During the briefing, the participant was also notified that there is no time limit, however, if it takes too long to finish a task, due to time constraints, the observer may have to intervene and ask the participant to move on to the next task. This usually happened if the participant's search session for a certain task exceeded 10 minutes.

[15] In the description of this design, I will be referring to the whole session (searching and interviewing) as “observation” and to the researcher guiding the stimulated recall and who will be present in the same room during the whole session along with the participant, “an observer”.

Next, the observer provided the participant with a list of 5 tasks printed on a sheet of paper, each task consisting of a question that the participant was instructed to answer by using Google (the tasks used in the pre-test can be found in Appendix 5). During the pre-test stage, the participants only received 4 questions; different sets from a pool of possible 7 questions as they appear in the appendix. The pre-test revealed that 1) there is enough time during each session to use 5 questions, 2) that the phrasing of one of the questions was confusing, while a different, improved phrasing was more clear to the participants, and 3) that one of the questions is too easy and does not lead to reformulations. As a result of the observations in this pre-test, in the actual study, 5 questions as shown in Appendix 8 were used. The order of the tasks was rotated between the participants. The observer asked the participant to write down the answer and save the URLs of websites that contain the answer to the question presented in the task, per each task. The participant simply had to copy & paste the URL from the address bar into a word document file[16]. The tasks used in this study are informational questions which have been adapted from questions assigned to participants who took part in a study exploring the strategies and behavior of successful searchers (Ageev, Guo, Lagun, & Agichtein, 2011). Ageev et al. selected their search tasks from community question answering sites such as wiki.answers.com and Yahoo! Answers. Their criteria for including the questions were that "the question should be clearly stated, had a clear answer, and that finding this answer was not a trivial task, that is, the answer was not retrieved simply by submitting the question verbatim to Google, Bing, or Yahoo! Search engines" (p. 347). Since the study described in this document relies on multiple reformulations, therefore it has a similar criteria for its tasks. Following several pilot sessions that focused

[16] It is important to keep URL saving simple, in a way that it would not interfere too much with the search process, intuitive for the participant, and would not change the interface of the SERP or the search engine in general. Using a simple notepad was the best compromise between simplicity and not interfering with the flow of the search sequence.

mainly on testing the questions in the tasks, I chose five questions and slightly modified the wording to make sure that the criterion of "finding the answer was not a trivial task" is satisfied and the questions are clear to the participants.

During each experimental session, screen capturing software was used to record everything that occurred on the screen (mouse movements, clicks, query (re)formulation) and to generate a video file of the whole session. Since the tasks were adapted from questions that were originally posted on community answering websites, results from these websites were hidden (the same solution was used by Ageev et al. (2011)). In order to support the Stimulated Recall process, a screen capturing tool, called Morae was used. It allowed the recorded session to be viewed in real time from another computer and let the observer mark the times at which reformulation occurred, without disturbing the participant. The participant was not aware that the observer was viewing the recording in real-time. Figure 3 shows an example of a screen shot taken from the observer's computer. The pink and orange diamonds indicate the markers placed by the observer. Pink markers were put at the beginning of each task and orange markers were inserted at the time of reformulation within each task. After the search session was over, the recording was saved together with the markers.

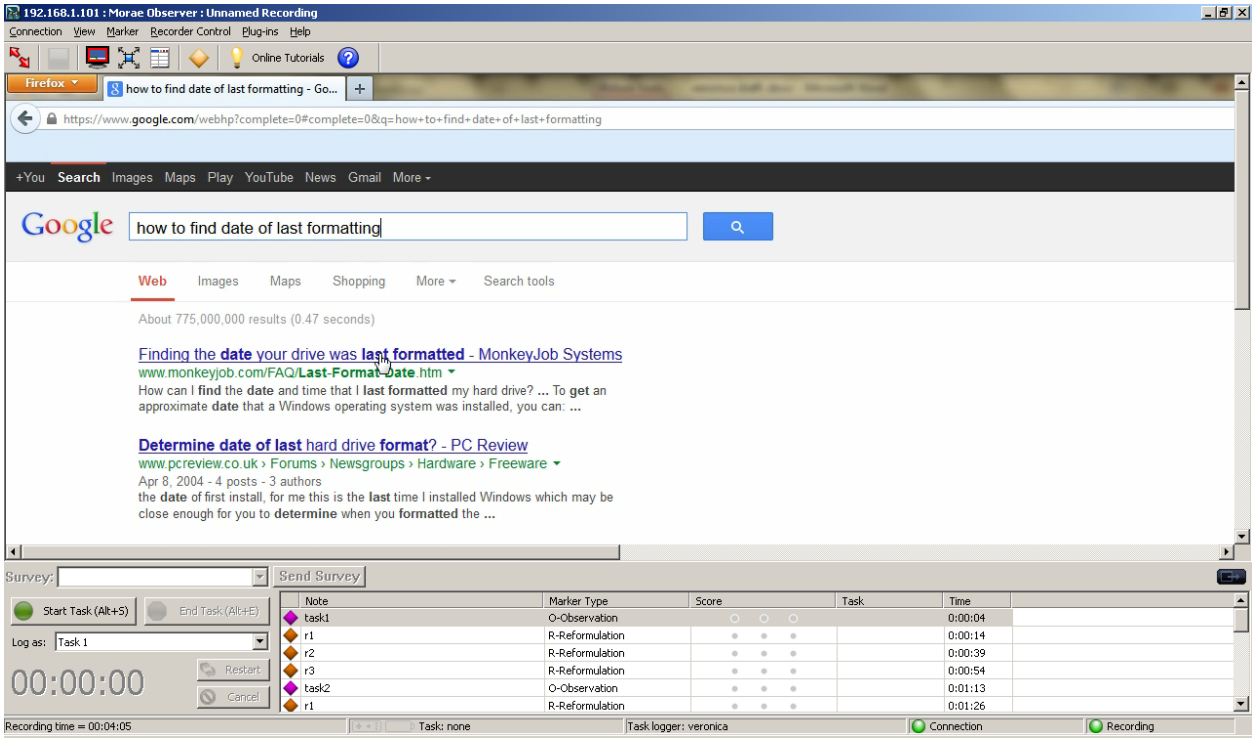

FIGURE 3: MORAE OBSERVER

In the next stage of the observation, the observer sat next to the participant and played the recording, which included the timeline markers. The observer navigated to the markers that represented reformulation points. During this part of the observation, the participant was able to fine-tune the navigation back and forth in a way that enabled her to see additional parts of the recording, which assisted her in answering the questions asked by the observer. The screen during this stage was also recorded by Morae, along with the audio through the computer's microphone. The cursor of the mouse was highlighted during this recording in order to emphasize which elements on the screen the participant was pointing at when discussing her decision making process. The highlighting is visible only once the recording of the questioning is being replayed. A screen shot of the recording during the questioning stage is shown in Figure 4. In the figure, the hand shaped cursor is the cursor from the original recorded search session and

the arrow shaped cursor is the cursor of the participant during the questioning session. The arrow shaped cursor appears highlighted in the recording of the questioning session.

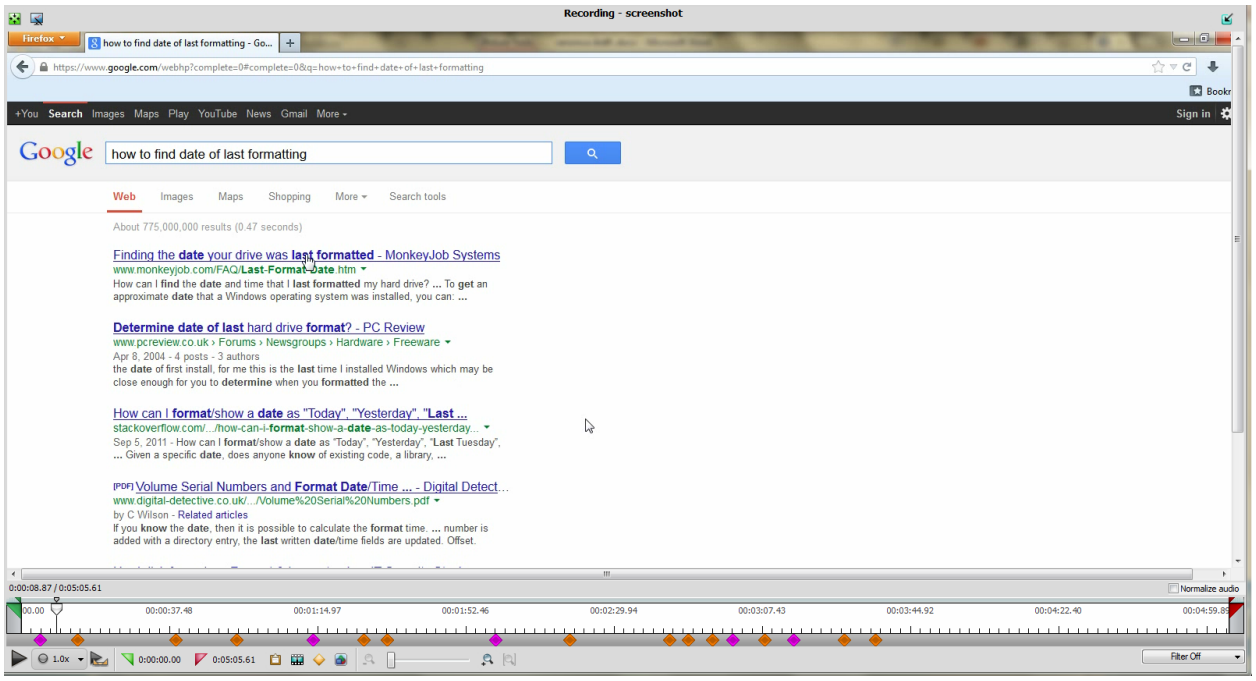

FIGURE 4: MORAE RECORDING DURING THE QUESTIONING STAGE

The following general questions (the full protocol can be found in Appendix 9) were asked by the observer at each query reformulation marker. The purpose of these questions was to elicit the cues that led the participants to reformulate their queries. It is important to probe as much as possible, because the reasoning behind reformulation is not necessarily a conscious process and thus requires extensive questioning in order to be able to elicit the decision-making behind it.

The questions asked during the questioning stage are:

- ***Why*** did you make the change in your query at this point?
- ***How*** did you know that you needed to make the change at this point? Additional probing if the participant said that they didn't see the answer: How did you know you were not getting the answer? What did you expect to see? Please give me specific examples in the

recording, of what helped you make this decision. Please point with mouse cursor at these examples on the results page or the full webpage.

- ***Is there anything on this page*** (results page or full webpage) that provided a hint that you needed to change the previous query? If so, please point with mouse cursor at specific examples on the results page or the full webpage, of what helped you make this decision. Additional probing: please point at the feature or the element that helped you realize that you needed to reformulate.
- ***Is there anything on this page*** (results page or full webpage) that helped you decide on what changes needed to be made to the previous query? Additional probing: please point with the mouse cursor at the feature or the element that helped you to reformulate.

The second question ("How did you know") is a modified version of the question as it was used in the pre-test (Appendix 6) and in the pilot (Appendix 2). This is due to the participants simply answering "I didn't see the answer", to this question in the pre-test. Probing further with "how did you know you were not getting the answer? what did you expect to see?" helped the participants think about the cues instead of the more general reason. The purpose of each question was to elicit the cues in an exploratory manner, without biasing the participant for particular cues. The questioning starts with a general question and becomes more specific and more cue oriented with each question. Both the pre-test and the actual study took place in the fall of 2012. The pilot study was conducted earlier, in the spring of 2011 and was discussed in section 4.3.

After the observation session was complete, the participants filled out a post-observation questionnaire (Appendix 10) to record their demographic details, domain knowledge on the

topics of the tasks, their search engine mental model, and search engine preference. Following are the questions used to record the self reported domain knowledge:

- How would you rate your level of familiarity with the world of **computers**[17] (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____
- How would you rate your level of familiarity with the world of **chemistry or chemical engineering**[18] (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____
- How would you rate your level of familiarity with the world of **history of sports**[19] (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____
- How would you rate your level of familiarity with the world of **material science**[20] (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____
- How would you rate your level of familiarity with the domain of **chemical compounds**[21] (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____

These are the questions which were asked in order to elicit the participants' search engine mental model:

Please answer the following questions with regard to your preferred web search engine:

1. How does this search engine know what you're looking for?
2. How does this search engine select and present the list of results?
3. With as much detail as possible, explain how this search engine works. In other words, what does the system "DO" with your search terms?

[17] represents task E (HD format)
[18] represents task C (partial pressure)
[19] represents task B (Australian sports)
[20] represents task D (Jominy test)
[21] represents task A (minerals)

4.6 Threats to validity and reliability

4.6.1 Internal validity

This study exhibits both internal and external threats to validity, which is quite typical with research that employs user-studies. I will start with a discussion on internal validity. First, this type of validity is challenged by the fact that the computer screen during the sessions is being captured and there is also audio recording of the questioning. This makes the participant more aware of the lab environment and she might perform differently than she would have without the recording. Second, the artificial lab-like setting in which the participants are placed during the search session and the fact that they are not using their own computers constitutes a threat to validity. In addition, the assignment of tasks also poses a threat to internal validity in this research. Assigned tasks make the environment more artificial, since these are not the participants' actual information needs and the tasks are subject to their ability to understand the task correctly and execute it as expected. Finally, since I utilized a commercial search engine, which is constantly going through changes (such as bucket testing, where only some of the users are affected by it), there is a risk that the SERP and the general layout changed from one participant to the next.

The threats described above were mitigated with several measures. First, the artificial setting was made less artificial by encouraging an informal and friendly atmosphere by emphasizing to the participant that this is not a test and that I am more interested in the process and their reasoning about it rather than in the final results. The idea is to decrease their anxiety and fear of not being able to perform as expected, something that they usually do not experience in their natural setting. The participants were also given an opportunity before the session to get acquainted with

the layout of the keyboard. Second, the screen capturing was rather unobtrusive and the recording of the reformulations was even more unobtrusive than in the pilot, with the help of Morae, which allows seeing the recording on a separate computer and marking reformulation time stamps on it.

The threat of using assigned tasks is the most crucial threat because previous research has shown that it can have a substantial effect on the way participants formulate queries. Therefore, the pilot and the pre-test consisted of many tasks of different types so that the tasks and their descriptions could be tested and the best ones chosen. The order of the assigned tasks was rotated between participants to eliminate the possibility of a "learning effect", where the participants might perform differently with tasks presented later in the search session compared to tasks performed earlier. It was also supposed to deal with a fatigue effect, where performance on later tasks might be affected by fatigue. Another option considered for resolving threats associated with assigned tasks contributing to the artificial setting, was to provide the participant with a greater number of tasks and ask her to choose only five. This way, the participant could choose only the tasks she identifies with and feels that they are more representative of her real needs. At the same time, this would have made it more difficult to compare between participants who did not choose the same task. Therefore this option was abandoned and all the participant received the same tasks.

The threat of a changing SERP or layout in general cannot be completely eliminated, because Google does tend to change its SERPs and the general appearance of the search results, but measures were taken to minimize it. In order to make sure that changes in the search engine minimally affect the participants and their interaction with the SERP, several actions were performed. The participants were not signed in with a user name to prevent any kind of personalization and bucket testing effects, instant search option was disabled, and query-auto-

complete was disabled as well[22]. The main reason for this is to control for the trigger of the query reformulations. The query reformulations that stand at the focus of this study are only those that are generated by the searcher and are triggered without the help of any auto-complete functionality of a search engine. In addition, this functionality may distract the searcher from cue utilization and create bias. Other features such as ads and facets (images, video, news, shopping, and others, which appear on the left hand side of the search results) which were not excluded, but are out of the scope of this research. In order to decrease the chance for changes in the search engine, the time frame for conducting the observations was relatively short (September through December of 2012).

4.6.2 External validity

The issue of external validity threats or the ability to generalize is related to the tasks, their assignment and the lab setting in general. Findings of a study with a set of assigned tasks may be valid for the context of these particular tasks, but it is difficult to determine whether the same effect would be achieved with different tasks or in different experimental or natural settings. Testing the tasks ahead of time and asking pilot participants which tasks are more realistic and could potentially be real information needs helps deal with this particular threat. Also, picking tasks that have already been tested in another study and have produced satisfactorily generalizable results helps mitigate this threat.

[22] The instant search option is a type of query reformulation, which isn't initiated by the participant and therefore brings noise and variables which are difficult to control, into the observation. It may also distract the participant from focusing on the results that were retrieved by the query that the participant did intend to execute. Same goes for the auto-complete option which was disabled by using the following search page URL: https://www.google.com/webhp?complete=0. This URL was saved as a shortcut in the browser and the participants were instructed to only use the shortcut to navigate to the search engine. As long as they started their search with this shortcut, all the queries, including the reformulations were written without the auto-complete function.

Choosing the participants also affects the extent to which the study can be representative of the real population. Ultimately, to get a representative sample, random sampling from the search engine users' population is needed, but this is hardly possible due to resource constraints. A more realistic option was to recruit the participants for this study from a population of university students, which limits subjects' variability. However, the claim for generalizability is stronger if stratified sampling is applied and participants can be sampled from different schools and majors. Therefore, to deal with this threat, the participants were recruited from different universities, home schools, majors, and graduate/undergraduate levels.

Threats to reliability are related to the fact that the interviews are conducted by a human being and therefore differences in the way briefing or questioning are performed, may exist from one session to the next. This could result in a lack of consistency in the gathered data across participants. This is why, in order to deal with this, I made sure that the sessions were performed in a similar environment and the scripts used for briefing and questioning were followed closely.

4.7 Initial data preparation and analysis

4.7.1 Cue elicitation and coding

The exploratory nature of this study suggests an iterative and inductive approach to the identification of cues present in the data and the development of a coding scheme for those cues. The collected data was very rich in detail due to its video and audio content and required human researchers and coders in order to develop the coding scheme, code the recordings, and validate the coding. Both the video and the audio were used in coding scheme development and the assignment of cues, once the scheme was in place. Initially, anything that the participant talked

about leading up to a reformulation, could be developed into a cue, but as cue development progressed, there was a need to identify recurring themes which eventually turned into cues.

I performed the initial inductive cue and coding scheme development and the scheme was later validated by two coders. I started systematically watching all the videos containing the observation material, including the participants' responses to the questions asked by the observer (the protocol can be found in Appendix 9). While listening to the participants' responses in the first 10 videos, I singled out the recurring cues and added a description for each cue. Some cues were discovered later in the process of watching those 10 videos. For example, cues that didn't appear until the videos of participant 6 or 7 and therefore were added to the scheme later. I based my cue identification efforts on the answers to the three questions asked by the observer ("why?", "how?", "what exactly?") before each reformulation point. I was operating under the guidance that a cue may be any element on the SERP or a full page that the participant paid attention to and which triggered query reformulation, similar to the potential cues described in the list of possible cues in section 1.3. Each cue assignment was accompanied by a quote of the participant which related to that cue. These quotes and their timestamps were included in the spreadsheet which contained data on all the participants, tasks, queries, and cues.

This was the method for creating the initial coding scheme. I then watched the first 15 videos again (10 videos for the second time and 5 for the first time) to see if I could identify the cues from the coding scheme again, code cues that were added to the scheme later in the process (for example, cues that were discovered for the first time in the videos of participants 6 or 7 and had to be applied for participants 1-5 as well), and check whether new cues were found in the additional five videos.

Three additional cues (O, S, and X) were discovered in the 5 videos that were watched for the first time. These 5 additional videos were coded with the cues developed from the first 10 videos and the 3 new cues discovered. After the discovery of 3 additional cues, I went over the 10 initial videos to make sure that those 3 newly discovered cues did not occur in these videos. At this point I decided to keep the cues which didn't appear frequently, because they could still appear in videos which hadn't been coded yet or could potentially be merged with other cues later on. I continued to code all 33 videos. No new codes were discovered in this subsequent phase.

The coding procedure described above yielded a cue (or a set of cues) that preceded each reformulation category. Overall, 19 different types of cues (Table 3) were discovered, but some of them were infrequent and were either combined together or discarded. For example, each of the cues: L, N, M, O, T, S, V, and X had less than 3% frequency relative to the total number of cues. Also, cues P, Q, R were combined into a new Z cue, because all three were representing "different meaning" of the query term and because on their own, P, Q, and R had low frequency.

TABLE 3: CODING SCHEME FOR THE CUES

Cue label	Cue description [Abbreviated description in parenthesis]	Cue explanation and examples
F	scanning for and unable to find certain word/s even though it was not explicitly part of the query [expecting a word]	The participant reports she was looking for/expecting/hoping to find a certain word/words that did not appear in the preceding query. For example: names of people, year, date, list of sport names, list of mineral names.
G	words that were not part of the query but appear in the results and indicate the character/type of the	The participant indicates that the snippet or the full page suggest that this page belongs to a certain general

	page [unwanted page type]	unwanted category of pages that will not lead to the answer. For example: talks about current events instead of history, definitions instead of history, travel websites instead of technical websites, websites that contain information that is too technical, forums, q&a website.
H	word/s that don't appear in the query but appear in the results and provide an idea for a change in the query [provide an idea]	The participant says that she saw a particular word (or words) in the results and that gave her an idea for the next query. The participants should mention the exact word/s that triggered the reformulation. For example: saying that they saw 'end quench test' together with jominy and that's what gave them an idea to add 'end quench test' to the next query.
I	query terms spread out [terms spread out]	The participant indicates that the query terms are scattered and appear separately, when she expected them to appear together. For example: saying that they are too disconnected or appear in separate sentences or paragraphs.
J	similar results as before [similar results]	The participant says that she has already gotten these same results in previous queries. For example: "I was getting the same pages as before".

K	query terms appear together with unwanted word/s which were not part of the query [unwanted words]	The participant says that there are certain word(s) in the results which they didn't want to see there. For example, when they say that they wanted to get away from something or that there were some words that they were not interested in seeing. For example eggs or grapes floating (instead of minerals).
L	focusing too much on a certain term/phrase [too much focus]	The participant says that a certain term/phrase is too dominant and appears more than desired. For example, saying that there was too much emphasis on salt water.
M	the results are too all over the place or too spaced out [results too diverse]	The participant reports that the results are 'too over the place' or 'too spaced out' to indicate that they cover content that is too diverse. For example: saying that the results were "everywhere, got stuff on the dead sea, on science experiments using salt, it was all over the place."
N	implying a different relationship between two terms [different relationship]	The participant indicates that the relationship between two terms was opposite from the desired one. For example: Australia adopted something from Britain as opposed to Britain adopting from Australia.
O	different order of query terms than in the query [different order]	The participant reports that the query terms appear in the results in an order different than they appeared in

		the query. For example: 'cricket' and then 'adopted' and then 'British' in the results for a query "Britain adopted cricket from Australia".
P	different meaning of one of the query terms (because of verb vs. adjective) [different meaning]	The participant reports that a certain word is not suitable because it has a different meaning than intended as a result from it being an adjective and not a verb as expected. For example: adopted star vs. was adopted yesterday.
Q	different meaning of one of the query terms (because of verb vs. noun) [different meaning]	The participant reports that a certain word is not suitable because it has a different meaning than intended as a result from it being a noun and not a verb as expected. For example: player vs. played.
R	different meaning of one of the query terms (due to it being next to a word that changes its meaning) [different meaning]	The participant reports that a certain word is not suitable because it has a different meaning than intended as a result of it being next to a word that changes its meaning. For example: time it takes to format vs. time it was last formatted
S	no results found [no results]	The exact query yields no results at all (as indicated by Google and verbally by the participant).
T	too many results/list [too many results]	The participant indicates that there was too much information to deal with or too much to read through. For example: mentioning that the list

		was too long to go over.
U	found an intermediate answer [intermediate answer]	The participant says that they found an answer to an intermediate question they had during the search (the answer was not an answer to the original question, but answered a certain claim they had during the search). For example: they wondered whether rugby was originated in Australia and found that it wasn't.
V	one or more of the query terms doesn't appear frequently enough [not frequent enough]	The participant says that a certain term from the query is not dominant enough, that it doesn't appear as frequently as they would like it to be. For example: the word 'inventor' appears in the query, but doesn't appear in the results at all or appears very infrequently.
W	no cues [no cues]	There were no evident cues mentioned by the participant, saying that they just used their prior knowledge, a hunch, or they couldn't remember what the cues were.
X	not authoritative results [not authoritative]	The participant didn't think that the previous results came from credible sources. For example: a certain website is not authoritative because it's a forum and not an academic source.

The unit of observation in this study is on the query level, from the moment that a query has been entered (either initial or the previous query) and until the query has been changed. The questions about the participants' decision making process revolve around each query reformulation point and everything that leads to it. On the one hand, this is the smallest unit that can be observed with the given design, but on the other hand it also contains all the cues that led to the subsequent reformulation and which are later elicited in the analysis stage. At this stage, the unit of analysis is on cue level. This is the smallest unit that can be discovered with the given collected data and once elicited, an association between a single cue or a set of cues and a subsequent reformulation can be made. Therefore, the most suitable unit of analysis for the research questions is on the cue level, although some of the variables which were used to answer the research questions are on the participant level or the task level, such as search engine expertise or domain knowledge and answer correctness, respectively.

4.7.2 Categorizing query reformulations

Query reformulation comprises the changes made between one query representation and the next. Current web search engines offer query (re)formulation aids. Google's auto-complete is one example of an algorithm that works as the searcher types in her query and offers additional terms that might be similar to the initial query ("Google Autocomplete," 2012). Its purpose is to allow faster query input or to catch common mistakes, but it may also suggest suitable query terms that the searcher may adopt as part of her query (re)formulation. Google's "related searches" is a type of query suggestion and is similar to auto-complete, because it helps the user to expand the query with suggested terms. Instead of being suggested as the query is being typed, it is presented at the end of the search results with additional or alternative terms to the original query. Both aids generate suggested terms automatically, based on similarity to the original

query and on popularity of query terms in other searchers' search history (Burke, Fan, Wada, Malhotra, & Coe, 2008; Sahami & Heilman, 2011; Sareen, Kumar, Yu, & Wang, 2008). These tools are intended to assist in query (re)formulation and on many occasions they do trigger query reformulation or help the searcher come up with a term that she had not thought of originally. These aids, however, are based on query term popularity among other searchers and they don't take into consideration factors such as the elements that the searcher pays attention to when reformulating a query. Query reformulations that stand at the center of this research are those that are made without the help of (re)formulation aids and are only triggered by the searcher's attention to elements on the SERP or the full webpage.

In this study, each reformulation was categorized into one of several possible categories. The first reformulation category in the search sequence was always assigned starting with the second query in the sequence, representing the transition between the first and the second query. This way, reformulation categories always represented the transition between each query and the one preceding it. For categorizing each query reformulation, a category scheme was developed based on earlier research. The first classification scheme was borrowed from Huang and Efthimiadis (2009), whose scheme was already mentioned in the literature review section 2.4. They utilized the following categories in their research and used a Python script to categorize their data into the following reformulation categories:

TABLE 4: ORIGINAL REFORMULATION CATEGORIES FROM HUANG AND EFTHIMIADIS (2009)

Reformulation Category	Description
Word Reorder	The words in the initial query are reordered but unchanged otherwise, producing the reformulated query. Example: seattle pizza palace --> pizza seattle palace

Whitespace and Punctuation	The reformulated query is a whitespace and punctuation reformulation of the initial query if only whitespace and punctuation are changed in the reformulation. Example: wal mart, tomatoprices --> walmart tomato prices
Remove words	When any number of words is removed from the initial query resulting in the same words in both queries. This reformulation neglects word order. Example: yahoo stock price --> price yahoo
Add words	When one or more words are added to the initial query. This reformulation applies even if words are reordered in the reformulated query. Example: eastlake home --> eastlake home price index
URL Stripping	If the initial query is a URL and the reformulated query is the URL stripped of its URL specific strings. Example: http www.yahoo.com --> yahoo
Stemming	This reformulation involves changing the word stems in the initial query. Example: running over bridges --> run over bridge
Form Acronym	When the reformulated query is an acronym formed from the initial query's words. Example: personal computer --> pc
Expand Acronym	When the first query is an acronym and the reformulation is a query consisting of the words that form the acronym. Example: pda --> personal digital assistant
Substring	An instance where the reformulated query is a strict prefix or suffix of the initial query. Example: is there spyware on my computer --> is there spywa
Superstring	An instance where the reformulated query contains the initial query as a prefix or suffix. Example: nevada police rec --> nevada police records 2008
Abbreviation	When corresponding words from the initial and reformulated queries are prefixes of each other.

	Example: shortened dict --> short dictionary
Word Substitution	When one or more words in the initial query are substituted with semantically related words, taken from the Wordnet database. Two words are related if one is a semantic relation (synonym, hyponym, hypernym, meronym, or holonym) of the other after both are converted to their base morphological form. Example for a synonym: easter egg search --> easter egg hunt

The rule for this categorization scheme was that anything that didn't fall into one of the above categories, was considered to be a new reformulation. Given the nature of the reformulations performed by the participants in this study, there were several categories that didn't apply to any of the reformulations (such as URL stripping, form acronym, expand acronym) and the number of reformulations that fit into the "new" category was quite large (more than half). As a result, I tried to add more categories that would be more suitable to the nature of the type of reformulations appearing in this study. This was in an attempt to accommodate the type of reformulations performed by the participants, but not covered by the original scheme. Another reason was to create a more lenient version of a category that already existed in the original scheme, such as Word Reorder which allowed for queries that didn't contain the exact same query terms, but did have a significant overlap and reordered overlapped terms. The original Python script that was borrowed from Huang and Efthimiadis (2009), was modified to include these additional categories:

TABLE 5: ADDITIONAL CATEGORIES ADDED TO THE SCHEME

Word Replace	When the shared words in both queries have the same order AND the number of replaced terms in the previous query minus the number of replaced terms in the reformulated query is 3 or less AND the proportion between the shared words that stay together vs. all the words in the query is at least 0.15 (different proportions were attempted).

	Example: partial pressure of oxygen at 5000 feet--> partial pressure of oxygen at various altitudes
Phrase Formation	When double quotes are added around at least two words that appear in the initial query, in order to retrieve web pages that contain these adjacent query terms. Example: what minerals can float in salt water --> minerals float "salt water".
Word Reorder	When there are at least 50% common words in both queries and those common words have been reordered. Example: formatting hard drive 101 information --> information of hard drive on pc

These additional categories helped reduce the number of reformulations that were classified as "new", but the new class still accounted for, around 30% of all reformulations. Also, the following categories stemCompare, substring, superstring, whitespacePunctionation, wordReplace, wordSubstitution applied to very few reformulations (1-3). Therefore a new, more concise, reformulation classification was adopted from Liu et al. (2010):

TABLE 6: REFORMULATION CATEGORIES DESCRIPTION FROM LIU ET AL. (2010)

Reformulation Category	Description
Generalization	Qi and Qi+1 contain at least one term in common; Qi+1 contains fewer terms than Qi Example: “harmful chemicals in food” --> “chemicals in food“
Specialization	Qi and Qi+1 contain at least one term in common; Qi+1 contains more terms than Qi Example: “2007 car” --> “2007 car sales”
Word Substitution	Qi and Qi+1 contain at least one term in common; Qi+1 has the same length as Qi, but contains some terms that are not in Qi. Example: “castle in canada” --> “fortress in canada”

Repeat	Qi and Qi+1 contain exactly the same terms, but the format of these terms may be different Example: “Denmark fortress” --> “fortress, Denmark”
New	Qi and Qi+1 do not contain any common terms Example: “anthill” --> “ant bites”

With this classification scheme the new category wasn't as large, its proportion was only 15%, however, the "repeat" category contained only two reformulations. Therefore, since "repeat", in the way it is defined above and the way it is manifested in those reformulations, is the closest to "substitution", these two reformulations were folded into "substitution" as well. There were no other reformulation categories that applied to such a small number of reformulations. In addition to these four categories, I added another category from the previous scheme: phrase formation. This category is distinguishable enough from the 4 other categories and is a common reformulation technique, which was used by some of the participants.

4.7.3 Coding verification

When this research was proposed, the initial idea was to let external judges (coders) inductively develop the cues and assign them to all the reformulations. However, once the data collection was over and I reevaluated its richness and quantity, this option seemed unfeasible. There is a trade-off between the best possible procedure and what can be done given limited resources. Requesting external coders to develop these codes based on many video and audio data files and then assigning the codes to all reformulations, would have required a lot of time on the coders' side and resources on my side. Therefore, there was a need for a compromise, where the coding scheme would be developed by me, but multiple coders can reliably identify and assign the same codes. One advantage of this method is that the cues could be discussed and if there was a

systematic problem with a definition of the cue, the existing coding could be systematically corrected, due to the quotes and the time stamps that accompanied the coding.

Consequently, the next step in this study was to determine if the coding scheme developed by me was also usable by other people and to measure the degree to which my coding aligned with the coding of an external judge. The unit of analysis for the verification was performed on the participant-task level, meaning that the tasks for which coding was to be verified, were chosen from a pool of all the participant-task pairs (33 participants times 5 tasks = 165 participant-task pairs). Two coders[23] were recruited for this work. This way, each coder's work could be compared to my coding, but also, they could be compared to each other in case there is high disagreement between me and both coders. In a case like this, it allowed to check whether the disagreement is mostly among myself and the coders or there is still agreement between both of them in spite of disagreement with me.

Ideally, all the instances coded by me should have been verified by an external judge, but given budget limitations, a sample of participant-task pairs was randomly selected. When deciding how many participant-task units to select, several considerations came into play. General guidance regarding the possible range of subsample size was taken from the Content Analysis Guidebook (Neuendorf, 2002), which states: "if one could attempt to make a general statement from the accumulated knowledge so far, it would be that the reliability subsample should probably never be smaller than 50 and should rarely need to be larger than about 300". Given the limited resources, I was able to fund coding verification of 50 distinct tasks between the two coders, which is within the recommended range, but on its lower side.

[23] One coder was a graduate student at the School of Information Studies at Syracuse University and the other coder was a senior undergraduate student at the Department of Communication at Cornell University.

The tasks to be coded, were randomly sampled from the list of participant-task pairs which had at least one reformulation. Nine tasks were selected separately (not randomly, but rather based on their instructional value) and used for training. The actual set to be verified contained 30 tasks in total, 10 of which overlapped, meaning that both coders coded 10 identical tasks and 20 different tasks. From the tasks used for training (not part of the 30 tasks in the actual set mentioned above), 4 tasks were used for initial training and 5 more (3 of them overlapping between the two coders) were used for a more advanced training. Appendix 13 contains the table with the participant-task pairs used for coder training and coding verification. The participant-task pairs that were used for training (either initial or advanced) were not taken into account in the agreement calculations.

The purpose of my first meeting with the coders was to introduce the cues and their descriptions/definitions. Both coders were present at the meeting, which lasted about 3 hours. They read the training document (Appendix 12) together with the coding scheme (Table 3) and afterwards I verbally repeated the instructions in the document. I emphasized the importance of understanding the descriptions in the coding scheme and asked whether they had any questions or clarifications. I also showed them the coding template which contained all the sampled tasks and their reformulations. Then I explained how to add the cues, the reasoning behind them, and the quote/timestamp that referred to the cue. Next, I answered their questions either regarding the cues or the template.

During the second part of the meeting, the coders coded 4 training tasks (the participant-tasks pairs picked for initial training). After each task, each of them presented their coding and we discussed any disagreement among them and compared to my existing codes. We also tried to pin point misunderstandings of cue descriptions and I answered any questions they had. At this

point, I also revealed my coding for these tasks. No inter-coder reliability was calculated at this stage, but as mentioned above, any disagreement was discussed in order to identify the source of disagreement. One of the purposes of this discussion was to bring up cases in which the description of the cues was misunderstood by the coder and required clarification to prevent this from reoccurring. We discussed the reasoning behind each cue, whether the coder agreed on it or not and put a lot of emphasis on the disagreement and tried to resolve it. At the end of the meeting, the coders were assigned the list of tasks they needed to code in the upcoming month and were asked to code 5 advanced training tasks before the next meeting.

The second coding meeting was conducted with each coder separately, after the coders completed coding of 5 advanced training tasks. During this meeting and all the others that followed, I went over their coding and listened to their reasoning behind each code. If there was a misunderstanding of the description of the cue, I explained it to them and how their interpretation differed from the intended meaning of the cue. During the second meeting, which was still considered training, these discussions were very elaborate and I revealed my coding for those tasks by the end of each task discussion. At the end of the second meeting, the coders were assigned 5 tasks from the actual set to code before the next meeting. From the third meeting and on, the coders were working on the actual set of tasks to be verified and therefore, they were not exposed to the actual codes that I had assigned. During these meetings, the discussion mostly revolved around the coders' reasoning, in an effort to understand whether it was consistent with the description of the cue they assigned. After this discussion, it was up to them whether they wanted to change their coding or stick to the original coding.

Following are some examples of incorrect interpretation of cues by the coders that had to be clarified:

One of the coders thought that when the participant was talking about getting away from "sports minister", it was because the word "sports" had a different meaning, but actually, it was an unwanted word -"minister" next to the word "sports", because "sports" still had the same meaning, but was part of a phrase together with "minister" and they wanted to get away from "minister".

The coders thought that when the participant talks about getting away from definitions, they should be explicitly referring to seeing the word "definition" actually on the full page or snippet, but according to the description of this cue (G [unwanted page type]), the word "definition" doesn't have to be part of the page, but it can rather be a general characteristic of the page. For example, it could be a wikipedia page containing a definition of what the jominy test is, but not necessarily contain the word "definition" in it.

At the same time, the coders also weren't sure whether when talking about the cue K ("query terms appear together with unwanted word/s which were not part of the query"), the participant should explicitly state the word that was unwanted. As part of the clarifications and discussions, I pointed out to them that when talking about an unwanted word, the participant should explicitly mention which word they were trying to get away from.

Another example of clarification that I had to make for the coders was regarding the "no results were found" (S) cue. Because the participants indicated quite often that they were not finding the answer (a point at which I usually probed further to figure out how they knew that they didn't get the answer), sometimes the coders interpreted this as "no results found". This is when I had to clarify that "no results found" was meant that the search engine returned no results at all and that the participant verbally indicated that the list of results was empty.

Also, after discussing the codes with the coders and hearing their input, I realized that some of my coding required a systematic change. The quotes of the participants provided a good basis to make a decision where the systematic change should be made. As a result, I changed certain instances from K (query terms appear together with unwanted word/s which were not part of the query) to G (words that were not part of the query but appear in the results and indicate the character/type of the page) because I realized that "how to" pages fit more to the description of cue G, since they belong to a certain type of pages, are of a certain character.

4.7.4 Measuring coder agreement

The coding method in this research allowed for each reformulation to have more than one cue assigned to it. This was also the instruction that was given to the coders that were supposed to verify my coding. They were told to add cues to a reformulation as long as the participant was describing different cues that referred to that reformulation. This influenced the way agreement between coders could be measured. There are two possible approaches. One way that would be more strict would be to require agreement on both the number of cues on each reformulation and the cue values. For example, in that case, if one of the coders has one or more extra cues in a particular reformulation, then all those extra cues would be considered as disagreement, even if all the common cues are in agreement. For instance, if one coder had the cues G and R and the other only R, then the fact that the other coder didn't think that G was also a cue for this reformulation, then this would be considered as a disagreement. Another way, which would be less strict, is not to require agreement on the number of cues, rather only on agreement of the minimum number of cues among the two coders. For example, if one coder had the cues G [unwanted page type] and R [different meaning] and the other only R, then it would be

considered as agreement in the less strict method, because they agreed on R, even though one of them thought that this reformulation was supposed to have G as well.

I calculated the agreement between myself and coder 1, myself and coder 2, as well as between coder 1 and coder 2. I measured agreement for the more strict and the less strict approaches by calculating the following measures using an online utility called ReCal2[24] (reliability for 2 coders): percent agreement, Scott's Pi, Cohen's Kappa, and Krippendorff's Alpha (Freelon, 2013). Krippendorff's Alpha is an attractive co-efficient, because it accounts for chance agreement and the magnitude of the misses, adjusting for whether the variable is measured as nominal, ordinal, interval, or ratio. However in the past it had been rarely used because its calculation is a rather tedious process (Neuendorf, 2002). The availability of online tools that calculate all these measure, made them more accessible and easy to compare to other studies. Table 7 presents the results of the comparison measures:

TABLE 7: CODER AGREEMENT MEASURES

Type of comparison	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha
Coder 2 & me - agreement on number of cues required (more strict)	80.82	0.7834	0.7835	0.7842
Coder 2 & me - agreement on number of cues NOT required (less strict)	92.91	0.9189	0.9189	0.9192
Coder 1 & me - agreement on number of cues required (more strict)	76.92	0.7409	0.7420	0.7416
Coder 1 & me - agreement on number of cues NOT required (less strict)	91.50	0.9035	0.9035	0.9038
Coder 1 & Coder 2 - agreement on number of cues required (more strict)	62.86	0.5870	0.5902	0.5900

[24] http://dfreelon.org/utils/recalfront/recal2/

Coder 1 & Coder 2 - agreement on number of cues NOT required (less strict)	88.00	0.8659	0.8661	0.8672

As expected, the less strict method of comparison resulted in higher values (percent agreement ranging from 88% to 93%) than the more strict comparison. The comparison of the more strict method, which required agreement on the number of cues, achieved lower values (percent agreement ranging from 63% to 81%). Lower agreement was observed between me and each of the coders and even worse so between the two coders.

One possible explanation for the low agreement between the two coders could be related to an observation that coder 1 tended to assign slightly more cues than me to each reformulation, while coder 2 tended to assign a bit less cues than me to each reformulation. This means there was a larger gap between the two coders in terms of cues per reformulation, than between me and each of the coders. For example, coder 1 assigned overall 180 cues, while I assigned 177. Coder 2 assigned overall 138 cues, while I assigned 140. Looking only at the overlapping participant-task pairs, out of 38 reformulations, coder 1 had 4 reformulations which had more cues assigned by her than by me vs. 1 that had more cues assigned by me than her. At the same time, coder 2 had 5 reformulations which had more cues assigned by me than her vs. 3 that had more cues assigned by her than me.

For the more strict method, which required match in the number of cues, these differences in the number of assigned cues may have caused better measures between each coder and myself, but worse measures among coder 1 and coder 2. Since the number of cues that each coder could elicit from participant explanations depended on the coders' interpretations on what should be considered a single cue, this number varied. It depended on to how many cues the coder decided to break down each sentence of explanation to. For example, the quote "it's still talking about

Australia, Australian culture, it's still not talking about Britain at all" was assigned by me to one cue "one or more of the query terms doesn't appear frequently enough". However, in addition to this cue, coder 2 also assigned the cue "query terms appear together with unwanted word/s which were not part of the query" to the same quote. For the less strict method of comparison, all the measures were higher than 0.8, which indicates high degree of agreement. This means that while the coders were agreeing on the actual codes they assigned to the reformulations, they didn't agree as much on the number of cues that should be assigned to each reformulation. In any case, both for the strict and the less strict method, the values for all the measures are in the 'good' to 'excellent' range and are comparable to other studies that used these measures.

Another possible explanation for the low agreement between the two coders, but higher between myself and the coders, could be the lower number of participant-task pairs that are overlapping between the two coders. There were 10 participant-task pairs that overlapped and compared between the two coders, while 30 participant-task pairs which were compared between each coder and myself. The lower agreement measures between the two coders could be due to the lower number of the overlapping participant-task pairs, which means that measures based on these 10 pairs were more sensitive to disagreement than the measures based on the 30 pairs used when comparing the coders to myself.

The decision whether or not these results are at an acceptable level must be made depending on the costs of drawing invalid conclusions based on these data. When human lives are at stake, the criteria for rejections have to be set far higher than when a content analysis is intended to support scholarly arguments (Krippendorff, 2004). For the latter case, Krippendorff suggests the data are at least similarly interpretable by other coders. In his writings, Krippendorff offered to use 0.8 and up, but also stated that tentative conclusions are still acceptable at 0.667 and up. Some

researchers applied the criteria for Cohen's Kappa (below 0.45 ‘poor’, 0.45 to 0.59 ‘fair’, 0.60 to 0.74 ‘good’, and 0.75 and above ‘excellent’) (Neuendorf, 2002) to Kripendorff's alpha results (Strijbos & Stahl, 2007). In their study, which explored methodological issues in developing a multi-dimensional coding procedure for small group chat communication, Kripendorff's alpha ranged from 0.367 to 0.857 and Cohen's Kappa ranged from 0.382 to 0.835, depending on the dimension and the pair of coders.

4.7.5 Categorizing the answers and search engine expertise

In order to measure the effectiveness of each search session, the answers reported by the participants had to be evaluated for their correctness. I evaluated the answers based on the answers deemed as true and false by Ageev et al. (2011) and also made sure that the URLs the participants saved, actually contained the answer they were reporting, by following the pages the saved URLs lead to. The web pages indicated by the participants didn't have to be exactly the same as the ones in Ageev et al. (2011), only the answers needed to be correct and the provided pages were expected to contain the answer. For questions which allowed more than one correct answer (such as the one with Australian sports), if the answer did not appear on the list of correct or incorrect answers by Ageev et al. (2011), they were evaluated by me, where I made sure that the answer indeed answers the question correctly (for example, that the sports suggested was indeed Australian). Tasks which were left without an answer, were also deemed as incorrect. When this research was first proposed, it was considered to incorporate into the measure of the level of effectiveness the number of sources containing the answer to each question, which was supposed to be provided by the participants. However, even though they were requested to do so, when the participants were able to find an answer to the questions, they provided only one source

with the answer and moved on. It could be due to the fact that the questions were too difficult and the participants were too tired by the time they had found an answer.

The post-questionnaire, which was filled out by the participants after the search session, also contained their self reported domain knowledge on each domain that related to each of the five questions they were asked to find answers to. In addition to that, there were three questions in the questionnaire that were supposed to assess the participants' mental model, which would in turn represent their search engine expertise. I evaluated their answers to the mental model questions and graded it on a scale of 1 to 10. As mentioned above, when evaluating their answers to the first two questions, I was expecting to see explanations that mentioned query terms and in the third question ("what does the system DO") how they are utilized in the search algorithm. The closer the explanation was to the way a basic search engine utilizes query terms to rank the results and the more detailed it was, the higher was the score of the participant's search expertise. Both evaluations were performed blindly, meaning no identification information was visible on the questionnaires or the answers.

4.7.6 Data set preparation

Once the analysis and manipulations described above were completed, I had a dataset which was almost ready for discovery of relationships and patterns, as were presented in the research questions. At this point, each row in the dataset represented a single cue per reformulation and additional steps were taken to represent the data in a way that would make it more suitable for data analysis. First, each cue and query reformulation category were represented as dummy variables (in other words - binary variable in which "1" meaning this cue appeared prior to the particular reformulation, "0" meaning this cue didn't appear in that reformulation). Following is a list of fields in the dataset:

TABLE 8: DATASET FIELD DESCRIPTION

Field	Description
Participant identification number	A number which represents the ID of a participant in the study
Question code	Represents the 5 different tasks: A - minerals that float, B - Australian sports adopted by Britain, C - partial pressure, D - Jominy test, E - finding time of hard drive format.
Question number	The number of the question, based on its order as it was presented to the particular participant (the questions were rotated).
Query number	The number of the query in that particular search sequence.
Cue number	In cases where there was more than one cue per reformulation, this was the indicator for the cue number in that particular search sequence.
Cue	The cue that was assigned to that particular reformulation. If more than one cue was assigned, another row in the dataset was created, with all the other data duplicated. The values that can appear in this field are: F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X
F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X	Each of these is a field which represents whether or not (0 or 1) each of these cues occurs prior to that reformulation.
Reformulation category	The category that was assigned to that

	particular reformulation. The values that can appear in this field are: new, generalization, substitution, repeat, specialization, and phrase formation.
new, generalization, substitution, repeat, specialization, phrase formation	Each of these is a field which represents whether or not (1 or 0) that reformulation category was assigned to that reformulation.
Search Expertise	A score that was yielded from the post-questionnaire, based on the mental model answers.
Domain Knowledge	Self reported domain knowledge score as it was recorded in the post-questionnaire.
Search Effectiveness (Answer correctness)	Whether or not the answer to the question was correct (1) or incorrect (0), represents the effectiveness of the search. The correctness of the answers was determined based on a list of correct and incorrect answers provided by Ageev et al. (2011). This variable is on a task level, meaning that search effectiveness is equal to 1 for the whole task if the question was answered correctly and 0 if it was answered incorrectly or not answered at all.

Second, for reformulations which had more than one cue, the rows containing those cues were aggregated together. As a result, the dummy representation of cues and their aggregation allowed more than one "1" value in each row. The number of "1"s was equivalent to the number of cues in that reformulation.

Also, some of the cues were combined together because of infrequent occurrence and similar definition. Cues that had to do with different meanings of query terms (P, Q, R) were combined into Z. Cues P, Q, and R represented occurrences in which the participants reported that a certain word is not suitable because it had a different meaning than intended as a result of 1) being an adjective and not a verb as expected; 2) being a noun and not a verb as expected 3) being next to a word that changes its meaning (respectively). These three cues were combined into a single cue (Z) which represents different meaning of query terms. Only the following fields remained after aggregation and combination of P, Q, R: participant ID, question code, question number, query number, search expertise, domain knowledge, answer, F, G, H, I, J, K, L, M, N, O, S, T, U, V, W, X, Z, new, generalization, substitution, specialization, phrase formation.

This data representation allowed me to discover the pattern, or in other words, the similarity between participants in the cues that they pay attention to and utilize for the purpose of reformulation. It allowed me to map the common cues and their frequency over all the subjects. Second, it also allowed me to identify the relationships between the set of cues (independent variable) and the reformulation categories (dependent variable) that they precede. Also, establishing the relationship between the set of cues (independent variable) and the effectiveness of the search (dependent variable) can also be done by using the above described dataset. Finally, adding search expertise and domain knowledge as the mediating variable for the above relationships can be performed with this dataset structure.

4.8 Data analysis

The research questions guiding this study are set out to explore the relationships between cues and other variables such as query reformulation categories and search effectiveness. This can be

done through several methods and in this section I will describe the methods I used to determine these relationships.

4.8.1 Conditional probability of a reformulation given a cue

The very first step of getting a sense of the pattern on which cue leads to which reformulation, can be achieved by a conditional probability of reformulation given a cue. However, since some reformulations are more likely to occur than others, there is a need to adjust for expected values of reformulation categories. One way to do it is to normalize the conditional probability by the probability of a reformulation. In other words, the ratio would be P(reformulation|cue) / P(reformulation). Since P(reformulation|cue) is equal to P(reformulation and cue)/P(cue), the ratio would be P(reformulation and cue)/(P(cue)*P(reformulation)).

The resulting value provides a measure of how much more or less likely it is to observe the reformulation-cue pair than it is to observe the reformulation in the overall population. For example, if the ratio is 2, then the reformulation is twice as likely to occur with the cue than it is in the overall population. If the ratio is 0.5 then the reformulation is half as likely to occur than it is in the overall population. The problem with this method is that there is a need for some kind of ratio threshold that would help determine which cues are likely to lead to a certain query reformulation. If the threshold is too high, then there may be several reformulation categories which will have no cues and if it's too low, then there will be too many cues leading to those categories. In addition, this method also does not take into account the differences between the participants and the tasks they were assigned to perform. Next, I will present a statistical approach that can address these issues as well and compare whether or not there is consistency in the results.

4.8.2 Mixed effects models

An alternative method to discover the pattern in the data, is to use a linear model to express the relationships in terms of a function. In other words, when trying to answer the first research question (RQ1) the idea is to predict each reformulation category (dependent variable) with cue values (independent variable). In the second research question (RQ2), the independent variable is the same, while the dependent variable is search effectiveness. One problem in attempting to fit a linear expression is the categorical nature of the dependent variables (reformulation type and search effectiveness). Therefore a logistic regression was a more suitable choice for analysis of this data.

The problem with regular logistic regression is its independence assumption, which is violated, because each participant has multiple observations (unit of observation being query reformulation). Every person has a different way of interpreting results and making sense of them and this is an idiosyncratic factor that affects the reformulations from the same participant. This is also what makes the multiple responses (reformulations) by each participant inter-dependent rather than independent. In addition, reformulations from different tasks are nested within each participant and each task has its own characteristics and difficulty. Therefore, similar to the case of by-participant variation, I also expected by-task variation among the reformulations. The way to deal with this situation, was to add a random effect for participant and task type. A random effect is generally something which is expected to have a nonsystematic, idiosyncratic, unpredictable, or "random" influence on the data. In many studies this is often "subject" and "item" and usually the researcher would want to generalize over the idiosyncrasies of these individual subjects and items. On the other hand, fixed effects are expected to have a systematic and predictable influence on the data. Adding random effect

allowed to resolve the non-independence by assuming a different "baseline" search style for each participant and task. This meant modeling the individual differences due to participant and task type by assuming different random intercepts for each participant and task type (Winter, 2013).

The regular linear models are "fixed-effects-only" models that usually have one or more fixed effects and a general error term "ε". With a linear model, the world is divided into things that are somewhat systematic (the fixed effects or explanatory variables) and the things that cannot be controlled for or cannot be understood (ε). In this type of model, this latter part, the unsystematic part of the model, does not have any interesting structure and is considered to be a general error term. In mixed models, however, one or more random effects are added to the fixed effects, which provide structure to the error term "ε". This is why this type of analysis is called "mixed" modeling, because in addition to the "fixed effects", the model also contains "random effects" (Winter, 2013). In the model developed for this study, I added a random effect for "participant" and nested within it, is "task". As mentioned before, these "random effects" are the ones that take into account the individual variation that is due to the differences between participants and task types.

In order to adequately represent reformulations with more than one cue, I coded the cues as dummy variables as described in 4.7, each cue had its own column (field). Rows that had more than one cue per reformulation, contained "1"s in those fields which represented the cues that were assigned to that reformulation. This way, the multiple rows that are used to represent multiple cues for the same reformulation could be grouped together and as a result, each row represented a reformulation of a certain participant and a specific task. Each reformulation category was also represented by a dummy variable and the model had to be run for each

reformulation category separately. An example of the relationship between the reformulation category specialization and the cues, without testing for moderation:

$$specialization \sim F+G+H+I+J+K+Z+U+W+(1|participant/task)$$

Where the (1|participant/task) incorporates the random effects of participants and the tasks nested within the participants. The left hand side of the formula represents the response variable and the right side represents the independent variables (fixed and random effects). When testing the relationship between the cues and search effectiveness, a similar formula can be used. Since effectiveness is measured in this study with the correctness of the participants' answers and it is on the task level instead of the reformulation level, the reformulations were aggregated together. Each cue was marked "1" if it appeared in the task and "0" if it was not used in that task at all. The formula for a mixed model trying to predict the effectiveness of the search with cues, is as following:

$$answers \sim F+G+H+I+J+K+Z+U+W+(1|participant)$$

Since all the reformulations were aggregated together, each row represents a task and therefore the random effects are only of participants, not tasks.

4.8.3 Interaction with moderating variables

The research questions in this study also inquired about the moderation of search expertise and domain knowledge. Testing for moderation of search expertise and domain knowledge, means that I would like to know whether the effect of cues on the response (reformulation category or search effectiveness) depends on the level of search expertise/domain knowledge. For example, a possible outcome could be that people with higher search expertise are more likely to use the cue K before the "specialization" category. In order to discover the moderating role of domain

knowledge and search expertise in the relationship between the cues and the response variable, there was a need to introduce interaction terms into the model. The interaction as it is represented in the formula below, is between each cue and the two moderating variables - domain knowledge (DK) and search expertise (SE):

specialization ~ F+G+H+K+W+Z+I+J+U+W+

+F*DK+G*DK+H*DK+K*DK+W*DK+Z*DK+I*DK+J*DK+U*DK +

+F*SE+G*SE+H*SE+K*SE+W*SE+Z*SE+I*SE+J*SE+U*SE +(1|participant/task)

The formula for search effectiveness including interaction terms:

answer ~ F+G+H+K+W+Z+I+J+U+W+

+F*DK+G*DK+H*DK+K*DK+W*DK+Z*DK+I*DK+J*DK+U*DK +

+F*SE+G*SE+H*SE+K*SE+W*SE+Z*SE+I*SE+J*SE+U*SE +(1|participant)

These analyses and some of the data preparation were performed in R, which is a language and environment for statistical computing and graphics[25]. R allows the use of a package, called lme4 (D. Bates, Maechler, & Bolker, 2013), which provides functions for fitting and analyzing mixed models. I used this package in order to fit the models for both the main effects (effects of F, G, H, K, W, Z, I, J, U, W) and the interactions (F*DK, G*DK, H*DK, K*DK, W*DK, Z*DK, I*DK, J*DK, U*DK, F*SE, G*SE, H*SE, K*SE, W*SE, Z*SE, I*SE, J*SE, U*SE).

One preparation that had to be made, was to center the moderating variables (search expertise and domain knowledge) (Cohen, Cohen, West, & Aiken, 2003, p. 301). This is due to the fact that the main effect of a variable in the presence of an interaction is the effect of the variable when the other variable with which it is interacting with equals to 0. In other words, the main

[25] http://www.r-project.org/

effect of each cue is when search expertise and domain knowledge are equal to 0. Since the scale used to represent search expertise and domain knowledge didn't include 0, in order to make the results more meaningful, I centered both domain knowledge (domain knowledge minus its mean) and search expertise (search expertise minus its mean) and used these values instead of the original values. This way, when the model is fitted, the main effects represent a case when both search expertise and domain knowledge are at their average. This is not a necessity but it helps provide a more useful interpretation of the main effects of cues in the presence of interactions. When running a model with interactions, the output includes both the main effects (cues only) and the interactions of search expertise/domain knowledge with the cues.

When only the main effect is significant and does not take part in any interaction, this means that that particular cue has an effect on the response variable (reformulation category or search effectiveness) regardless of the moderating variable (domain knowledge and search expertise). In order to reach a final parsimonious model, there is a need, to "clean" the output iteratively, by gradually removing the insignificant interactions from the model, starting with the most insignificant and until only the significant main effects and interactions remain. The cues that eventually comprise the model are the ones that are significant regardless of any moderating variables (are not part of any significant interactions) and the ones that participate in significant interactions. If a cue participates in a significant interaction, but is insignificant as a main effect, this means that when the moderating factor (search expertise or domain knowledge) is at its mean level, this cue has no significant effect at this level. This means that an interaction between the two variables exists, but not at the mean level of the moderating variable. At this point, there is a need to show at which level of the moderating variable, the cue that participates in the interaction becomes significant. In order to show at which level the main effect becomes

significant, the model also needs to be run with the moderating effects centered around mean plus one standard deviation (high) and around mean minus one standard deviation (low). This allows to show at which level of domain knowledge and search expertise, the cues that take part in the interaction have a significant effect on the response variable. Once the main effects that are part of the interaction are found significant, then the interpretation should be performed only on the interaction level.

If the main effect which participates in the interaction is insignificant at the mean level of the moderating factors, but is found to have a significant effect when the moderating variable is held at one standard deviation above or below the mean, an appropriate interpretation can be made regarding the interaction. For example, I run the model with the response variable reformulation category = "generalization". The outcome shows that cue K interacts with search expertise, but is insignificant when search expertise is at is mean. At this point, I can check what happens when search expertise is held at the mean search expertise plus one standard deviation. If it indeed is significant at that level, then this means that cue K has an effect on "generalization" when search expertise is high.

If the main effect (cue K) is insignificant at all the levels (mean, high, and low), this means that most probably cue K is significant outside of the meaningful range of the moderating variables, when they are too high or too low. For example, they could be significant at 2 standard deviations above the mean, but if this would mean a search expertise value of 13, while the range is only 1 through 10, then this interaction is not too meaningful or useful and shouldn't be entered into the model.

The research design and the analysis presented above were meant to address the exploratory nature of this study and to operationalize the research questions described in section 1.3. The

tasks have been chosen from different sources and tested sufficiently to provide enough confidence that they are suitable tasks for this study. The research questions in this study involve exploration of both systematic cue usage in general and the relationship between cues and other variables such as query reformulation categories, search effectiveness, domain knowledge, and search expertise. Therefore, the rich video and audio data that was collected during the observation stage required extensive work on detection of cues and their coding. In order to confirm that the coding scheme was adequate and can be performed by external judges, verification of the coding was performed and the data was manipulated to accommodate the analysis that would help answer the research questions.

5 Results

This chapter includes descriptive statistics for reformulation categories, query characteristics, cues, and cues in each reformulation category. The results of the systematic cue usage and patterns appear as part of conditional probability analysis results and mixed models results, for both query reformulation category prediction and search effectiveness prediction. This chapter focuses mainly on the plain raw results, while the discussion of these results and their interpretation is to follow in the discussion chapter 6.

5.1 Descriptive statistics

As mentioned above, overall 165 tasks (33 participants x 5 tasks) were performed by the participants in this study. Out of those, 142 tasks had 1 or more query reformulations and the remaining 23 tasks had no reformulations at all. Table 9 shows the frequency of each reformulation category. The "specialization" category was the most frequent and "new" was the least frequent.

TABLE 9: REFORMULATION CATEGORIES FREQUENCY

Reformulation Category	Frequency	Percentage of Total
New	117	15%
Generalization	217	29%
Substitution	117	15%
Specialization	279	37%
Phrase Formation	29	4%
Total	759	100%

The number of cues per reformulation ranged from 1 to 4, where 601 reformulations had only one cue and the remaining 158 reformulations had more than one cue per reformulation (Table 10). For reformulations that had more than one cue, the most frequent pair of cues were F [expecting a word] and G [unwanted page type], which appeared 19 times together. Cues F and K [unwanted words] co-occurred 14 times together. The most frequent cue overall was H [provide an idea], while the most frequent cue only for reformulations that had more than one cue was K [unwanted words].

TABLE 10: NUMBER OF CUES PER REFORMULATION

Number of cues per reformulation	Number of reformulations containing this number of cues
1	601
2	136
3	21
4	1

In terms of the characteristic of the tasks, the average search sequence length (number of reformulations) was 5 and is distributed between the 5 different tasks as depicted in Table 11. It can be noted that the shortest average sequence belongs to task C (partial pressure question) and the longest to B (Australian sports).

TABLE 11: SEARCH SEQUENCE AVERAGE LENGTH (NUMBER OF REFORMULATIONS)

A (minerals)	B (Australian sports)	C (Partial pressure)	D (Jominy test)	E (Hard drive)
4.9	6.6	2.7	4.7	4.7

In terms of query length measured in words, Table 12 shows the distribution of average query length over the tasks. The longest average query was for task E (hard drive) and the shortest was for task D (Jominy test). Naturally, tasks that had longer descriptions, had longer average query length (in words), while tasks that had shorter descriptions had shorter queries on average. For example, the long description of task E ("How do you find the exact time a hard drive was last formatted on a PC?") corresponded to the rather long average queries (7.62) for that task. While and the short description of task D ("When was the Jominy test invented?"), corresponded to the short average queries (3.82) of that task. It is important to note that all these averages are higher than the average web query length of 3.08 reported recently (Taghavi, Patel, Schmidt, Wills, & Tew, 2012). Since the tasks in the current study are intentionally difficult, it is plausible that the queries that they generate would be longer than the average web-query.

TABLE 12: AVERAGE QUERY LENGTH BY TASK

Task	Average of Query Length (words)	Max of Query Length	Min of Query Length
A (minerals)	5.05	10	1
B (Australian sports)	4.16	12	1
C (partial pressure)	6.60	13	2
D (Jominy test)	3.82	10	1
E (hard drive)	7.62	17	1
Grand Total	5.08	17	1

Table 13 presents cue occurrences per task. It can be seen from the table that some cues are more frequent for certain tasks and less frequent or absent for others. For example, the cue U [intermediate answer] did not occur at all in tasks D and E and occurred only once in task C.

TABLE 13: CUE COUNTS PER TASK

Cues\Tasks	A	B	C	D	E	Total cues
F [expecting a word]	37	24	27	50	10	148
G [unwanted page type]	14	34	11	41	22	122
H [[provide an idea]	35	68	18	39	21	181
I [terms spread out]	14	5	6	1	4	30
J [similar results]	9	7	3	14	6	39
K [unwanted words]	72	27	10	27	18	154
L [too much focus]	4	4	1	0	2	11
M [results too diverse]	2	3	0	1	0	6
N [different relationship]	0	15	0	0	0	15
O [different order]	0	1	1	0	0	2
P [different meaning]	0	3	0	0	0	3
Q [different meaning]	0	2	0	3	3	8
R [different meaning]	9	10	0	6	15	40
S [no results]	6	4	0	0	1	11
T [too many results]	1	3	0	1	0	5
U [intermediate answer]	6	21	1	0	0	28

V [not frequent enough]	7	4	2	10	3	26
W [no cues]	35	22	7	23	11	98
X [not authoritative]	3	2	2	0	6	13
Total cues per task	254	259	89	216	122	940

Although no statistical comparison was performed, initial numbers in the tables above do show that different tasks yielded different number of cues, query lengths, and search sequence length. This supports the claim that the analysis should include task as random effects.

Table 14 shows the frequencies of the cues for each reformulation category. As mentioned before, in section 4.7, cues P, Q, R, which represent "different meaning of one of the query terms" were all folded into one cue Z, because P and Q had very infrequent occurrence. Other cues which had a very low frequency (less than 3%) are: L, M, N, O, S, T, V, and X. These cues were not included in the analysis, because such a low sample size for those cues means very low power, which in turn means that it would not lead to significant results. None of these cues were combined into larger categories, because they were all very different from each other and did not form semantically cohesive groups. After filtering out the low frequency cues which were never included in the model, this left only the following cues suitable for analysis: F [expecting a word], G [unwanted page type], H [provide an idea], I [terms spread out], J [similar results], K [unwanted words], Z [different meaning], U [intermediate answer], and W [no cues].

TABLE 14: FREQUENCY OF CUES IN EACH REFORMULATION CATEGORY

Cues	Generalization	New	Phrase formation	Specialization	Substitution	Grand Total	Percentage of total number of cues
F	41	11	1	76	19	148	15.7%
G	38	6	3	53	22	122	13.0%
H	42	57	3	53	26	181	19.3%
I	17	1	7	3	2	30	3.2%
J	13	2	3	15	6	39	4.1%
K	49	12	8	56	29	154	16.4%
L	3	2	1	3	2	11	**1.2%**
M	3	1	1	1	0	6	**0.6%**
N	8	2	1	1	3	15	**1.6%**
O	2	0	0	0	0	2	**0.2%**
Z	25	2	0	15	9	51	5.4%
P	1	0	0	0	2	3	0.3%
Q	4	0	0	2	2	8	0.9%
R	20	2	0	13	5	40	4.3%
S	0	2	5	2	2	11	**1.2%**
T	0	0	0	5	0	5	**0.5%**
U	5	15	0	6	2	28	3.0%
V	9	0	1	8	8	26	**2.8%**
W	22	19	2	39	16	98	10.4%
X	5	0	0	4	4	13	**1.4%**
Total	282	132	36	340	150	940	

Table 15 shows a list of the remaining cues, which were eventually included in the model:

TABLE 15: CUES SUITABLE FOR ANALYSIS

Cue label	Cue description
F	scanning for and unable to find a certain word even though it was not explicitly part of the query
G	words that were not part of the query but appear in the results and indicate the character/type of the page
H	word/s that don't appear in the query but appear in the results and provide an idea for a change in the query
I	query terms spread out
J	similar results as before
K	query terms appear together with unwanted word/s which were not part of the query
Z	different meaning of one of the query terms
U	found an intermediate answer
W	no cues

5.2 Conditional probability results

Table 16 shows the probability of getting a certain reformulation category given a cue, normalized by the overall probability of a reformulation. I selected a threshold of 1.2 and all the ratios that are higher than this number are marked in bold.

TABLE 16: NORMALIZED CONDITIONAL PROBABILITY OF A REFORMULATION GIVEN A CUE

	New	Generalization	Substitution	Specialization	Phrase Formation
F	0.482	0.969	0.833	**1.397**	0.177
G	0.319	1.089	1.17	1.182	0.644
H	**2.043**	0.812	0.932	0.797	0.434
I	0.216	**1.982**	0.432	0.272	**6.107**
J	0.333	1.166	0.998	1.046	**2.013**

K	0.505	1.113	**1.222**	0.989	**1.36**
U	**3.475**	0.625	0.463	0.583	0
W	**1.258**	0.785	1.059	1.083	0.534
Z	0.254	**1.715**	1.145	0.8	0

If a threshold of 1.2 is selected, this means that the reformulation with is at least 1.2 times more likely to occur together with the cue than in the overall population. For example, generalization is 1.72 times more likely to occur with cue Z than it is in the overall population. According to these ratio values and given a threshold of 1.2, the following cues and reformulations are likely to appear together:

New: H, U, W; Generalization: I and Z; Substitution: K; Specialization: F; Phrase Formation: I, J, K;

This type of analysis, however, does not take into account the variation between different participants and tasks. It also does not account for the fact that observations from the same participants and tasks may be correlated. To address this issue, I ran mixed models as first described in 4.8. Mixed models can account for differences between participants and tasks, given multiple instances for each task and participant.

5.3 Mixed effects models results

In order to fit a good model with interaction terms, I had to make sure there were enough data points at the intersection of the moderating variable and each cue, given a certain reformulation (for example, search expertise and cue F for the reformulation category "generalization"). Basing a model on insufficient data points, especially if they are all grouped in the same level of domain knowledge or search expertise, would cause an exaggerated extrapolation. Therefore, I generated descriptive statistics for each reformulation category along with the frequency of each cue at

various levels of search engine expertise and domain knowledge. Table 17 shows the cues which had very low frequencies at various levels of domain knowledge. As can be seen from the table, for reformulation category "new" (Table 17), cue Z [different meaning] appeared (was equal to 1) only twice for domain knowledge=1, cue I appeared once for domain knowledge=4, and cue J appeared twice for domain knowledge=1. Due to their low frequency, the interactions of these cues (Z, I, and J) with domain knowledge were never included in the model.

TABLE 17: FREQUENCY OF CUES AT VARIOUS LEVELS OF DOMAIN KNOWLEDGE FOR REFORMULATION CATEGORY = NEW

Domain Knowledge	cue Z=0	cue Z=1	cue I=0	cue I=1	cue J=0	cue J=1
1	51	2	53	0	51	2
1.5	0	0	0	0	0	0
2	32	0	32	0	32	0
2.5	0	0	0	0	0	0
3	13	0	13	0	13	0
3.5	0	0	0	0	0	0
4	19	0	18	1	19	0
5	0	0	0	0	0	0
Grand Total	115	2	116	1	115	2

The results below (Table 18 through Table 22, the full model results can be found in Appendix 13) present the main and interaction effects for each of the five reformulation categories. The statistically significant ($p<0.1$) main effects and interactions in these tables are marked in bold. Similarly to regular logistic regression, the coefficients (estimates) are in log-odds units. If a main effect does not participate in any significant interactions (such as cue Z [different meaning] in Table 18), this means that it has the same effect on the reformulation category (generalization category in Table 18), regardless of the level of search expertise and domain knowledge. In other words, if cue Z is observed, it means that it has the same odds to lead to generalization, no matter

what level of search expertise or domain knowledge the participant has. A positive value of the coefficient (for example, 0.987 for Z) indicates that the odds for a certain reformulation increase if a particular cue is observed and a negative value indicates that the odds for that reformulation decrease.

If the coefficient is converted to odds ratio by calculating $e^{coefficient}$, the odds ratio represent the change in the odds of performing a particular reformulation category when a certain cue is observed compared to the odds in the absence of that cue (Browner, 2006). Once converted to odds ratio ($e^{coefficient}$), values larger than 1 indicate that the odds for that particular reformulation would increase with the usage of this cue and values smaller than 1 indicate that the odds for that particular reformulation would decrease with the usage of this cue.

As described in section 4.8.3, the mixed effects model was run for each reformulation category, where search expertise and domain knowledge were centered around three different levels: mean, mean minus one standard deviation (low), and mean plus one standard deviation (high). If a cue participated in a significant interaction, its main effect was checked at these three levels to determine at which level it becomes significant. If the coefficient of the interaction is positive, this means that for each increase in the moderating variable (search expertise or domain knowledge), the effect of the main effect (cue) on the reformulation category increases. If the coefficient is negative, for each increase in the moderating variable, the effect of the main effect on the reformulation category decreases. For example, in Table 18 the coefficient of K*SE is negative (-0.28), which means that for each increase in SE, the effect of cue K on getting "generalization" decreases. Only at the low level of SE (at mean minus one standard deviation) the effect becomes positive (0.584) and significant. This means that for participants with low search expertise, the odds for "generalization" after observing cue K are $e^{0.584}$ times higher

compared to the odds without cue K usage. The data presented in the tables below was used to construct a model that will be presented and discussed in section 6.2.

In order to determine the effects of cues on search effectiveness (correctness of answers), a similar mixed effects model was run. The response variable was expressed by the correctness of the answers. The results appear in Table 23 and show that most of the cues have a negative effect on the ability to answer a question correctly. The only cue that increases the odds for a correct answer is cue J [similar results], only when used by participants with low search expertise.

TABLE 18: REFORMULATION CATEGORY = GENERALIZATION

	main effects held at mean minus one standard deviation of domain knowledge (DK) and search expertise (SE)			main effects held at mean SE and DK			main effects held at mean plus one standard deviation of SE and DK		
effects	coefficient	Odds ratio= $e^{coefficient}$	p-value	coefficient	Odds ratio= $e^{coefficient}$	p-value	coefficient	Odds ratio= $e^{coefficient}$	p-value
G	-0.21635	0.80545	p>=0.1	0.1598	1.17328	p>=0.1	**0.53594**	**1.70905**	**p<0.1**
I	0.13449	1.14395	p>=0.1	**1.3061**	**3.69175**	**p<0.01**	**2.47772**	**11.91407**	**p<0.001**
K	**0.58414**	**1.79345**	**p<0.05**	0.16111	1.17481	p>=0.1	-0.26195	0.76955	p>=0.1
Z	**0.98719**	**2.68368**	**p<0.01**	**0.98722**	**2.68376**	**p<0.01**	**0.98721**	**2.68374**	**p<0.01**
SE	0.05068	1.05199	p>=0.1	0.05069	1.05200	p>=0.1	0.05069	1.05200	p>=0.1
DK	-0.09708	0.90748	p>=0.1	-0.09707	0.90749	p>=0.1	-0.09708	0.90748	p>=0.1
I*SE	**0.77679**	**2.17448**	**p<0.1**	**0.77678**	**2.17446**	**p<0.1**	**0.77679**	**2.17448**	**p<0.1**
K*SE	**-0.28048**	**0.75542**	**p<0.1**	**-0.28049**	**0.75541**	**p<0.1**	**-0.28049**	**0.75541**	**p<0.1**
G*DK	**0.33579**	**1.39905**	**p<0.1**	**0.33569**	**1.39891**	**p<0.1**	**0.33573**	**1.39896**	**p<0.1**

TABLE 19: REFORMULATION CATEGORY = SPECIALIZATION

	main effects held at mean minus one standard deviation of search expertise (SE)			main effects held at mean of SE			main effects held at the mean plus one standard deviation of SE		
effects	coefficient	Odds ratio= $e^{coefficient}$	p-value	coefficient	Odds ratio= $e^{coefficient}$	p-value	coefficient	Odds ratio= $e^{coefficient}$	p-value
F	**0.76638**	**2.151962034**	**p<0.001**	**0.76636**	**2.151918995**	**p<0.001**	**0.76636**	**2.151918995**	**p<0.001**
I	-0.55974	0.571357598	p>=0.1	**-2.81249**	**0.060055268**	**p<0.5**	**-5.06517**	**0.006312838**	**p<0.01**
K	**-0.74243**	**0.475955936**	**p<0.05**	0.05464	1.056160329	p>=0.1	**0.85171**	**2.343651071**	**p<0.01**
SE	-0.02182	0.978416334	p>=0.1	-0.02182	0.978416334	p>=0.1	-0.02182	0.978416334	p>=0.1
I:SE	**-1.49354**	**0.224576247**	**p<0.05**	**-1.49358**	**0.224567264**	**p<0.05**	**-1.49356**	**0.224571755**	**p<0.05**
K:SE	**0.52842**	**1.696250115**	**p<0.001**	**0.52838**	**1.696182266**	**p<0.001**	**0.52842**	**1.696250115**	**p<0.001**

TABLE 20: REFORMULATION CATEGORY = SUBSTITUTION

	main effects held at mean minus one standard deviation of search expertise (SE)			main effects held at mean of SE			main effects held at mean plus one standard deviation of SE		
effects	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value
J	-1.59776	0.202349273	p>=0.1	-0.33324	0.71659819	p>=0.1	0.9313	2.537806182	p>=0.1
SE	-0.06002	0.941745698	p>=0.1	-0.06002	0.941745698	p>=0.1	-0.06002	0.941745698	p>=0.1
J:SE	**0.83832**	**2.312478747**	**p<0.1**	**0.83824**	**2.312293756**	**p<0.1**	**0.83825**	**2.312316879**	**p<0.1**

TABLE 21: REFORMULATION CATEGORY = NEW

	main effects held at mean minus one	main effects held at mean of search	main effects held at mean plus one

	standard deviation of SE			expertise (SE)			standard deviation of SE		
effects	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value
F	-0.8772	0.415945931	p>=0.1	-0.023	0.977262484	p>=0.1	0.8312	2.296072375	p>=0.1
H	**2.0403**	**7.692916728**	**p<0.001**	**2.0403**	**7.692916728**	**p<0.001**	**2.0403**	**7.692916728**	**p<0.001**
U	**2.6185**	**13.71513544**	**p<0.001**	**2.6185**	**13.71513544**	**p<0.001**	**2.6185**	**13.71513544**	**p<0.001**
W	**1.5087**	**4.520849868**	**p<0.001**	**1.5087**	**4.520849868**	**p<0.001**	**1.5087**	**4.520849868**	**p<0.001**
SE	-0.1425	0.867187554	p>=0.1	-0.1425	0.867187554	p>=0.1	-0.1425	0.867187554	p>=0.1
F:SE	**0.5663**	**1.761736552**	**p<0.05**	**0.5663**	**1.761736552**	**p<0.05**	**0.5663**	**1.761736552**	**p<0.05**

TABLE 22: REFORMULATION CATEGORY = PHRASE FORMATION

	main effects held at mean minus one standard deviation of domain knowledge (DK)			main effects held at mean of DK			main effects held at mean plus one standard deviation of DK		
effects	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value
G	**-1.6313**	**0.195675031**	**p<0.1**	**-1.6313**	**0.195675031**	**p<0.1**	**-1.6313**	**0.195675031**	**p<0.1**
F	**-2.6323**	**0.071912872**	**p<0.1**	**-2.6323**	**0.071912872**	**p<0.1**	**-2.6323**	**0.071912872**	**p<0.1**
H	-1.0692	0.343283034	p>=0.1	-1.0692	0.343283034	p>=0.1	-1.0692	0.343283034	p>=0.1
I	**5.402**	**221.8496721**	**p<0.01**	**3.2661**	**26.20892496**	**p<0.01**	**1.1303**	**3.096585336**	**p>=0.1**
J	-0.3721	0.689285311	p>=0.1	-0.3721	0.689285311	p>=0.1	-0.3721	0.689285311	p>=0.1
K	**2.3453**	**10.43640318**	**p<0.05**	**-0.2772**	**0.757902901**	**p>=0.1**	**-2.8997**	**0.055039729**	**p<0.1**
DK	0.296	1.344470157	p>=0.1	0.296	1.344470157	p>=0.1	0.296	1.344470157	p>=0.1
I:DK	**-1.9067**	**0.148569859**	**p<0.1**	**-1.9067**	**0.148569859**	**p<0.1**	**-1.9067**	**0.148569859**	**p<0.1**
K:DK	**-2.3413**	**0.096202494**	**p<0.05**	**-2.3413**	**0.096202494**	**p<0.05**	**-2.3413**	**0.096202494**	**p<0.05**

TABLE 23: RESPONSE VARIABLE = SEARCH EFFECTIVENESS (ANSWER CORRECTNESS)

	main effects held at mean minus one standard deviation of SE and DK			main effects held at mean of search expertise (SE) and domain knowledge (DK)			main effects held at mean plus one standard deviation of SE and DK		
effects	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value	coefficient	$e^{coefficient}$	p-value
G	**-1.04346**	**0.352233842**	**p<0.1**	-0.22544	0.798164949	p>=0.1	0.59257	1.808630628	p>=0.1
F	0.97097	2.640504507	p>=0.1	-0.2393	0.787178693	p>=0.1	**-1.44956**	**0.234673522**	**p<0.1**
H	-0.96264	0.381883382	p>=0.1	-0.07093	0.931527097	p>=0.1	0.82078	2.272271518	p>=0.1
J	**2.05647**	**7.818322363**	**p<0.05**	0.10621	1.112055384	p>=0.1	**-1.84403**	**0.15817868**	**p<0.05**
K	**-1.63814**	**0.194341181**	**p<0.001**	**-1.63814**	**0.194341181**	**p<0.001**	**-1.63813**	**0.194343124**	**p<0.001**
U	**-1.3154**	**0.268366955**	**p<0.1**	**-1.31539**	**0.268369639**	**p<0.1**	**-1.31547**	**0.26834817**	**p<0.1**
G:SE	**0.54248**	**1.720267841**	**p<0.1**	**0.54248**	**1.720267841**	**p<0.1**	**0.54247**	**1.720250638**	**p<0.1**
J:SE	**-1.29336**	**0.274347425**	**p<0.01**	**-1.29337**	**0.274344682**	**p<0.01**	**-1.29335**	**0.274350169**	**p<0.01**
F:DK	**-1.07666**	**0.340731671**	**p<0.05**	**-1.07666**	**0.340731671**	**p<0.05**	**-1.07665**	**0.340735078**	**p<0.05**
H:DK	**0.79323**	**2.210524903**	**p<0.1**	**0.79322**	**2.210502798**	**p<0.1**	**0.79326**	**2.21059122**	**p<0.1**

TABLE 24: THE EFFECTS OF SEARCH EXPERTISE AND DOMAIN KNOWLEDGE (REGARDLESS OF CUES) ON SEARCH EFFECTIVENESS

effects	coefficient	$e^{coefficient}$	p-value
Search Expertise	0.2375	1.268075	**p<0.05**
Domain Knowledge	0.6653	1.945074	**p<0.001**

6 Discussion

This study set out to explore a broad question about the systematic usage of cues (on SERPs or on the target pages) in query reformulation. More specifically, I wanted to explore whether there are consistent relationships between the uses of cues and reformulation categories (RQ1). In addition, recognizing that search is a highly contextualized activity, I wanted to examine the role of domain knowledge and search expertise as mediating factors in the searchers' use of cues in query reformulation (RQ2). To answer these questions, however, I had first to establish whether there are established patterns of reliance on cues in search behavior – do searchers systematically pay attention to the same aspects of SERPs or the target pages? Thus, I will start the discussion with establishing this fundamental element of the study, which I refer to as cues in search. I then will move on to discussing each one of the research questions in relation to the findings presented in the previous chapter; I refer to that part as cues in reformulations.

6.1 Cue discovery

The results presented in chapter 5 suggest that searchers are reliably able to point to particular aspects of the SERPs or the full pages when reformulating their queries. Moreover, it appears that they systematically rely on similar cues in similar search situations and it was possible to identify patterns of similar behavior across participants. Those cues can be roughly arranged in three groups based on the frequency of their occurrence.

6.1.1 High frequency cues

One of the behaviors in the common cue group happened when a participant tried answering a question that in their mind had a specific type of answer, in which case the participants were

scanning the results (SERP or full pages) for certain words, which were not part of the query (cue F). This cues was more about not what they saw, but what they didn't see and had trouble locating on the SERPs and pages. That was not necessarily because they forgot to include those terms in the query, but because they expected those terms to be part of the answer. These words usually belonged to a certain category of words, such as: names, dates, numbers, etc., which typically were the target of the search task in hand. For example, when evaluating results in their search for an answer to question about Australian sports, participants did not focus specifically on the word "Australian" or "sports", but were scanning the pages for names of sports, such as “rugby.” At times, they would find such a list of names and it would trigger an idea for a new query or a change in the existing query (cue H). This behavior included participants stumbling upon a word which was not part of their previous query, but appeared in the results and provided an idea for a change in the query. For example, after looking at a list of sport names, they noticed "rugby" and decided to check whether or not this sport originated in Australia.

Another consistent pattern of behavior could be observed when participants were assessing the relevance of the search results. For example, when they saw query terms appearing together with unwanted word/s which were not part of the query (cue K), a single word or a phrase were often enough to signal participants that the result as a whole was irrelevant. So, when looking for minerals that float on salt water, any pages that talked about "grapes" or "people" deterred the participants. Interestingly, out of all the participants who expressed their intention to get away from certain words in results, only one participant used the "-" operator to resolve that. In other cases, when the participants tried to stay away from unwanted categories of pages (e.g. forums, Q&A websites, travel websites) they often paid attention to words that were not part of the query but appeared in the results and indicated the character/type of the page (cue G).

6.1.2 Medium frequency cues

While cues F, G, H, and K were the most common elements of the search repertoire across participants, a number of other cues were frequent enough in my sample to make it into the final model. In many cases, the participants noticed that their query terms took on different meaning when presented in the results (cue Z). For example, "time" meaning "duration" vs. "time" meaning "when something occurred". Participants usually strived to get away from those words with a different meaning. Another common occurrence was when participants noticing that some of the retrieved web pages were repeating, similar to the results they had received before (cue J). For example, getting the same wikipedia page about the jominy test across different queries.

In numerous instances, participants noticed that their query terms were spread out in the results, contrary to their anticipation for the terms to appear next to each other (cue I). Many participants used the quotes operator to bring the query terms together, but for others this was not a trivial solution. Finally, in some cases, the participants relied on answering an intermediate question, different than the assigned task, and used that as a stepping stone in reformulation (cue U). One example of such behavior occurred when, for whatever reason (either came up with it on their own or got the idea in previous results, a participant thought that rugby had originated in Australia, but the results of that query showed that was not true. The participant then relied on that information as she moved on to the next query. At times, participants noted that they didn't base their decision to reformulate on anything particular that appeared on any of the pages, but rather followed a hunch, their prior knowledge or simply didn't recall if they had noticed anything particular in the results that lead them to reformulate. This situation was labeled as W ("no cues").

6.1.3 Low frequency cues

The last group of the rarer behaviors still contains valuable insights into the elements that searchers pay attention to. One of those behaviors usually came up when the participants noted that a certain query term or phrase was too dominant in the results (cue L). For example, cue L would occur when a participant decided that the phrase "salt-water" appeared too frequently in the search results about floating minerals. For them, too many results revolved specifically around salt-water, but not about the floating properties. This is a particularly interesting cue, as it does not occur on a single snippet level, but rather on an overall impression the participant had gotten from an entire page of results. Another observation made by participants, and which occurred at the level of an entire SERP, suggested that "the results were too all over the place" (cue M). In other words, the participants noticed that the results were too diverse in their nature, e.g. a few scientific web-pages, a few travel pages, and a few pages containing experiments for kids. This type of "entire SERP" cues only occurred in less than 2% of the cases, but does provide an insight on the participants' ability to evaluate the whole SERP vs. each snippet individually. Finally, in a number of cases, the participants have indicated that the relationship between two search terms in the results was opposite from the desired relationship (cue N). For example, they noticed when the results contained web pages about Australia adopting sports from Britain, as opposed to the desired relationship of Britain adopting from Australia. Although not many participants paid attention to this cue, it is important to note, because Google's search engine currently does not have any indication for such mix up. The above mentioned uncommon cues were too infrequent to enter the model, but are still worth noting as they offer an insight into searchers' behavior and can be potentially useful for future research.

6.2 Effect of cues on reformulations

This section discusses the results that were used to answer RQ1. After the process of cue discovery was completed, I used the cues and the reformulation categories to create a model which predicts the way cues drive the reformulations of subsequent queries. In other words, determine which cues lead to which reformulation categories. This model also shows how search expertise and domain knowledge moderate the relationships between the cues and query reformulation categories. The model is based on the systematic behavior exhibited by the participants, taking into account only the main effects and interactions regarding which there was enough evidence to show that these relationships exist and contribute to the model. In other words, including only the effects which were statistically significant. The cues which decrease the chance of a certain reformulation category (odd ratios less than 1) are not represented in this model. These relationships were left out because information on which cues decrease the chance for a certain reformulation is not very useful. It merely gives information on which reformulation is less likely to occur with usage of that cue, while the whole purpose of this study is to discover which reformulations are likely to occur after certain cue usage. Knowing that a certain reformulation is less likely to occur after a certain cue was observed, just means that another reformulation category is more likely to occur given that cue.

Figure 5 depicts the model, showing which cues lead to which reformulation categories (gen=generalization, phrase=phrase formation, spec=specialization), along with the moderating effects of search expertise and domain knowledge. The numbers on each connector indicate the odds ratio of a particular reformulation after certain cue usage was observed, compared to the absence of that cue. For example, observing cue Z [different meaning] is associated with 2.68-fold greater odds of reformulation using "generalization", compared to the odds of

"generalization" without observing cue Z [different meaning]. The boxes which contain either SE or DK, represent the interaction effects of search expertise (SE) and domain knowledge (DK). If the arrow near SE or DK is upward pointing, this indicates that the higher the moderating variable is, the larger is the effect of the cue on the reformulation. If it's pointing downward, this means that the larger the moderating variable is, the lower is the effect of the cue on the reformulation. For example, for participants with higher search expertise, the usage of I [terms spread out] is associated with 11.9-fold greater odds of getting generalization, compared to the odds of getting generalization without cue I [terms spread out]. It is important to note that for the "substitution" category, there was only one interaction which were also borderline significant (p-value = 0.0959) and its main effect (cue J [similar results]) was not significant in the meaningful range of search expertise. Therefore it was not included in the model below.

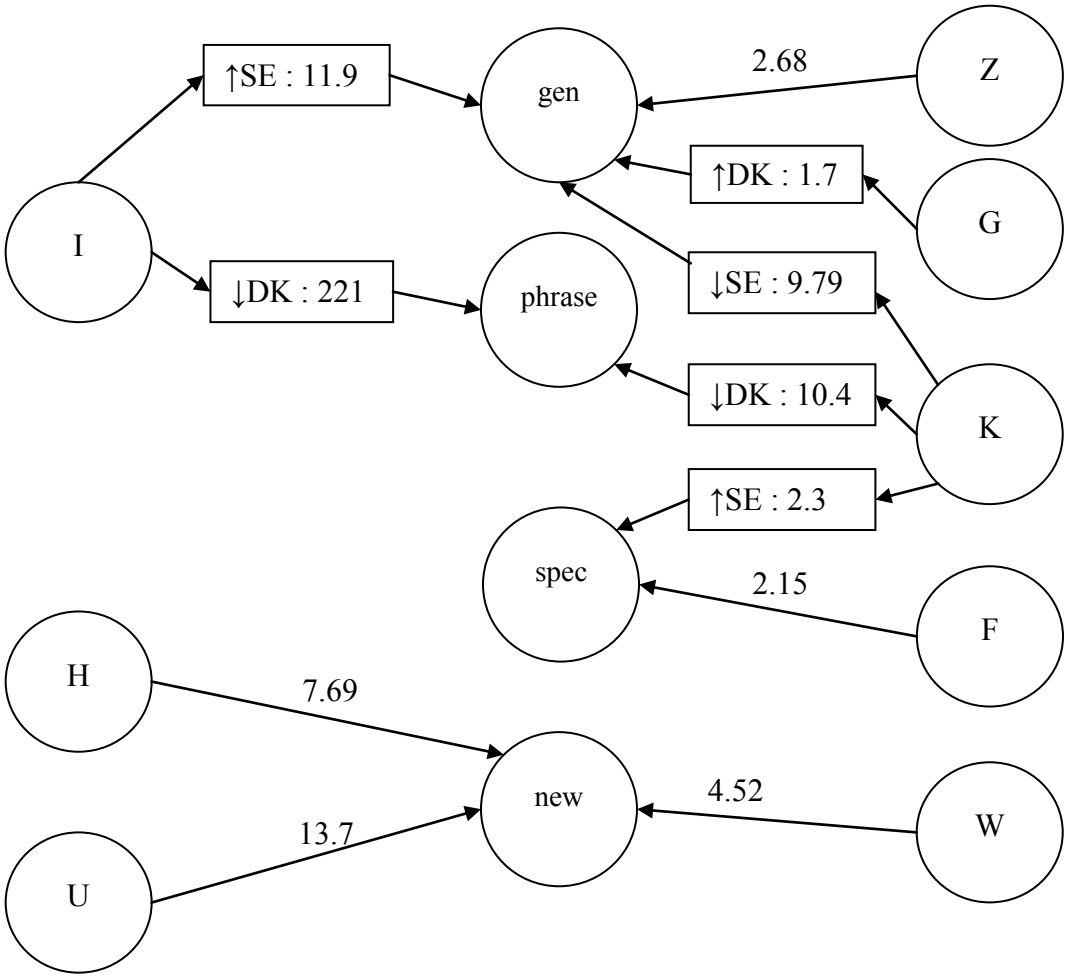

FIGURE 5: CUES AND REFORMULATION CATEGORIES RELATIONSHIPS WITH MODERATING FACTORS

Figure 5 depicts the following relationships between cues and reformulation categories:

Cue I [terms spread out]:

For participants with lower domain knowledge, the usage of cue I [terms spread out] is associated with 221-fold greater odds of performing phrase formation, compared to the odds of performing phrase formation without cue I [terms spread out]. At the same time, for participants with higher search expertise, the usage of cue I [terms spread out] is associated with 11.9-fold greater odds of performing generalization, compared to the odds of getting performing without cue I [terms spread out].

Cue K [unwanted words]:

For participants with lower search expertise, the usage of cue K [unwanted words] is associated with 9.8-fold greater odds of performing generalization, compared to the odds of performing generalization without cue K [unwanted words]. At the same time, for participants with higher search expertise, the usage of cue K [unwanted words] is associated with 2.3-fold greater odds of performing specialization, compared to the odds of performing specialization without cue K [unwanted words]. For participants with lower domain knowledge, the usage of cue K [unwanted words] is associated with 10.4-fold greater odds of performing phrase formation, compared to the odds of performing phrase formation without cue K [unwanted words].

Cue G [unwanted page type]:

For participants with higher domain knowledge, the usage of cue G [unwanted page type] is associated with 1.7-fold greater odds of performing generalization, compared to the odds of performing generalization without cue G [unwanted page type].

Cue Z [different meaning]:

Regardless of domain knowledge or search expertise levels, the usage of cue Z [different meaning] is associated with 2.68-fold greater odds of performing generalization, compared to the odds of performing generalization without cue Z [different meaning].

Cue F [expecting a word]:

Regardless of domain knowledge or search expertise levels, the usage of cue F [expecting a word] is associated with 2.15-fold greater odds of performing specialization, compared to the odds of performing specialization without cue F [expecting a word].

Cue U [intermediate answer]:

Regardless of domain knowledge or search expertise levels, the usage of cue U [intermediate answer] is associated with 13.7-fold greater odds of starting a new query, compared to the odds of starting a new query without cue U [intermediate answer].

Cue H [provide an idea]:

Regardless of domain knowledge or search expertise levels, the usage of cue H [provide an idea] is associated with 7.69-fold greater odds of starting a new query, compared to the odds of starting a new query without cue H [provide an idea].

Cue W [no cues]:

Regardless of domain knowledge or search expertise levels, when the searcher doesn't notice any cues, this situation is associated with 4.52-fold greater odds of starting a new query, compared to the odds of starting a new query when cues are used.

In the remainder of this section, I will interpret the relationships depicted in the model and explain the possible reasoning behind these relationships with examples and quotes from the collected data.

6.2.1 Cues that lead to phrase formation

One of the most obvious outcomes is the finding that cue I [terms spread out] is likely to lead to phrase formation, especially for participants with lower domain knowledge. It is reasonable to believe that when a participant sees that her query terms are not next to each other, her next step would be to force those terms together by phrase formation. One explanation for the increased effect of cue I [terms spread out] for participants with low domain expertise is overuse of phrase formation due to lack of familiarity with relevant terminology. It is possible that those

participants tried to stick to the same familiar phrases which appeared in the description of the question, because they had difficulty using their own terms in the reformulation, since they were not familiar with any additional domain specific terms. For example, one participant changed the query from 'what minerals can float in salt water,' which was also the exact description of the question, to '"salt water" "minerals" "float".' In this instance, the query terms remained practically identical to the original query, but the participant tried to force the same terms as they appeared in the description of the question. The participant stated: "Like here [points at a result of 'Great Salt Lake'] it's just 'salt', not 'salt water' and then water, I wanted salt water together."

The reasoning behind cue K [unwanted words] leading to phrase formation for participants with low domain knowledge could also be explicated. Some participants were trying to get away from pages that contained unwanted words, by insisting that certain query terms appear together. They were possibly hoping to force a certain context on some or all of the query terms and by this keep away the unwanted words from appearing next to the query terms. For example, an initial query of 'britain adopting australian sports' was changed to 'britain "adopting australian sports"', because this participant wanted to get away from the word "constitution" in the results, which appeared as 'adopting a constitution'. The participant explained: " they just weren't talking about… like it was like 'adopting a constitution', like that's not a sport and things like that."The reason behind this type of relationship in participants with low domain knowledge could also be related to their lack of familiarity with relevant terminology. As a result, they were lacking ideas on which terms to add and tended to reuse of the same terms from previous queries, with an addition of phrase formation around some of the terms.

The finding that cues I [terms spread out] and K [unwanted words] are likely to lead to phrase formation is consistent with the results of the conditional probability analysis, however, that

analysis also suggested that cue J [similar results] is also likely to lead to phrase formation. Even though the mixed effects model did not support this relationship, cue J [similar results] leading to phrase formation could be explained by cases in which participants tried to filter out the results they had seen before, by forcing some of the query terms together. This was their attempt to exclude some of the words which they believed were causing the recurring results. For example, changing the query '"Britain adopts" Australian sport' to '"Britain adopts" "Australian sport"', by forcing 'Australian' and 'sport' together. Since the results that kept reappearing were consistently mentioning 'australian sporting bodies', the participant was trying to exclude the 'sporting bodies' by placing double quotes around 'australian' and 'sport'. As the participant stated: " [reading] 'Australian sporting bodies' - it is not talking about australian sports, it is talking about australian body or person that plays the sport."

6.2.2 Cues that lead to generalization

Another mixed effects model finding is the outcome that cues Z [different meaning], as well as I [terms spread out] (moderated by search expertise), K [unwanted words] (moderated by search expertise), and G [unwanted page type] (moderated by domain knowledge), are likely to lead to "generalization". When one of the query terms had a different meaning than intended (Z), some participants preferred to remove that term all together (generalization). For example, in initial query of 'which australian sport is now found in Britain' the phrase 'now found' was removed and the query changed to 'what australian sport was adopted in britain', because the meaning of the word 'found' appeared in a sense of 'founded' instead of 'currently exists'. As described by the participant: "this is the word 'founded' is in the sense of 'which is founded' and then I knew that this is isn't the one I wanted to look at."

Cue G [unwanted page type] leading to generalization is represented by situations in which participants who noticed pages that belong to an unwanted category (G) wanted to get away from this category. They often tried to do that by removing the query terms they thought may have brought up those pages. Given that this relationship is moderated by domain knowledge, it is difficult to say what led participants with higher domain knowledge to generalize their queries after seeing pages that belong to an unwanted category. Since at times participants wanted to get away from news websites, it could be that participants with higher domain knowledge preferred more official sources. For example, one participant started with a query " what australian sport is now played in britain" and then moved on to " history of sports australia britain". Due to the realization that the first query yielded news pages that dealt mostly with current issues and wanted to get away from that. The participant explained: "I wanted a website that had a history of a sport, because that way, it would be 'this sport originated in australia, but england participates in this sport now'. Stuff like that, rather than current sport updates, that's not really important" me: is this something that you saw on previous pages 'current sport updates'? Participant: yeah, like 'golf news', 'live cricket scores' - I just wanted to get rid of current issues."

When query terms do not appear next to each other (spread out, cue I) and lead to "generalization", this can be explained by the possible perception of participants that if the terms don't appear next to each other in the results, they should be removed from the query and only the terms they wanted to see together should remain in the query. For example, starting with 'what happens when minerals are less dense than salt water' led to results with query terms spread out ('minerals' in one part of the snippet, 'salt' and 'water' in another). This may have made the participant think that by leaving only 'minerals in salt water' in the query, the terms 'minerals' and 'salt water' would appear next to each other. The participant indicated that: "...it's all my

keywords, and it does what it's supposed to do, it just doesn't apply to what I'm looking for. Like water's up here, minerals and salt down there." It is difficult to explain why cue I [terms spread out] has higher odds leading to "generalization" for participants with higher search expertise.

According to the mixed effects model, cue K [unwanted words] is also likely to lead to "generalization" for participants with lower search expertise. Cue K represents situations where participants saw a certain word that was unwanted and they wanted to get away from it. Similarly to the relationship between cue G [unwanted page type] and "generalization", they were trying to get away from the unwanted words, by removing the query terms they thought had led to the unwanted words. For example, a participant changed their query from 'minerals floating in salt water' to 'minerals in salt water', because they believed that the term 'float' was the one that led to pages with phrases such as 'eggs floating' and she wanted to get away from eggs. When asked why she removed the word 'float', the participant responded: "it went more to the eggs instead of the actual minerals..." It is possible that participants with lower search expertise were not familiar with other ways of filtering out unwanted words (by the "-" operator for example) and used "generalization" instead.

These findings were partially consistent with the conditional probability analysis of the generalization category, which showed that cue Z [different meaning] and cue I are likely to lead to generalization. Cues G [unwanted page type] and K [unwanted words] were not supported by this analysis, however these cues did get a ratio higher than 1 (1.089 and 1.113 respectively) for the "generalization" category, so a smaller threshold would include those cues as well.

6.2.3 Cues that lead to specialization

Based on the mixed model analysis, cues F [expecting a word] and K [unwanted words] are likely to lead to specialization. The relationship between K [unwanted words] and specialization is moderated by search expertise, meaning that participants with higher search expertise are more likely to add more terms when they see unwanted word/s in their results. These instances occurred when participants started their search too broadly and when they noticed words that indicated the broad nature of the results, they wished to narrow these general results by adding more specific query terms. For example, following the query "Australian influence" the participant saw the words 'languages', 'climate', 'China' and realized she needed to also add 'sport' to narrow down the results to be sports specific and further away from 'languages' and other unwanted words. As the participant reported: "because looking at this 'Australian climate influences' and then I saw 'language' and then 'China', and then I looked at this one here and I saw that these results were too broad and I needed to go back." It is difficult to hypothesize why this type of relationship has a higher chance of occurring for participants with high search expertise, since as opposed to participants with lower search expertise, one would expect of participants with higher search expertise to use more advanced operators such as "-" instead of simply adding words to the query. This type of behavior should be more characteristic of low expertise participants.

Cue F represents situations in which the participant is scanning for certain words that were not part of their query. Specialization is very plausible with this cue, because this means that the participant may simply add the term they were scanning for and was missing. Alternatively, in case they were scanning for something that could not be added directly, such as names of minerals, they could be attempting to indirectly get those words to show up in the results, by

adding additional terms. For instance, participant who started with "minerals float water" was scanning the results for a list of mineral names and tried to get to this list of names by changing the query to "minerals float water list". These results are partially consistent with the conditional probability analysis, which also suggested cue F as the cue which is likely to lead to specialization. Cues K [unwanted words] and G [unwanted page type] were not suggested by the conditional probability analysis, although cue G was very close to the threshold (1.182).

6.2.4 Cues that lead to a new query

The cues that are likely to lead to the "new" category are H [provide an idea], U [intermediate answer], and W [no cues]. This is consistent with the results of the conditional probability analysis, which resulted in the same cues. Cue H represents situations in which a participant notices a certain word or a phrase, which give her the idea to reformulate. In cases where the idea is completely different than the previous query, it is acceptable that the following query would be a completely new one. For example, going from the query 'australian sport adopted in britain' to 'where did water polo originate' because the participant had seen 'water polo' on a general list of sports in the results. The participant explained how he got the idea: "in one of the things I had found something about water polo so I thought that maybe that if I just searched a few of the sports that I found that had been in Australia."

The explanation of the cue U [intermediate answer] leading to the "new" reformulation category has a similar explanation. Since cue U represents instances in which the participant found an intermediate answer to an intermediate question, it is possible that once the intermediate question is answered, the following query would be completely different. For example, once the participant receives an answer to a query "is Rugby Australian?", the next query would have nothing to do with rugby anymore, because the participant would have answered that question

and be trying a new direction. Finally, there is cue W, which isn't really a cue, but rather signifies the absence of cues, means that the participant did not identify any cues. This instance is also a probable candidate to lead to a new query, because when there are no more cues to use, it is reasonable to assume that the participant would seek a new direction to follow.

6.3 Cues and search effectiveness

This section discusses the results that were used to answer RQ2. Search effectiveness was measured at the task level and marked as '1' or '0', depending on the correct answer or incorrect answer (respectively) that the participant provided for that question. The goal for this part of the research was to determine what effect cues had on the correctness of the answers and whether there was a mediating effect of search expertise and domain knowledge on this relationship. Results showed that participants with low search engine expertise who use cue J [similar results] in any of the reformulations are more likely to answer the question correctly than without cue J [similar results]. At the same time, for participants with high search expertise, J [similar results] has a negative effect (the odds of getting a correct answer with these cues are lower than the odds of getting a correct answer without these cues) on search effectiveness. Domain knowledge on its own was found to have a positive effect on the ability to get a correct answer. All the other effects are either insignificant at the three levels (mean, low, and high) of the moderating variable or have a negative effect on search effectiveness.

For example, results showed cues K [unwanted words] and U [intermediate answer] have a negative effect regardless of search expertise or domain knowledge. Cue G [unwanted page type] has a negative effect on search effectiveness for participants with lower search expertise. Observing cue F [expecting a word] is likely to lead to incorrect answers as well, but only for

participants with high domain knowledge. Cue H [provide an idea] has an interaction with domain knowledge, but does not have a significant effect on search effectiveness in the meaningful range (within the scale of these moderating variables). These results indicate that for participants with higher search expertise and domain knowledge there are no cues that have a positive effect on search effectiveness.

The finding that most of the significant relationships between cues and search effectiveness (the ability to answer a question correctly) are negative, could be related to the fact that all the cues represent negative situations such as unwanted word, unwanted category, missing query term, etc. that the searcher wishes to move away from by reformulating. Therefore, it is reasonable to believe that encountering these negative situations negatively affects search success. In addition, each reformulation indicates a failure of the sequence of queries up to that point. This means that there are necessarily a lot more cue-reformulation observations for tasks that had an incorrect or missing answer. In the collected data, based on reformulation frequency only, there were on average 6.09 reformulations for search sequences that resulted in incorrect answers and 4.59 reformulations on average for search sequences that resulted in correct answer. The difference between these averages is in fact statistically significant at the level of $p<0.05$. Consequently, since the number of cues in a task correlates with the number of reformulations, it is not surprising that most of the cues have a negative effect on search effectiveness.

It is important to note that domain knowledge on its own, regardless of cues has a statistically significant effect on search effectiveness. Results (Table 24) show that with each unit of increase in domain knowledge, the odds for a correct answer almost double (odds ratio = 1.95). This effect is statistically significant with $p< 0.001$. The results are similar for search expertise. With

each unit of increase in search expertise, the odds for getting a correct answer also increase (odds ratio = 1.27). This effect is statistically significant with $p < 0.05$.

This section discussed the results of analysis which aimed at describing the cues discovered in this study and answering the two research questions posed in section 1.3. RQ1 dealt with identifying the relationships between cues and reformulation categories. The second part of the question inquires about the moderation effect of search expertise and domain knowledge on these relationships. The relationships presented in this discussion answer both parts of this question by showing which cues lead to which reformulations and how the level of search expertise and domain knowledge affect the odds of getting each reformulation after observing a certain cue. RQ2 inquired about the contribution of cues to the level of search effectiveness. This question was answered by showing which cues had a positive effect on search effectiveness and which had a negative effect, moderated by search expertise and domain knowledge.

7 Conclusions

The overall goal of this study was to start unpacking the decision-making process behind query reformulation. It was driven by two broad questions about the use of cues (on the SERP or the full web pages) in the searchers' decisions regarding query reformulation (RQ1) and the effects of cue usage on search effectiveness (RQ2). In both cases I was interested in the role that search expertise and domain knowledge play in those relationships. The study utilized observations of stimulated recall, inductive analysis of recalled cues, and construction of mixed effects models that link cues to query reformulation and search effectiveness. As a whole, this work offers several contributions to the field of user-centered information retrieval.

First, it reaffirms some of the earlier conceptual work about the general role of cues in search behavior, then expands on it by proposing specific relationships between cues and reformulations, and finally, it serves a certain trailblazing capacity for future research in this field. Additionally, it highlights potential design considerations in creating SERPs and query term suggestions. Next, it offers training suggestions for educators working on digital literacy.

The conclusions of this study are best viewed as addressing four distinct, but interrelated, aspects of the use of cues in query reformulation. First, this study illustrates and explains some of the common dynamics of reliance on cues in generic search behavior. Second, it zooms onto a much more nuanced activity of query reformulation and the utility of cues in this behavior. Third, further unpacking the reformulation behavior, this study raises questions regarding the function of traditionally significant factors such as domain knowledge and search expertise. Finally, it addresses search effectiveness, which is the most commonly used way of evaluating and qualifying search behavior.

In this chapter, when discussing each aspect of the use of cues in query reformulation, I follow the following structure. I start with recapping the core of relevant findings. Then, I position these findings within existing understanding of the phenomenon, particularly highlighting unexpected and potentially provocative results. Lastly, I offer an explanation for these results and discuss how future research may achieve clearer understanding of that particular aspect.

7.1 Cues in search

In this study, I view search as an iterative process. Fundamental to this process is the act of query reformulation, for it is through revising their query that searchers move towards achieving or refining their initial information needs. Studying query reformulation, however, is difficult. The

act of query reformulation is largely tacit, it occurs as part of the searchers' mundane activities, and without much reflection or conscious consideration of what goes into it. Thus, my challenge here was to make the implicit mechanisms of query reformulation explicit, to uncover the pieces of information – the cues - that searchers pick from the SERPs or the target pages in order to evaluate the adequacy of their results and the relation of those cues to search behavior patterns.

In his description of the THOMAS system, which was based on his view of a dialogue structure, Oddy (1977) talks about forming an image of the searcher's interest. This image is continuously being modified like a query, based on the searcher's rejection or acceptance of different elements in the results. Each result carries new pieces of information, which help shape the image or in the case of this study, the subsequent query. These "pieces of information," are pivotal to the dynamic and constantly evolving process of search, as the individual notices elements to be accepted or rejected, which refine or shift her thinking and subsequent queries. In this sense, such elements are analogues to cues discussed in this study.

Reliance on cues, inductively elicited in this study in a reliable fashion, confirms that such dialogue dynamics indeed takes place in search behavior. Study participants consistently identified new pieces of information on SERPs and full web-pages, and then used this new information to reformulate their preceding queries. These pieces of information were new to the participants, not necessarily part of their initial thinking about the search task, and discovered through evaluation of results yielded by preceding queries. Sometimes those pieces of information indicated that the participants wanted to get away from them and at other times provided them new ideas for the subsequent query. This research, however, takes Oddy's (1977) limited list of elements that shape the image and expands it to a more elaborate list of cues, along with mapping out specific relationships between cues and search behaviors.

Some of the cues identified in this study are consistent with other, similar elements, mentioned in the literature. Cues K [unwanted words], G [unwanted page type], and Z [different meaning], for example, can be linked to the notion of relevance. The participants identified certain characteristics (unwanted words, unwanted webpage category, unwanted meaning) which signaled for them that the result containing these cues was not relevant to their search task. Comparing these cues to the classes of relevance clues compiled by Saracevic (2007b), all three would fall into the "topic" subcategory (under the "content" category). These cues are closest to the "content" clue, because they represent relevance (or in this case, irrelevance) based on the content and are dissimilar to the other subcategories (such as: quality, depth, scope, currency, treatment, and clarity). Similarly, cue H [provide an idea], which occurred when a participant got an idea for a reformulation based on the content of the search results can be mapped on to one of the users' feedback categories identified by Spink and Saracevic (1998). The 'term relevance feedback' category in their typology represented cases in which the subsequent query included new search terms from the retrieved output. This resonates with the idea that when searchers examine the search results, they are constantly learning new things and apply them to their subsequent queries, in order to shift the results into a more relevant direction.

Observing consistency with previous research supports the validity of my observations about what cues participants have relied on. Previous research, however, does not offer an insight into how exactly the cues are being used in the process of query reformulation. To address that gap, I constructed a multilevel model that accounts for both the use of cues and the moderating role of search expertise and domain knowledge. The final model included nine cues, eight of which ended up being significant either as main effects or moderated by search expertise or domain knowledge. The results of the model support the fundamental intuition underlying this research -

cues matter in query reformulation. And not only that. The ability to construct a model with statistically significant effects shows that when refining queries, reformulations are not done arbitrarily. Instead, searchers systematically elicit and use cues with regard to query reformulation. The systematic nature is expressed in a pattern, in terms of which cues lead to which reformulations. Not all the relationships support the initial intuition, but speculations can be raised to explain either the obvious or not so obvious relationships.

7.2 Cues and reformulations

Once mapped out, some of the relationships between cues and query reformulation may appear straight forward and easy to explain. For example, when participants saw that some of their query terms were spread out (cue I), they revised their query so that those terms necessarily appear together (phrase formation). Another obvious relationship is the likelihood of cue H [provide an idea] to lead to a new query. Since in many cases the new idea which rose from the results may be completely different than the previous query, it is reasonable that the following query would be a completely new one. Findings that reaffirm common sense intuitions about the relationship between cues and reformulations are an important piece, making existing conceptual models (e.g. Oddy (1977)) more robust and valid. They help forming sound foundations for future research in human-centric IR.

Other relationships between cues and query reformulation are not as straight forward and may call for a more complex explanation. For example, the model suggests that cue K [unwanted words] is likely to lead to "generalization" for participants with lower search expertise. One explanation for this occurrence is that participants were trying to get away from the unwanted words by removing the terms they thought had caused that result and without realizing that a

more general query can lead to results with more unwanted words. More unexpected relationships like this will be described and explained in the subsequent sections.

The less intuitive findings offer fruitful ground for future research. In the example above, examining the content of the queries (as opposed to relying on the more technical definition used in this study) can offer alternative explanations to the logic underlying such behavior. For example, if a user removes common words from the query, in her mind she might be performing specialization. Additionally, in-depth interviews interrogating how experienced and inexperienced searchers interpret search results can shed light on how these two populations go about their decision-making process. Another line of research can explore the learning patters of search behavior by examining sequences of reformulations. In the case when the searchers are trying to get rid of unwanted words by crafting a more generic query, it will be interesting to see whether they subsequently fall on to a different strategy in a systematic fashion and then adopt it for future searches. By showing that the relationships between cues and query reformulation are systematic, this study offers an initial stepping stone for such further exploration.

Future research may also look into behaviors associated with cues L [too much focus], M [results too diverse], and N [different relationship]. In this study, those were too infrequent to be included in the final statistical model, but nevertheless they were part of the search repertoire observed in the sample. It is interesting that participants were noticing which terms were receiving more focus than the others, perhaps because this focus was undesired and they were hoping to see more focus on different terms. Some mentioned that they would like to indicate which terms they would like the search engine to focus on, but did not know how to perform this. It could be beneficial to study these cues, because they may raise new query formulation functionalities when it comes to helping the searchers deal with the cues they encounter.

Studying this particular set of cues will require developing tasks dedicated to exhibiting these cues as opposed to others discussed in this study.

7.3 The role of domain knowledge and search expertise

Previous research has highlighted that both domain knowledge and search expertise have an impact on search behavior. This study yields mixed results when both domain knowledge and search expertise are examined in the context of cue usage in query reformulation.

7.3.1 Domain knowledge and query reformulation

IR literature suggests that domain knowledge has influence on both initial formulation and reformulation strategies. For instance, Hembrooke et al. (2005) found that searchers with high domain knowledge used more elaboration (the level of detail and sophistication intrinsic to search attempts) in their techniques. Their searches were also complex and they incorporated significantly more unique terms than for those with low domain knowledge. At the same time, low domain knowledge searchers used less effective techniques (redundancy, plural making). In my findings, when search behavior is examined through the lens of cues, domain knowledge has only limited influence on query reformulation categories.

My findings are not directly comparable to Hembrooke et al. because of the way (re)formulation strategy was measured (reformulation categories vs. elaboration and complexity). Nevertheless, since their results showed such a significant difference between domain knowledge novices and experts, one would expect to see domain knowledge influence on more cues-reformulation type pairs. In the model devised in this study, only three cue-reformulations pairs were affected by domain knowledge. The model showed that domain knowledge played a role in relationships

between G [unwanted page type] and generalization, I [terms spread out] and phrase formation, as well as K [unwanted words] and phrase formation.

Some of these findings are rather consistent with Hembrooke et al. (2005) who also found that domain knowledge novices had more redundancy in their reformulations when compared to the domain knowledge experts. They measured redundancy by the extent to which new search terms were introduced into the queries. Similarly here, for participants with lower domain knowledge, using cue K [unwanted words], was likely to lead to phrase formation. Meaning that instead of removing terms or adding new terms that would help steer away from the unwanted context, the participants reused the same words and only added quotes and gathered the existing terms into phrases. This was also true for cue I [terms spread out], as it was also likely to lead to phrase formation for the same type of participants. Hembrooke et al. (2005) hypothesized that redundancy reflects a limited cognitive understanding about the subject matter. According to this hypothesis, limited knowledge structure complexity constrains the novice, increases the likelihood that they will resort to reusing search terms, thus lowering the number of new terms in their queries. The findings discussed above reaffirm this hypothesis as well.

Moving forward, deeper understanding of the relationship between cue K [unwanted words] and phrase formation can be instrumental both for design and training purposes. For example, one can speculate that without appropriate domain knowledge, participants have difficulties coming up with new query terms when wishing to get away from a certain word and resort to syntactic changes to the query, that make it in many cases as similar as possible to the formal question they were asked to investigate. While this may be a flaw in my research design (search task being an artificial question, as opposed to an organically developed one), or it may also be an indicator of a more substantive behavioral pattern.

A similar explanation could apply to the relationship between cue I [terms spread out] and phrase formation. It is possible that those participants tried to stick to the same familiar phrases which appeared in the description of the question, because they had difficulty using their own terms in the reformulation, since they were not familiar with any additional domain specific terms. As a result, they tended to overuse the phrase formation functionality. At this point, it is hard to offer a definitive explanation of these behaviors without discussing it further with the participants. Future research should look deeper into how different populations of searchers (defined in terms of their domain knowledge) perceive cues which were found to lead to particular types of reformulations in this study, and what they expect to achieve by those types of reformulations.

While the relationships between domain knowledge and query reformulation described above can be explained with logic offered by prior research, another observation in this study does not necessarily follow that intuition. When cue G [unwanted page type] was used by participants with higher domain knowledge, they were more likely to perform "generalization" (removing query terms), compared to other types of reformulation. Given that generalization is a rather simplistic reformulation technique, these results are inconsistent with some of the Hembrook's research, who found that searchers with high domain knowledge used more elaboration and detail in their techniques. In this study, instead of adding detail and elaboration to the queries, when cue G was present, participants with higher domain expertise made their queries more general.

One can view users with greater domain knowledge resorting to generalization in order to get away from unwanted page types as stepping back from a path that a searcher deemed as wrong for the particular task in hand, and restarting the search from a higher level of abstraction and a larger pool of search result types. Inherent to this behavior might be the searcher's confidence

that she has enough domain knowledge to make sense of that broader pool of results. Another potential explanation suggests that people who know what they are looking for, may value less sources such as forum, Q&A, travel, and news websites and prefer more formal sources. Thus, they tried to steer away from particular types of pages by removing terms they deemed as leading to those kinds of pages. For example, a participant who noticed that her query yielded news websites, wanted to get away from those by removing the terms "now" and "played", as well as a few other terms. This in turn means that highlighting the social context, as it has become popular with many search engines, may not always be consistent with people's needs and established behaviors. Further analysis of the content of the query (i.e. what kinds of words got removed in the process of generalization) and examining why exactly the searchers think removing those words may improve the results, could offer a more robust explanation to this rather counterintuitive behavior.

7.3.2 Search expertise and query reformulation

Previous research that investigates search engine expertise, as it is expressed through mental models, suggested that it may influence reformulation behavior. For example, in Holman (2011), students who had stronger mental models of search mechanisms managed to construct more complex searches (with multiple Boolean operators) compared to students with weaker mental models. In the model developed in this study, search expertise interacted with cues K [unwanted words] and I [terms spread out] with regard to query reformulation. For participants with high levels of search expertise, cue I [terms spread out] increased the odds for using the “generalization” type of reformulation and cue K [unwanted words] increased the odds for using “specialization.” On the other hand, for participants with low level of search expertise, cue K [unwanted words] increased the odds for using "generalization." A possible explanation for the

way search expertise affects those relationships, could be that search engine experts are better at spotting the right words in the results, that in their opinion take them in an unwanted direction and are able to either remove or add the words they think would lead to better results.

One interesting observation related to search expertise is that the same cue K [unwanted words] leads to completely opposite types of query reformulations - "generalization" and "specialization" – for participants with different levels of search expertise. This suggests that participants with different levels of search expertise employ different strategies to avoid unwanted words in future search results, some by removing query terms and others by adding query terms. Such a dynamic has design implications for query suggestions as a function of searchers' profile in terms of their search expertise and based on perceived desired search behavior one is willing to model.

An observation of the same cue leading to conceptually different reformulation categories requires further attention. Future research may include a different classification scheme for query reformulations. The current scheme, which is very technical, tends to classify many reformulations under the "generalization" and "specialization" categories and a more fine-grained categorization may explain this phenomenon. For instance, a reformulation that is categorized as "specialization" according to the current scheme, could be classified by a different scheme as "phrase replacement", where a two word phrase is replaced with a three word phrase. It is possible that in such a case, many of the G [unwanted page type]-> specialization relationships would be classified as G [unwanted page type]->"phrase replacement" and yield a different model. Another option would be to try a different measure of search expertise, one that would also include additional factors, for example the time spent on query reformulation (as mentioned in section 2.2.1).

It is difficult to directly compare the results of this study to Holman's conclusions. In her study, the complexity of queries was defined through the use of Boolean operators, but participants in this study didn't use those at all. The most sophisticated behavior in terms of query reformulation in my sample included phrase formation using quotation marks. If one uses phrase formation as a proxy for query complexity, the results of the model are inconsistent with Holman. In the current study, search expertise was not a moderating factor in the relationship between the cues and phrase formation. Domain knowledge was the only moderating factor in those relationships. These findings are perplexing, but as I discussed earlier, phrase formation could be the participants' way of dealing with inadequate results when they struggle to come up with additional terms, which indicates their lack of domain knowledge, not their search engine expertise. These findings raise a new question related to the way search expertise should be measured. If one wants to explore the complexity of the queries generated by searchers with varying levels of search expertise, there is a need for a different measure for search expertise, one that pertains to the ability of the searchers to formulate queries and use various operators to express their need. This is yet another avenue for future research, which should include other ways of measuring search expertise, in terms of the ability to use operators when formulating queries.

7.3.3 Re-evaluating the role domain knowledge and search expertise

As mentioned at the beginning of this section, aside from the relationships between cues and reformulations mentioned above, for majority of the relationships, domain knowledge as well as search expertise had no moderating effects. No significant effects were found on the use of cues F [expecting a word], Z [different meaning], as well as W [no cues], H [provide an idea], and U [intermediate answer] in predicting the use of specialization, generalization, and new queries.

Since previous research did not explicitly examine those relationships, one way to explain these results, is that the use of these cues as predictors of reformulation behavior is largely independent from levels of domain knowledge and search expertise. This is an important observation and a somewhat ambitious conclusion, as it undermines the fundamental view of domain knowledge and search expertise as factors that are shaping search behaviors.

One possible explanation to the relatively marginal role of domain knowledge in query reformulation could indicate that some cues are not tied to familiarity with the terminology or the ability to thoroughly understand the content of the results. Therefore, they are utilized by searchers regardless of the level of their knowledge in the specific domain. Another potential explanation may suggest that participants at all levels of domain knowledge utilized the same cues when reformulating in a certain way, but used them in a different manner. For instance, it is possible that for cue F [expecting a word], participants with high domain expertise were expecting a certain word that relates to the subject matter more than the participants with low domain knowledge. At the same time, participants with low domain knowledge were also expecting certain words to appear, but those were more general, less terminological words. It's reasonable to assume that both types of participants were using the same cue (expecting to see a certain word), but the words they were expecting to see depended on their level of domain knowledge. The current study does not cover a deeper examination of this behavior, but future research can look into whether in spite of the same cue usage, the exact words that were used by both groups did differ. It is possible that search expertise also had no effect in this relationship, because for many of the instances in which the participants expected to see certain words, it didn't matter how much search expertise they had, there were no tools available to emphasize names or dates they were expecting to see.

By the way of another example, cue Z [different meaning] could have a similar explanation and it's possible that both high and low domain knowledge participants were using this cue, but the query terms that had a different meaning than intended, were not the same type of query terms. It could be that participants with high level domain knowledge were noticing the different meaning of query terms more specific to their domain expertise, while participants with low level domain expertise were noticing the different meaning of general query terms which were not domain specific. Since this study did not go into this much detail regarding which words exactly were expected by each type of participants, it is hard to determine whether this is indeed the case and further research is required. In terms of search expertise effect, there is a possibility that query disambiguation is a rather intuitive methods and could be performed in the same manner both by search engine expert and novices.

All other cues that are independent of domain knowledge and search expertise levels are those that lead to the reformulation category 'new'. Those are W [no cues], H [provide an idea], and U [intermediate answer] and it may not be a coincidence that they all lead to the 'new' category. These cues lead to a change in search direction significantly, a step that can be taken both by domain experts and novices. The same as a participant with low domain knowledge may not notice any cues in the results and move on to the next query, a participant with high domain knowledge may not notice any important cues as well and have the same outcome of changing the query completely. Both types of participants may have an intermediate question in mind, such as: "Is Rugby Australian" and get an answer which would take their further reformulations in a completely different direction. In terms of search engine expertise, it is reasonable to believe that since the response to these cues was simply changing the query significantly (or generating a completely new query) and didn't require any advanced expertise in query formulation or

understanding of how the search engine words, this change could be done equally well both by participants with high and low search engine expertise.

What unites all these cues is that they are rather broad. Not something specific one can point to on the page, but something that has to do with the mental process in the searcher's head. It's hard to systemize those in order to have any specific design suggestions, but it is important to pay attention to those stepping stones when talking about searching literacy. This means that searchers of all levels of domain knowledge and search expertise can benefit from learning how to identify those cues more quickly, so they can move on to the next direction in their search sooner. In addition, in terms of search engine design, highlighting in the results words that could possibly constitute an answer to the question written in the query, could help the searcher identify the answer that relates to cue U [intermediate answer] faster.

In summary, the explanation to lack of influence of domain knowledge and search expertise can be that people have established search micro-routines that have developed over time and are "ways to do things" in a semi-subconscious fashion. Those are tacit behavioral patterns, which the searchers perform automatically, and into which they fill content depending on the question in hand and contextual factors such as search expertise and domain knowledge. So, for example, when the searcher sees nothing to build on in the search results (cue W), she just starts a new query. Alternatively, when she does not see a specific word she expected in the search results (cue F), the searcher tries to focus her search in the direction of the expected word, by making the query more specialized. In performing any of these behaviors the searcher may use different vocabularies or techniques depending on their domain knowledge and search expertise, but the very basic behavioral patterns persist. Future research should examine such micro-behaviors and

routines involved in the search process. With that, knowing that those relationships exist already has potential implications for SERP design.

It is important to note that before reaching decisive conclusions, it is essential to consider the methodological limitations of this study, because these observations can also be explained by the way that query reformulation categories were constructed. My classification scheme for reformulation categories did not take into account any usage of terminology or complexity of reformulation, but rather a simpler, more mechanic manipulation of the query of adding/removing words. This level of abstraction of the reformulation behavior could be the reason for such limited effect on the relationships between the cues and the reformulation categories. Future research should account not only for the mechanics, but also for qualitative use of language in query reformulation. Again, this study offers an initial "map of the land," future research needs to add granularity to this mapping and explore additional operationalizations.

7.4 Cues and search effectiveness

Exploring the link between the use of cues and search effectiveness proved to be challenging in this study. By definition, reformulations occur when searchers are not satisfied with the retrieved results. Thus, unsurprisingly, since each reformulation was preceded by at least one cue, many cues in the model had a negative effect on search effectiveness. Some of these effects were amplified by domain knowledge and search expertise. In other words, from the onset, the mere use of cues and reformulations in each case suggests that the participant is struggling with finding an answer to her task.

Against that background, it is particularly interesting to note one cue that helped participants with low search expertise to find better answers. On the one hand, participants with low search expertise were trying to steer away from queries that have repeatedly yielded similar results (cue J), which helped them find the right answers. On the other hand, for participants with high search expertise the same cue actually reduced the chances to get a correct answer. Both observations highlight that searchers are not a homogenous group and that information communicated through SERP design for the less experienced and the less knowledgeable searchers may have a significant impact on their search experience. Thus, practically speaking, searchers with low search expertise would benefit if search results that keep repeating are highlighted and a way to steer away from those results, is suggested. For searchers with high search expertise, such service may be unnecessary.

Regardless of search expertise and domain knowledge, usage of some cues such as U [intermediate answer] and K [unwanted words] had a negative effect on search effectiveness. This is rather counterintuitive, since this means that finding an intermediate answer doesn't necessarily grant the participant with the ability to find the final answer to the question. Such behavior may be due to the fact that intermediate answers steered the participant away from the main question, by looking at intermediate questions and their answers, thus lowering her chance of finding the actual answer; this is in comparison to the participants who stuck to the same path and did not get destructed by intermediate answers. A similar explanation could possibly apply to cue K leading to unsuccessful searches. It could be that the attempt to steer away from unwanted words leads to a significant deviation from the path towards, and therefore the ability to find, the correct answer.

Regardless of cue usage, search expertise as it was measured in this study, was found to have a positive effect on search effectiveness. Since previous literature hasn't explored search effectiveness as a function of search expertise measured through mental models of search engines, it is difficult to compare these findings to other studies, which used a different method to measure search expertise (such as: Hargittai & Hsieh, 2012; Hölscher & Strube, 2000; Hsieh-Yee, 1993; Jenkins et al., 2003; Lazonder et al., 2000; Saracevic et al., 1988a).

In terms of domain knowledge, it was found that observing cue F [expecting a word] is likely to lead to incorrect answers, but only for participants with high domain knowledge. This finding requires more research, since it could be that if certain expected words were highlighted (such as names, dates, etc.) the outcomes would be different both for high and low level domain knowledge. In the future, a study can be designed which will examine the effect of highlighting certain elements on the page on the outcomes of search by domain knowledge experts and novices.

As mentioned in the discussion section above, domain knowledge on its own, regardless of cue usage did affect search effectiveness. It makes sense, because searchers with certain domain knowledge level are expected to succeed more, because they have a deeper understanding of the content and the terminology. Consequently they are probably able to utilize this understanding to eventually find the answer, regardless of the cues they use. In a way, these observations are in line with the inconclusive results of previous research. As Wildemuth (2004) reviewed in her work, some researchers found that the effect of domain knowledge on search success is positive, while others did not find such relationship.

7.5 Design and training implications

In addition to its conceptual contributions described above, this study also highlights potential modest design and training implications. Ultimately, those should be considered in tandem as at its core, search is a socio-technical endeavor. In other words, improving search requires attention from multiple partners – particularly those who design the technology and those who use it.

Some potential design implications arise directly from the type of cues that were discovered in this study. Most implications focus on presentation of search results on the SERP. For instance, Cue F [expecting a word] represents a situation when participants scanned the results (on SERP or on the full pages) for certain words, which were not part of the query. This cue was found significant when leading to specialization and could be explained by the participants' attempt to make the words they were scanning for, more visible, by additional query terms. This relationship suggests that highlighting certain words which are not part of the query, but relevant to the particular search, could be useful to the searcher. For example, if the searcher is using a query in a form of a question, then the possible answer can be highlighted. More specifically, if a query starts with the phrase "when did...," highlighting any date that appears in the results may be potentially useful for the searcher. Similarly, in the case of Cue J [similar results], it may be useful for the searcher to see the repeated or similar results highlighted on the SERP (with potential suggestions for a reformulation type to steer away from such, presumably unwanted, results).

Another set of implication is related to query terms and query syntax suggestions. For example, in cases where the cue can be identified automatically, such as cue I, which indicates that some of the query terms are spread out and suggest using quotes around those terms. In cases where the cue cannot be identified automatically, for instance cue G [unwanted page type], one

suggestion would be to solicit explicit feedback - a clarification for cases in which the results contain different types of pages. For example, if the results include both scientific pages and travel pages, the search engine can detect the type of web-pages in the results and present the searcher with a question, inquiring about the type of pages she intended to receive.

Finally, with the growing personalization of the search experience, the third set of design implications highlights the simple truth that searchers are not a homogenous group. Thus, provided that the search provider "knows" its users, at least in terms of their search expertise, it can offer or highlight different cues for query reformulation based on the user type. Thus, for example, while highlighting cue J [similar results] may help users with low search expertise improve their search effectiveness, such design may yield the opposite results for the more experienced searchers.

With regard to training, searchers may benefit from better familiarity with search engine operators (in this particular case Google's). In other words, the searchers need not only a strong mental model of how the search engine is functioning, but also a clear picture of what the interface of communicating with mechanism looks like. As this study showed, many participants wanted to "get away" from certain words and certain types of pages, but very few of them really knew how to exclude words from their consequent searches. Until we have design solutions that make use of operators effortless for the user, it falls on the shoulders of digital literacy experts to educate the public about tools for effective search.

7.6 Limitations and future research

As any other study, this work has certain, primarily methodological limitations, and as is the case with any other study, those limitations offer potentially fruitful avenues for future research.

Some of the limitations of this project are rather generic and standard for a lab study; other limitations are more specific to the methodology used and to the particular research design and operationalization decisions. Suggestions for future research, thus, are partially focused on mitigating those methodological limitations. More suggestions, however, focus on the substantive findings of this research.

First, starting with the generic and standard limitations of a lab study, the participants in this research were all university students and therefore, the findings of this study can be generalized only to people with characteristics similar to that demographic (e.g. young people from affluent background, who have continuously used information technology in their daily routines for most or even all of their lives). Future research will benefit from a more heterogeneous sample, representative of the general population of search engine users. Although recruiting members of the public for lab experiments is more costly than working with student participants, such study will benefit from a greater external validity. Similarly, the artificial environment of a lab and a customized SERP must have affected the behaviors of my participants. Future work might benefit from observing searchers in their "natural" settings in terms of their working environment, the technology that they use, and the nature of questions motivating their search (more on that later).

In addition, since this research was performed on a PC, its findings can't necessarily be generalized to mobile platforms. First, search on mobile devices occur in different settings, compared to search on a PC, and it is likely to be motivate by a different information need. Second, mobile search result snippets are smaller and can fit only some of the content and highlighted query terms compared to the results on a PC. Therefore, cues that are noticed by participants on a PC, will not necessarily be noticed on a mobile device. Finally, since typing on

a mobile device is more limited than on a PC, it can also affect the amount and the type of reformulations performed. As a result, more research is needed to be done in order to compare the way cues are used on a mobile device as opposed to a PC.

Another limitation stems from the use of the *Stimulated Recall* procedure. Here I ran the risk that by the time the participants had finished performing the 5 tasks, they would not necessarily remember their precise thought process at the moment of each query reformulation. The length of the participation, combined with the cognitive load of dealing with difficult tasks and the overall low domain knowledge, could have potentially impaired the participants' ability to recall parts of their reasoning. Showing participants the recording of their search and thus stimulating the recall, should have ameliorated this problem to a certain degree. Future research, however, will benefit from relying on real-time measurement of cue-related activities. Eye tracking, for example, can be used to compare the cues reported by the participants, with the actual elements on the page they looked at. This would allow verifying at least for some of the cues that the cues which were mentioned by the participant were actually looked at by them.

The low cost of reformulation, although accurately capturing the real world behavior, imposed another potential limitation on the quality of recall. One of the observations in this study showed that when faced with a difficult question, participants often scrambled and instead of slowing down their search and paying more attention to the actual results they were getting, they preferred to move from one reformulation to the next, impatiently waiting for the answer to just 'pop up', without dedicating much thought in between the reformulations. The Stimulated Recall process forced the participants to analyze the search results and interpret them with regard to the previous query, but it is unclear as to how much attention searchers pay to the cues when they search in their natural environment. This could explain why the cues elicited in this study were

not as sophisticated as one would hope they would be (for example: position of the query terms on the SERP, parts of speech of the query terms, etc.) This type of behavior has implications on how the searchers' thought process can be studied. Given that the cost of query reformulation is so low (results are returned instantly), this contributes to the frantic searching. As recent research by Azzopardi, Kelly and, Brennan (2013) showed, subjects who used an interface that required high physical cost to compose a query, were more thoughtful about their query reformulation. They submitted significantly fewer queries, spent more time on search results pages, examined significantly more documents per query, and went to greater depths in the search results list. Therefore, future studies could benefit from limiting the ease of reformulation and by this enabling a more developed thought process not only in the Stimulated Recall stage, but also during the search session as well.

The search tasks in this study, although thoroughly tested, have some inherent limitations as well. All the tasks this study came in a form of a question, which means that my findings can only be generalized for tasks posed as questions. Although the option of using different types of tasks was considered, given that the pilot and the pre-test showed that among the types tested, this was the most suitable type of task for generating multiple reformulations and sufficient evaluation of the results. I decided to control for the type of task and use a single type tasks in a form of a question. Moving forward, it would be important to test additional types of tasks that, if found equally suitable for discovering cues, could be incorporated in future studies and compared to question type tasks.

On a related note, using difficult tasks, instead of complex tasks (in terms of the number of paths involved while engaging in the task) means that in this study the reformulations were mostly triggered by the participant's inability to find the answer. Alternatively, if complex tasks were

used, then the trigger for reformulation would have been different, most likely in order to generate a parallel path to satisfy another facet of the task. This would make a difference in the way cues would have been used, because those query reformulation would not be based off previous search results, but rather an independent parallel path to deal with one of the facets of the task. As before, future research should examine the relationship between difficulty and complexity of the tasks and the use of cues in query reformulation.

Finally, in terms of technical limitations, the way I operationalized a number of variables offer a fruitful ground for future exploration. For example, alternative ways of conceptualizing and measuring search expertise and domain knowledge could potentially validate my or lead to different results. As mentioned earlier, in terms of search expertise, the mental model evaluated in this study is more of a "search engine mental model" - how a search engine works in general, instead of a "query reformulation mental model", because it doesn't inquire about the participants' knowledge on how to use query syntax and reformulation techniques. Therefore, future studies should involve more comprehensive mental models of search expertise, those that would also measure the searchers' familiarity with query formulation and reformulation techniques.

With regard to domain knowledge, which was self reported, its limitation has to do with the fact that participants may have estimated their own domain knowledge inaccurately. Particularly, given the fact that they were filling out the questionnaire after completing the search session, which was challenging and may have lowered their self esteem and self evaluation regarding their domain knowledge, their self reported domain knowledge might have been underestimated. One way to improve it in the future and allow a better reflection of the participants' real domain knowledge is to measure it by adding knowledge questions for each of the five topics and

administering this questionnaire prior to the search session. Another approach can include additional factors, such as time spent on reformulation, as part of the search expertise measure.

Another potential limitation and an opportunity for future research stems from my operationalization of query reformulation. As I notice earlier, this study uses a rather technical definition of query reformulation. Thus, it is limited to relying on technical changes such as adding, removing, or replacing words, but it does not unpack the meaning of the queries. Given the provocative nature of some of the findings, which are inconsistent with previous research, it will be beneficial to conduct additional analysis of existing data, one that takes into account the terminology used in the query or complexity of reformulation. Moreover, more detailed analysis of the meaning of the query will help better explain newly observed phenomena, such as behavior of searchers with high domain knowledge, who use generalization in order to steer away from unwanted types of web pages in their results.

Beyond, technical limitations that signal fruitful avenues for research, this study puts forward a series of substantive claims that warrant future investigation. Particularly interesting is the general sense that domain knowledge and search expertise are not as significant in query reformulation as earlier research suggests. Further exploring this claim will require examining queries beyond their technical definition. Future research should look into whether there is a significant difference in terms that more knowledgeable or more experienced users operate with. Along the same lines, future research should look into the expectations of these different groups of searchers from the search results, because it is against that expectation that they evaluate the SERP and the target pages.

Echoing more of Oddy's (1977) dialogical approach, future research should take a more comprehensive view of the query reformulation process. My current approach treats this process

as a collection of standalone acts of reformulation. It does not account for the learning that occurs as the searcher moves from one query to another and does not account for potential path dependency resulting from the searcher building on a set of results as she continues her quest. Related to this, future research should take a deeper look into the meaning making by searchers. i.e. not only what cues they rely on, but how they interpret those cues. In such inquiry, a researcher should also pay close attention to the context in which cues are interpreted (being it the task, the search environment, the information need or any other contextual factor). Finally, such a comprehensive view calls for an understanding that the searchers are not a homogenous group and may rely on various repertoires of interpretation and practices when they make sense of queries and search results or when they reformulate. Utilizing a more comprehensive view of the search process may explain puzzling cases observed in this study, such as when the same cue leads to different reformulation behaviors.

An even more comprehensive view of the search process would view search as a process that unfolds across multiple search sessions and over a lengthy period of time. It would also account for the micro-practices – potentially unconscious behaviors acquired by the searchers through their life-long experience of interacting with technology. Designing a study that those factors into account will require a longitudinal analysis on the one hand, and, on the other, additional metrics that will capture micro-behaviors. Panel data that incorporate biometrical measures, such as eye tracking or cognitive activity, would be a treasure trove for a researcher undertaking this task.

7.7 Searching as thinking

Search has become an integral part of our online, and in many cases even offline, experiences. Today, people engage in search not only on their computers, but also on their phones and other handheld devices. As information technology continues penetrating various aspects of our lives, understanding people's search behavior is pivotal not only for designing better search experiences, but also for understanding how search is being integrated into the social fabric. Yet, search behaviors seem to be tacit, almost subconscious, and thus very difficult to study systematically. When observing searchers' behavior online it is often difficult to separate their thought process from their query reformulation behavior – people search as they think. Unpacking this link, however, is essential for moving the field of IR to its next frontier. This study offers one of the first glimpses into the decision-making mechanisms behind query reformulation. The conceptual and methodological constructs put forward through this work should be instrumental for future research and the findings discussed in this study should offer a rich map for furthering user-centric research in IR.

8 Appendix 1 – Tasks and instructions used in the pilot

Participant #1:

Please use Google's search engine to find answers to the following questions:

1. What is the name of an iPhone app that schedules a shutdown of other apps at a certain time in the day
2. How much caffeine is there in maple syrup?
3. How to prevent a cat from getting asthma?
4. What are the names of dinosaurs that had a small skull, big body structure, and used to eat plants?

Participant #2:

Please use Google's search engine to find answers to the following questions:

1. What types of crimes are committed by people who have been previously convicted and later released or paroled from prison before committing this crime?
2. What parents are doing to prepare for spiraling cost of college tuition?
3. What actions are being taken to make the quality of children's television in the U.S better
4. What actions are being taken by U.S. airplane manufacturers to improve the safety of their passenger aircraft?

Participant #3:

Please use Google's search engine to find answers to the following questions:

1. Why were camels domesticated?
2. Into what languages has the book "Bad Science" by Ben Goldacre been translated?

3. Where can I find a list of congress members of all times, which includes information about their ethnic races?
4. What is the name of an iPhone app that schedules a shutdown of other apps at a certain time in the day

Participant #4:

Please use Google's search engine to find answers to the following questions:

1. How much caffeine is there in maple syrup?
2. What types of crimes are committed by people who have been previously convicted and later released or paroled from prison before committing this crime?
3. What actions are being taken to make the quality of children's television in the U.S better
4. Why were camels domesticated?

Participant #5:

Please user Google's search engine to find information sources that support the following statements:

1. It is difficult to produce containers that maintain the freshness of vegetables during shipping
2. Fishermen find it difficult to earn net profit
3. Mints and treats that look like coins are favorite holiday candies
4. It is difficult to secure a mortgage or insurance for property directly on the bank of a river
5. For security, conductors carry radios as they move between stations

Participant #6:

In your job as a trainee, you support the journalists at the newspaper. Your responsibility is to find information about the journalists' article topics. Today you need to search for good sources of information about four different topics that the journalists are working on. Please find as many good information sources as possible. A good information source is a source that you could and would use to get information about the topic. Any source with information that will inform the journalist on the topic can be considered a "good" source. You need to find as many "good" information sources as you can, but it is also important to avoid choosing information sources that are not good. Please user Google's search engine to find good information sources.

1. Drinking water helps you stay well
2. Mints and treats that look like coins are favorite holiday candies
3. It is difficult to produce containers that maintain the freshness of vegetables during shipping
4. Fishermen find it difficult to earn net profit

Participant #7:

Please use Google's search engine to find web pages that support or refute the following statements:

1. Mints and treats that look like coins are favorite holiday candies
2. It is difficult to produce containers that maintain the freshness of vegetables during shipping
3. Fishermen find it difficult to earn net profit
4. It is easy to tire when driving a car

Participant #8:

Please user Google's search engine to find web-pages that support the following statements:

1. It is difficult to secure a mortgage or insurance for property directly on the bank of a river
2. For security, conductors carry radios as they move between stations
3. In some cultures it is common to hire a band for the wedding
4. It is difficult to produce containers that maintain the freshness of vegetables during shipping

Participant #9:

Please use Google's search engine to find answers to the following questions:

1. Why is it difficult to secure a mortgage or insurance for property directly on the bank of a river?
2. Why is it difficult to produce containers that maintain the freshness of vegetables during shipping?
3. For what purposes do train conductors in the USA carry radios as they move between stations?
4. Why do fishermen have difficulties to earn net profit?

Participant #10:

Please use Google's search engine to find as many web-pages as possible that contain the answer to the following questions:

5. What are the difficulties of producing containers that maintain the freshness of vegetables during shipping?
6. Why is it difficult to secure a mortgage or insurance for property directly on the bank of a river?
7. Into what languages has the book “Bad Science” by Ben Goldacre been translated?
8. Where can I find a list of members of the US Congress of all times, which includes information about their ethnicity or race?

9 Appendix 2 – The protocol used in the pilot study

The protocol used in the pilot study (used with most participants, first few participants got a slightly different protocol, in which the first "why" question was not asked):

[Briefing of the participant:] This is a study of how people use search engines. I will give you 4 search tasks, please perform the search by using Google's search engine. Software installed on this computer will capture everything that you do during this session. After you are done with all the tasks, you and I will go over the recording and I will ask you some questions about your decision-making during the search process. It is important to note that although you need to try and find an answer to the search tasks that will be presented to you – this is not a test and I'm more interested in the process that you're going through when performing the search and not how well you perform.

[Make sure that Google's search engine is not logged in with any user. Delete browser cookies. Turn off recording from microphone. Make the pointer green.]

[For the purpose of this pilot and task testing, only 4 different questions will be used per participant, from a pool of 20 different tasks and 5 different types of instructions]. Please copy and paste (into a notepad file) the URLs of the pages that in your opinion answer these questions. [While the participant is searching, sit behind the participant and start timing once the screen capture begins. Record the exact timestamp of every reformulation to allow going back to it later in the questioning stage. Create separate recordings for every task (start recording when the task begins and stop when it ends.]

[After the participant is finished with the tasks, before the questioning begins, explain the following terminology to the participant: query terms – the words that you typed in the query;

results page – the page that you get after running a query with the search engine; full webpage – the webpage followed from one of the search engine results]

General Questioning:

[Make sure that cursor is red. Make sure that recording from microphone is turned on].

[Replay the recording of each task, followed by general questions the purpose of which is to elicit cues without any guidance/leading of the participant. Start recording with screen capture, including voice.]

After reformulation has been presented in the recording – for every query revision, ask:]

- **Why** did you make the change in your query at this point?
- **How** did you know that you needed to make the change at this point? Please give me specific examples in the recording, of what helped you make this decision. Please point with mouse cursor at these examples on the results page or the full webpage.
- **What exactly on this page** (results page or full webpage) provided a hint that you needed to change the previous query? Please point with mouse cursor at specific examples on the results page or the full webpage, of what helped you make this decision. Additional probing: please point at the feature or the element that helped you realize that you needed to reformulate.
- **What exactly on this page** (results page or full webpage) helped you decide on what changes needed to be made to the previous query? Additional probing: please point with the mouse cursor at the feature or the element that helped you to reformulate.

Cue-specific Questioning[26]**:**

[Show a printed list of the specific cues (appears on the next page) to the participant and while playing the recording and ask:] Please give an example of each of these cues (if exist) **before the query revisions** that appear in the recording, in order to demonstrate how it helped you decide on the changes that needed to be made to the previous query. Please show on the screen by pointing with the mouse pointer. On a scale of 1 to 5, how important was each cue for this specific reformulation?

[If any time is left, ask the participant if s/he has any comments regarding the pilot itself, whether or not there was something confusing, annoying or unclear about the instructions or the procedure in general].

The list of printed cues to be presented in the cue-specific questioning part will include:

- **Distance** between the query terms on the results page or the full webpage.
- **Absence of** query terms or other words that you expected to appear in the results page or the full webpage.
- Words that are **immediately preceding or following** your query terms in the results page of the full webpage.
- **Parts of speech** of the query terms that appear in the results page or full webpage.
- **Position** of the query terms on the results page (in the body of the snippet, in the title, or the URL).
- **Order of appearance** of the query terms on the results page or the full webpage.

[26] As mentioned earlier, this part was used only in the pilot and will be removed from the protocol of the actual study.

- Can you think of other cues that you may have used in this session when reformulating a query?

10 Appendix 3 –The questionnaire used in the pilot study

A pilot study on how searchers interact with web search engine results

General Background Questionnaire

Your Gender: ( ) Male ( ) Female

Age: _____

If English is not your native language, how would you rate your fluency in English on the scale of 1 to 5 (1 being not fluent at all and 5 being as fluent as a native speaker): _____

How many years have you been using search engines? _____

What is your preferred search engine? _____

11 Appendix 4 – Consent form used in the pilot

A pilot study on how searchers interact with web search engine results

Study consent form

You are invited to take part in a pilot research study about how searchers interact with search engine results. We asked you to participate because we want you to show us how you perform search tasks, interpret the results and interact with them. Please read this form carefully and ask any questions you may have before agreeing to take part in the study.

What the study is about: The purpose of this study is to learn how searchers interpret web search engine results and interact with them in order to make decisions regarding their future searches. We are interested in observing the ways in which you perform search tasks with the use of a search engine and will ask you some questions regarding your decision making process.

What I will ask you to do: If you agree to be in this study, I will ask you to participate in a session, envisioned to last about an hour and a half to two hours. We will start with a list of 4-5 search tasks that I will provide you and you will need to perform with Google's search engine. Software installed on the computer will capture everything that you will do during this session. Once you are done, we will go over the captured recording and I will ask you questions regarding your decision-making process. I will ask you to point to some elements on the screen to demonstrate the process you went through when making these decisions. This questioning step will also be captured with a screen capturing and your answers will be audio recorded. Finally, I will ask that you complete a short questionnaire with some demographic data about yourself and about your experience with search engines.

Risks, benefits, and compensation: I do not anticipate any risks for you participating in this study, other than those encountered in day-to-day life. The study will not have any direct benefits for you other than a sense of contribution to advancement of scientific research. In addition, you will receive extra credit in IST 637 equivalent to one homework assignment as a token of appreciation for your participation in this study. If you decide to withdraw in the process, you will be compensated with extra credit pro-rated for the amount of work completed.

Taking part is voluntary: Taking part in this study is completely voluntary. You may skip any questions that you do not want to answer. If you decide not to take part or to skip some of the questions, it will not affect your current or future relationship with Syracuse University. If you decide to take part, you are free to withdraw at any time.

Confidentiality: Your responses will remain confidential. The information from the recordings and questionnaires will be aggregated for purposes of analysis and reporting. Specific quotes or instances of your behavior may be used in future publications. In those cases we will use aliases and at no point any identifiable information about you will be released. The records of this study will be kept private and all the video and audio recordings will be used for research purposes only. All data will be securely stored on several hard disks. Hard copies of data will remain in my office. All data will be destroyed (i.e., shredded or erased) when their use is no longer needed but not before a minimum of two years after the completion of the study.

I**f you have questions:** The researchers conducting this study are Veronica Maidel and Howard Turtle. Please ask any questions you may have now. If you have questions later, you may contact Veronica Maidel at vmaidel@syr.edu. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) at 315-4433013 or access their website at http://orip.syr.edu/humanresearch.php.

You will be given a copy of this form to keep for your records.

Statement of Consent: I have read the above information, and have received answers to any questions I asked. I am 18 years or older and I consent to take part in the study.

[ ] I agree to be audio taped and for the screen to be captured.

[ ] I do not agree to be audio taped and for the screen to be captured.

Your Signature _____ Date _____

Your Name (printed) _____

Signature of person obtaining consent _____ Date _____

Printed name of person obtaining consent _____ _____

This consent form will be kept by the researcher for at least three years beyond the end of the study and was approved by the IRB on 03/30/2011.

12 Appendix 5 – Tasks and instructions used in the pre-test

Please use Google's search engine to find answers to the following questions[27] (try to find as many sources as possible):

[copy and paste the **answer** and the **URLs that contain the answer** into the word document]

1) How do you find out the date of the most recent hard drive formatting on a PC machine?

2) What Australian sport was adopted in Britain?

3) What minerals can float in salt water?

4) When was the Jominy test invented?

5) What is the partial pressure of oxygen at an altitude of 5000 feet?

6) How do you find the exact time a hard drive was last formatted on a PC?

7) What animal is smaller than a bear but eats a plant called bearberry?

[27] The first ten participants (the pre-test) participants received different questions (some got 4 and some got 5) in order to identify the most suitable questions for the actual study, the number of questions, and the right phrasing that won't be confusing to the participants (the hard drive question was phrased in two different ways).

13 Appendix 6 - The protocol used in the pre-test

Study Procedures for "A study on how people interact with web search engine results":

[Briefing of the participant:] This is a study of how people use search engines. You will receive 4 questions, please find the answers to these question by performing a Google search. Software installed on this computer will capture everything that you do during this session. After you are done with all the tasks, you and I will go over the recording and I will ask you some questions about your decision-making during the search process. It is important to note that although you need to try and find an answer to the questions that will be presented to you – this is not a test and I'm more interested in the process that you're going through when performing the search and not how well you perform. I should note[28] that every searcher usually goes through a different process - one example of such process would be starting with an initial query that the user thinks may lead her to the answer, looking at the retrieved summaries and web-pages and then revising the query based on these results until the answer is found.

[Make sure that : 1) Google's search engine is not logged in with any user, 2) instant search is disabled, 3) the URL that disables auto-complete is used, 4) the order of the tasks has been randomized, 5) browser cookies and history have been deleted. Turn off recording from microphone.]

Please use Google's search engine to find as many web-pages as possible that contain the answer to the following questions [present a page with the 4 questions printed out]. Please copy and

[28] The purpose of this note is to assure the participant that reformulation does not mean that they failed, but rather a part of the natural process that all searchers go through. The idea is to prevent the participants from stopping themselves from reformulating, just because they are being watched and assessed.

paste (into the notepad file) the URLs of the web-pages that in your opinion answer these questions.

[While the participant is searching, use the screen capturing software on a parallel computer to observe the search process and to mark time stamps of query reformulation points. Make sure to mark the initial query formulation time-stamp along with the reformulation points].

[Once the participant is done with all the tasks, before the questioning begins, explain the following terminology to the participant: query terms – the words that you typed in the query; results page – the page that you get after running a query with the search engine; full webpage – the webpage followed from one of the search engine results]

Questioning for exploratory elicitation of cues:

[Set the mouse cursor to yellow. Make sure that recording from microphone is turned on].

[Navigate in the video, to the time when the initial query was made and then to every reformulation point in each task. After reformulation has been presented in the recording – for every query revision, ask the following question (if needed, allow the participant to scroll the video back or forward, in order to answer the question):]

- ***Why*** did you make the change in your query at this point?
- ***How*** did you know that you needed to make the change at this point? Please give me specific examples in the recording, of what helped you make this decision. Please point with mouse cursor at these examples on the results page or the full webpage.
- ***Is there anything on this page*** (results page or full webpage) that provided a hint that you needed to change the previous query? If so, please point with mouse cursor at specific examples on the results page or the full webpage, of what helped you make this

decision. Additional probing: please point at the feature or the element that helped you realize that you needed to reformulate.

- ***Is there anything on this page*** (results page or full webpage) that helped you decide on what changes needed to be made to the previous query? Additional probing: please point with the mouse cursor at the feature or the element that helped you to reformulate.

14 Appendix 7 - The consent form used in the pre-test and the actual study

Consent form for a study on how people interact with web search engine results

You are invited to take part in a research study about how searchers interact with search engine results. We asked you to participate because we want you to show us how you use Google's search engine, interpret its results and interact with them. Please read this form carefully and ask any questions you may have before agreeing to take part in the study.

What the study is about: The purpose of this study is to learn how searchers interpret web search engine results and interact with them in order to make decisions regarding their future search-related actions. For this purpose, we are interested in observing the ways in which you use Google's search engine to find answers to certain questions and would like to observe your decision making process and talk to you about it.

What I will ask you to do: If you agree to be in this study, I will ask you to participate in a session, envisioned to last one and a half to two hours. We will start with a list of 4-5 questions the answer to which you will need to find with Google's search engine. Software installed on the computer will record your actions during this session. Once you are done, we will review the recording and I will ask you questions regarding your decision-making process. I will ask you to point to some elements on the screen to demonstrate the process you went through when making these decisions. The interview will be captured with screen capturing software and an audio recording. Finally, I will ask that you complete a

short questionnaire with some demographic data about yourself and about your experience and knowledge.

Risks, benefits, and compensation: I do not anticipate any risks for you participating in this study, other than those encountered in day-to-day life. Your participation in the study will contribute to scientific knowledge and to our understanding of human search behavior. In addition, you will receive $15[29] as a token of appreciation for your participation in this study, plus you will get a chance of about 1 in 25 to win a $50 Amazon gift card. If you decide to withdraw in the process, your compensation will be pro-rated for the amount of work completed.

Taking part is voluntary: Taking part in this study is completely voluntary. You may skip any questions that you do not want to answer. If you decide not to take part or to skip some of the questions, it will not affect your current or future relationship with Syracuse University. If you decide to take part, you are free to withdraw at any time.

Confidentiality: Your responses will remain confidential. The information from the recordings and questionnaires will be aggregated for purposes of analysis and reporting. Specific quotes or instances of your behavior may be used in future publications. In those cases we will use aliases and at no point will any identifiable information about you be released. The records of this study will be kept private and all the video and audio recordings will be used for research purposes only. All data will be securely stored on several hard disks. Hard copies of data will remain in my office. All data will be destroyed (i.e., shredded or erased) when their use is no longer needed but not before a minimum of two years after the completion of the study.

I**f you have questions:** The researchers conducting this study are Veronica Maidel and Howard Turtle. Please ask any questions you may have now. If you have questions later, you may contact Veronica

[29] In the pre-test, the compensation amount was $10 and then increased to $15 for the actual study to increase response rate.

Maidel at vmaidel@syr.edu. If you have any questions or concerns regarding your rights as a subject in this study, you may contact the Institutional Review Board (IRB) at 315-4433013 or access their website at http://orip.syr.edu/humanresearch.php.

You will be given a copy of this form to keep for your records.

Statement of Consent: I have read the above information, and have received answers to any questions I asked. I am 18 years or older and I consent to take part in the study.

[ ] I agree to be audio taped and for the screen to be captured.

[ ] I do not agree to be audio taped and for the screen to be captured.

Your Signature _____ Date _____

Your Name (printed) _____

Signature of person obtaining consent _____ Date _____

Printed name of person obtaining consent _____ _____

15 Appendix 8 - The tasks used in the actual study

Please use Google's search engine to find answers to the following questions (try to find as many sources as possible):

[copy and paste the **answer** and the **URLs that contain the answer** into the word document]

6) What minerals can float in salt water?

7) What Australian sport was adopted in Britain?

8) What is the partial pressure of oxygen at an altitude of 5000 feet?

9) When was the Jominy test invented?

10) How do you find the exact time a hard drive was last formatted on a PC?

16 Appendix 9 - The protocol used in the actual study

Study Procedures for "A study on how people interact with web search engine results":

[Briefing of the participant:] This is a study of how people use search engines. You will receive 5 questions. Please find the answers to these questions by performing a Google search. Software installed on this computer will capture everything that you do during this session. After you are done with all the tasks, you and I will go over the recording and I will ask you some questions about your decision-making during the search process. It is important to note that although you need to try and find an answer to the questions that will be presented to you – this is not a test and I'm more interested in the process that you're going through when performing the search and not how well you perform or if you find the answer. I should note[30] that every searcher may go through a different process. For example, you may start with an initial query that you think may lead to the answer, then look at the retrieved summaries and web-pages and then revise the query based on these results until the answer is found. Or, you may have your own way of doing it that's different from this. I appreciate learning about how ***you*** go about finding the answers.

[Make sure that : 1) Google's search engine is not logged in with any user, 2) instant search is disabled, 3) the URL that disables auto-complete is used, 4) the order of the tasks has been randomized, 5) browser cookies and history have been deleted. Turn off recording from microphone.]

[30] The purpose of this note is to assure the participant that reformulation does not mean that they failed, but rather a part of the natural process that all searchers go through. The idea is to prevent the participants from stopping themselves from reformulating, just because they are being watched and assessed.

Please use Google's search engine to find as many web-pages as possible that contain the answer to the following questions [present a page with the 5 questions printed out]. Please copy and paste (into the notepad file) the URLs of the web-pages that in your opinion answer these questions.

[While the participant is searching, use the screen capturing software on a parallel computer to observe the search process and to mark time stamps of query reformulation points. Make sure to mark the initial query formulation time-stamp along with the reformulation points].

[Once the participant is done with all the tasks, before the questioning begins, explain the following terminology to the participant: query terms – the words that you typed in the query box; results page – the page that you get after running a query with the search engine; full webpage – the webpage followed from one of the search engine results]

Questioning for exploratory elicitation of cues:

[Set the mouse cursor to yellow. Make sure that recording from microphone is turned on].

[Navigate in the video, to the time when the initial query was made and then to every reformulation point in each task. After reformulation has been presented in the recording – for every query revision, ask the following question (if needed, allow the participant to scroll the video back or forward, in order to answer the question):]

I am going to ask you a few questions about the points in which you made changes to your query, so that you can walk me through your decision making process.

- ***Why*** did you make the change in your query at this point?
- ***How*** did you know that you needed to make the change at this point? Additional probing: **How did you know you were not getting the answer? What did you expect to see?**

Please give me specific examples in the recording, of what helped you make this decision. Please point with mouse cursor at these examples on the results page or the full webpage.

- ***Is there anything on this page*** (results page or full webpage) that provided a hint that you needed to change the previous query? If so, please point with mouse cursor at specific examples on the results page or the full webpage, of what helped you make this decision. Additional probing: please point at the feature or the element that helped you realize that you needed to reformulate.
- ***Is there anything on this page*** (results page or full webpage) that helped you decide on what changes needed to be made to the previous query? Additional probing: please point with the mouse cursor at the feature or the element that helped you to reformulate.

17 Appendix 10 – Post-questionnaire used in the actual study

A study on how people interact with web search engine results

Background Questionnaire

Search Engine Questions:

What is your preferred web search engine? _____

Please answer the following questions with regard to your preferred web search engine:

1. How does this search engine know what you're looking for?

_____

_____

_____

_____

2. How does this search engine select and present the list of results?

_____

_____

_____

_____

3. With as much detail as possible, explain how this search engine works. In other words, what does the system "DO" with your search terms?

_____

_____

_____

_____

Domain Knowledge Question:

How would you rate your level of familiarity with the world of **computers** (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____

How would you rate your level of familiarity with the world of **chemistry or chemical engineering** (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____

How would you rate your level of familiarity with the world of **history of sports** (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____

How would you rate your level of familiarity with the world of **material science** (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____

How would you rate your level of familiarity with the domain of **chemical compounds** (from 1 to 5, where 1 is the lowest and 5 being the highest)? _____

General Questions:

Your Gender: ( ) Male ( ) Female

What year were you born in? _____

Are you a ( ) Undergraduate student( ) Graduate student ( ) Other (specify) _____?

What is your home college or school? _____

What is your major? _____

If you're an undergraduate student, are you a:

( ) Freshman ( ) Sophomore ( ) Junior ( ) Senior

If you're a graduate student, what year are you in your program? _____

18 Appendix 11 - Training document for the coders

The purpose of my research is to identify which elements (or as I call them 'cues') in web search results, searchers are using when they make a decision to reformulate (change) their query. In other words, which cues trigger the change from one query to the next.

In order to observe how the searchers reformulate their queries, they were given a list of five questions and asked to use Google's web search engine to find the answers to these questions:

1) How do you find the exact time a hard drive was last formatted on a PC?
2) What minerals can float in salt water?
3) What Australian sport was adopted in Britain?
4) What is the partial pressure of oxygen at an altitude of 5000 feet?
5) When was the Jominy test invented?

These questions are quite difficult, which means that the answers usually didn't appear on the first search result page. The purpose of assigning difficult questions was to make sure that the searchers had at least one query reformulation during their search.

Overall, 33 participants took part in this study. On average, there were 6 reformulations per question.

I captured their computer screen during the search process and marked each query reformulation instance (I was able to observe their entire search process from a different computer and use a special software to mark various points in the recording).

In the next step, I asked the participant questions regarding their decision making process at each query reformulation point (marked during their search), while allowing them to go back and

forth in the recording . I was interested in finding out why they reformulated the query, or more specifically, what exactly they saw in the results preceding the reformulation, which hinted that they needed to reformulate. The participant was asked to point at that element on the results page or the full page. My goal was not to settle for general statements such as 'I didn't see the answer', but get the participants to reveal how they knew that they were not seeing the answer. These interviewing sessions were recorded as well.

In the next stage, I watched and listened to all these recordings in order to come up with an inductive coding scheme for the cues, based on the participants' explanations regarding their reasons to reformulate. Meaning, that I identified recurring themes in the reasons for reformulation mentioned by the participants and made a list of the possible cues. I also added description of the cues, which helps to understand how to identify them, and short cue labels (letters F through X). I then assigned these cues to the participants' explanations of what triggered each query reformulation.

In order to verify my coding, I randomly picked out 25 questions. Your task will be to assign cue labels (letters F through Y) to the explanations provided by the participants regarding their reason for each reformulation. I have prepared a template document with the reformulations for each question, which also includes the time stamp to indicate where in the video you can find the beginning of the dialog about this question. When assigning cue labels, please remember that there can be more than one cue per reformulation. You just need to add another line per each cue that you add.

Please pay attention to which results (after which query) the participant is referring to when talking. The questions were asked about the results that preceded the reformulation, but sometimes the participant preferred to talk about the results they got after the reformulation,

meaning that in this case their explanation addresses the following reformulation, since these are the results that influenced the following reformulation.

19 Appendix 12 - Participant-task pairs used for coder training and verification

The set that the participant-task pair belongs to (bolded are overlapping)	Participant-Task Pair
initial training for both coders	26A
initial training for both coders	13C
initial training for both coders	24C
initial training for both coders	23D
coder 1 advanced training	**3A**
coder 1 advanced training	5A
coder 1 advanced training	1B
coder 1 advanced training	**3B**
coder 1 advanced training	**3C**
coder 2 advanced training	**3A**
coder 2 advanced training	**3B**
coder 2 advanced training	**3C**
coder 2 advanced training	7C
coder 2 advanced training	2E
coder 1 actual set	11A
coder 1 actual set	18A
coder 1 actual set	20A
coder 1 actual set	**23A**
coder 1 actual set	26A

coder 1 actual set	30A
coder 1 actual set	5B
coder 1 actual set	**6B**
coder 1 actual set	**12B**
coder 1 actual set	**18B**
coder 1 actual set	**19B**
coder 1 actual set	24B
coder 1 actual set	27B
coder 1 actual set	29B
coder 1 actual set	17C
coder 1 actual set	**19C**
coder 1 actual set	**20C**
coder 1 actual set	23C
coder 1 actual set	34C
coder 1 actual set	19D
coder 1 actual set	**26D**
coder 1 actual set	**30D**
coder 1 actual set	31D
coder 1 actual set	5E
coder 1 actual set	10E
coder 1 actual set	**12E**
coder 1 actual set	13E
coder 1 actual set	16E
coder 1 actual set	28E
coder 1 actual set	31E

coder 2 actual set	12A
coder 2 actual set	13A
coder 2 actual set	**23A**
coder 2 actual set	**6B**
coder 2 actual set	**12B**
coder 2 actual set	**18B**
coder 2 actual set	**19B**
coder 2 actual set	21B
coder 2 actual set	23B
coder 2 actual set	31B
coder 2 actual set	9C
coder 2 actual set	15C
coder 2 actual set	**19C**
coder 2 actual set	**20C**
coder 2 actual set	21C
coder 2 actual set	27C
coder 2 actual set	2D
coder 2 actual set	10D
coder 2 actual set	18D
coder 2 actual set	20D
coder 2 actual set	22D
coder 2 actual set	24D
coder 2 actual set	25D
coder 2 actual set	**26D**
coder 2 actual set	29D

coder 2 actual set	**30D**
coder 2 actual set	9E
coder 2 actual set	**12E**
coder 2 actual set	21E
coder 2 actual set	26E

20 Appendix 13 - Mixed model results

Reformulation Category = Generalization

At the mean of domain knowledge and search expertise

Formula: generalization ~ G + I + K + Z + I * SEC + K * SEC + G * DKC + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC	BIC	logLik	deviance
898	953.5	-437	874

Random effects:

Groups	Name	Variance	Std.Dev.
questionCode:userID	(Intercept)	5.6861e-14	2.3846e-07
userID	(Intercept)	2.3926e-02	1.5468e-01

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.11249	0.11048	-10.070	< 2e-16	***
G	0.15980	0.22319	0.716	0.47401	
I	1.30610	0.41966	3.112	0.00186	**
K	0.16111	0.20455	0.788	0.43090	
Z	0.98722	0.30400	3.247	0.00116	**
SEC	0.05069	0.06623	0.765	0.44411	
DKC	-0.09707	0.08521	-1.139	0.25459	

I:SEC 0.77678 0.42244 1.839 0.06594 .

K:SEC -0.28049 0.14650 -1.915 0.05554 .

G:DKC 0.33569 0.19935 1.684 0.09221 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Search expertise and domain knowledge centered at the mean minus one standard deviation:

Formula: generalization ~ G + I + K + Z + I * SECminus1 + K * SECminus1 + G *

DKCminus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

898 953.5 -437 874

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 0.000000 0.00000

userID (Intercept) 0.023927 0.15468

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.08017 0.16500 -6.547 5.89e-11 ***

G -0.21635 0.31995 -0.676 0.49891

I 0.13449 0.78727 0.171 0.86436

K 0.58414 0.28750 2.032 0.04217 *

Z 0.98719 0.30400 3.247 0.00116 **

SECminus1 0.05068 0.06623 0.765 0.44414

DKCminus1 -0.09708 0.08521 -1.139 0.25456

I:SECminus1 0.77679 0.42244 1.839 0.06594 .

K:SECminus1 -0.28048 0.14650 -1.915 0.05554 .

G:DKCminus1 0.33579 0.19936 1.684 0.09211 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Search expertise and domain knowledge centered at the mean plus one standard deviation:

Formula: generalization ~ G + I + K + Z + I * SECplus1 + K * SECplus1 + G * DKCplus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

898 953.5 -437 874

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 3.4175e-11 5.8459e-06

userID (Intercept) 2.3927e-02 1.5468e-01

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.14477 0.16731 -6.842 7.81e-12 ***

G 0.53594 0.31144 1.721 0.085280 .

I 2.47772 0.73781 3.358 0.000784 ***

K -0.26195 0.31412 -0.834 0.404324

Z 0.98721 0.30400 3.247 0.001165 **

SECplus1 0.05069 0.06623 0.765 0.444071

DKCplus1 -0.09708 0.08521 -1.139 0.254579

I:SECplus1 0.77679 0.42244 1.839 0.065942 .

K:SECplus1 -0.28049 0.14650 -1.915 0.055533 .

G:DKCplus1 0.33573 0.19935 1.684 0.092166 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

<u>Reformulation Category = Specialization</u>

At the mean of search expertise:

Formula: specialization ~ F + I + K + I * SEC + K * SEC + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

963.6 1005 -472.8 945.6

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 0.031393 0.17718

userID (Intercept) 0.013389 0.11571

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.67680	0.10223	-6.621	3.58e-11	***
F	0.76636	0.19228	3.986	6.73e-05	***
I	-2.81249	1.09489	-2.569	0.010207	*
K	0.05464	0.20419	0.268	0.788995	
SEC	-0.02182	0.05985	-0.365	0.715442	
I:SEC	-1.49358	0.70485	-2.119	0.034089	*
K:SEC	0.52838	0.15872	3.329	0.000872	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Search expertise centered around the mean minus one standard deviation:

Formula: specialization ~ F + I + K + I * SECminus1 + K * SECminus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC	BIC	logLik	deviance
963.6	1005	-472.8	945.6

Random effects:

Groups	Name	Variance	Std.Dev.
questionCode:userID	(Intercept)	0.031395	0.17719

userID (Intercept) 0.013386 0.11570

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.64388 0.13742 -4.686 2.79e-06 ***

F 0.76638 0.19228 3.986 6.73e-05 ***

I -0.55974 1.01556 -0.551 0.581525

K -0.74243 0.33294 -2.230 0.025754 *

SECminus1 -0.02182 0.05985 -0.365 0.715448

I:SECminus1 -1.49354 0.70482 -2.119 0.034088 *

K:SECminus1 0.52842 0.15873 3.329 0.000871 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Search expertise centered at the mean plus one standard deviation:

Formula: specialization ~ F + I + K + I * SECplus1 + K * SECplus1 + (1 |

userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

963.6 1005 -472.8 945.6

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 0.031393 0.17718

userID (Intercept) 0.013388 0.11571

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.70970	0.13533	-5.244	1.57e-07	***
F	0.76636	0.19228	3.986	6.73e-05	***
I	-5.06517	1.90430	-2.660	0.007817	**
K	0.85171	0.29525	2.885	0.003917	**
SECplus1	-0.02182	0.05985	-0.365	0.715427	
I:SECplus1	-1.49356	0.70484	-2.119	0.034089	*
K:SECplus1	0.52842	0.15873	3.329	0.000871	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

<u>Reformulation Category = Substitution</u>

At the mean of search expertise:

Formula: substitution ~ J + J * SEC + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC	BIC	logLik	deviance
660.3	688	-324.1	648.3

Random effects:

Groups	Name	Variance	Std.Dev.

questionCode:userID (Intercept) 0.036099 0.19000

userID (Intercept) 0.030307 0.17409

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.73892 0.11138 -15.612 <2e-16 ***

J -0.33324 0.58490 -0.570 0.5689

SEC -0.06002 0.07311 -0.821 0.4116

J:SEC 0.83824 0.50340 1.665 0.0959 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Search expertise centered at the mean minus one standard deviation:

Formula: substitution ~ J + J * SECminus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

660.3 688 -324.1 648.3

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 0.036099 0.19000

userID (Intercept) 0.030308 0.17409

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.64839 0.15154 -10.878 <2e-16 ***

J -1.59776 1.19030 -1.342 0.1795

SECminus1 -0.06002 0.07311 -0.821 0.4116

J:SECminus1 0.83832 0.50342 1.665 0.0959 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Search expertise centered at the mean plus one standard deviation:

Formula: substitution ~ J + J * SECplus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

660.3 688 -324.1 648.3

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 0.036098 0.19000

userID (Intercept) 0.030308 0.17409

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.82945   0.16175 -11.310   <2e-16 ***

J   0.93130   0.64848   1.436   0.1510

SECplus1   -0.06002   0.07311 -0.821   0.4116

J:SECplus1 0.83825   0.50338   1.665   0.0959 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

<u>Reformulation Category = phrase formation</u>

At the mean of domain knowledge:

Formula: phraseFormation ~ G + F + H + I + J + K + I * DKC + K * DKC +   (1 |

userID/questionCode)

   Data: grouped_bool_matrix

   AIC   BIC logLik deviance

 196.9 252.5 -86.47   172.9

Random effects:

 Groups   Name   Variance Std.Dev.

 questionCode:userID (Intercept) 25.2446 5.0244

 userID   (Intercept) 5.4487 2.3343

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

            Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.2783   1.4426 -5.738 9.56e-09 ***

G -1.6313 0.9636 -1.693 0.09047 .

F -2.6323 1.4342 -1.835 0.06646 .

H -1.0692 0.9419 -1.135 0.25630

I 3.2661 1.1363 2.874 0.00405 **

J -0.3721 1.0236 -0.364 0.71621

K -0.2772 0.9317 -0.298 0.76608

DKC 0.2960 1.2439 0.238 0.81190

I:DKC -1.9067 1.1582 -1.646 0.09973 .

K:DKC -2.3413 0.9635 -2.430 0.01509 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Domain knowledge centered at the mean minus one standard deviation:

Formula: phraseFormation ~ G + F + H + I + J + K + I * DKCminus1 + K * DKCminus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

196.9 252.5 -86.47 172.9

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 25.2443 5.0244

userID (Intercept) 5.4489 2.3343

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8.6100	1.8506	-4.652	3.28e-06	***
G	-1.6313	0.9636	-1.693	0.09047	.
F	-2.6323	1.4342	-1.835	0.06646	.
H	-1.0692	0.9419	-1.135	0.25629	
I	5.4020	2.0508	2.634	0.00844	**
J	-0.3721	1.0236	-0.364	0.71621	
K	2.3453	1.0534	2.226	0.02598	*
DKCminus1	0.2960	1.2439	0.238	0.81189	
I:DKCminus1	-1.9067	1.1583	-1.646	0.09972	.
K:DKCminus1	-2.3413	0.9635	-2.430	0.01510	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Domain knowledge centered at the mean minus one standard deviation:

Formula: phraseFormation ~ G + F + H + I + J + K + I * DKCplus1 + K * DKCplus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC	BIC	logLik	deviance
196.9	252.5	-86.47	172.9

Random effects:

```
 Groups              Name        Variance Std.Dev.
 questionCode:userID (Intercept) 25.2446  5.0244
 userID              (Intercept)  5.4486  2.3342
Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.9467     2.1496  -3.697 0.000218 ***
G             -1.6313     0.9636  -1.693 0.090472 .
F             -2.6323     1.4342  -1.835 0.066458 .
H             -1.0692     0.9419  -1.135 0.256298
I              1.1303     1.3202   0.856 0.391906
J             -0.3721     1.0236  -0.364 0.716212
K             -2.8997     1.7193  -1.687 0.091677 .
DKCplus1       0.2960     1.2439   0.238 0.811890
I:DKCplus1    -1.9067     1.1582  -1.646 0.099723 .
K:DKCplus1    -2.3413     0.9634  -2.430 0.015094 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Reformulation Category = new

At the mean of search expertise:

Formula: new_b ~ F + H + U + W + F * SEC + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

566.7 608.4 -274.3 548.7

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 9.1816e-01 0.95820730

userID (Intercept) 4.0241e-10 0.00002006

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.1336 0.2736 -11.454 < 2e-16 ***

F -0.0230 0.4426 -0.052 0.9586

H 2.0403 0.3127 6.524 6.86e-11 ***

U 2.6185 0.4957 5.283 1.27e-07 ***

W 1.5087 0.3846 3.922 8.77e-05 ***

SEC -0.1425 0.1058 -1.347 0.1781

F:SEC 0.5663 0.2674 2.117 0.0342 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Search expertise centered at the mean minus one standard deviation:

Generalized linear mixed model fit by the Laplace approximation

Formula: new_b ~ F + H + U + W + F * SECminus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

566.7 608.4 -274.3 548.7

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 9.1818e-01 9.5821e-01

userID (Intercept) 1.3730e-12 1.1718e-06

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.9187 0.3042 -9.596 < 2e-16 ***

F -0.8772 0.6518 -1.346 0.1784

H 2.0403 0.3128 6.524 6.86e-11 ***

U 2.6185 0.4957 5.283 1.27e-07 ***

W 1.5087 0.3846 3.922 8.77e-05 ***

SECminus1 -0.1425 0.1058 -1.347 0.1781

F:SECminus1 0.5663 0.2674 2.117 0.0342 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Search expertise centered at the mean plus one standard deviation:

Formula: new_b ~ F + H + U + W + F * SECplus1 + (1 | userID/questionCode)

Data: grouped_bool_matrix

AIC BIC logLik deviance

566.7 608.4 -274.3 548.7

Random effects:

Groups Name Variance Std.Dev.

questionCode:userID (Intercept) 0.91817 0.95821

userID (Intercept) 0.00000 0.00000

Number of obs: 759, groups: questionCode:userID, 142; userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.3485 0.3288 -10.184 < 2e-16 ***

F 0.8312 0.5406 1.538 0.1242

H 2.0403 0.3128 6.524 6.86e-11 ***

U 2.6185 0.4957 5.283 1.27e-07 ***

W 1.5087 0.3846 3.922 8.77e-05 ***

SECplus1 -0.1425 0.1058 -1.347 0.1781

F:SECplus1 0.5663 0.2674 2.117 0.0342 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

<u>Search Effectiveness (response variable = answer)</u>

At the mean of domain knowledge and search expertise:

Formula: answer ~ G + F + H + J + K + U + G * SEC + J * SEC + F * DKC + H * DKC + (1 | userID)

Data: answer_matrix

AIC BIC logLik deviance

171.1 212.5 -71.57 143.1

Random effects:

Groups Name Variance Std.Dev.

userID (Intercept) 0.13784 0.37127

Number of obs: 142, groups: userID, 33

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.56968	0.59042	2.659	0.00785	**
G	-0.22544	0.42610	-0.529	0.59675	
F	-0.23930	0.50236	-0.476	0.63383	
H	-0.07093	0.45746	-0.155	0.87678	
J	0.10621	0.54931	0.193	0.84668	
K	-1.63814	0.47476	-3.450	0.00056	***
U	-1.31539	0.73041	-1.801	0.07172	.
SEC	0.05764	0.20360	0.283	0.77710	
DKC	1.00288	0.42833	2.341	0.01921	*
G:SEC	0.54248	0.31728	1.710	0.08731	.

J:SEC -1.29337 0.49119 -2.633 0.00846 **

F:DKC -1.07666 0.47217 -2.280 0.02259 *

H:DKC 0.79322 0.41566 1.908 0.05635 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Search expertise and domain knowledge centered at the mean minus one standard deviation

Formula: answer ~ G + F + H + J + K + U + G * SECminus1 + J * SECminus1 + F * DKCminus1 + H * DKCminus1 + (1 | userID)

Data: answer_matrix

AIC BIC logLik deviance

171.1 212.5 -71.57 143.1

Random effects:

Groups Name Variance Std.Dev.

userID (Intercept) 0.13783 0.37126

Number of obs: 142, groups: userID, 33

Fixed effects:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.35542 0.71939 0.494 0.62126

G -1.04346 0.61869 -1.687 0.09169 .

F 0.97097 0.61813 1.571 0.11622

H -0.96264 0.60896 -1.581 0.11393

J 2.05647 0.98011 2.098 0.03589 *

K -1.63814 0.47476 -3.450 0.00056 ***

U -1.31540 0.73042 -1.801 0.07172 .

SECminus1 0.05764 0.20360 0.283 0.77712

DKCminus1 1.00289 0.42833 2.341 0.01921 *

G:SECminus1 0.54248 0.31728 1.710 0.08731 .

J:SECminus1 -1.29336 0.49119 -2.633 0.00846 **

F:DKCminus1 -1.07666 0.47217 -2.280 0.02259 *

H:DKCminus1 0.79323 0.41566 1.908 0.05635 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Search expertise and domain knowledge centered at the mean plus one standard deviation

Formula: answer ~ G + F + H + J + K + U + G * SECplus1 + J * SECplus1 + F * DKCplus1 + H * DKCplus1 + (1 | userID)

Data: answer_matrix

AIC BIC logLik deviance

171.1 212.5 -71.57 143.1

Random effects:

Groups Name Variance Std.Dev.

userID (Intercept) 0.13783 0.37126

Number of obs: 142, groups: userID, 33

Fixed effects:

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.78392    0.93017   2.993  0.00276 **
G            0.59257    0.66193   0.895  0.37068
F           -1.44956    0.82830  -1.750  0.08011 .
H            0.82078    0.69596   1.179  0.23826
J           -1.84403    0.86025  -2.144  0.03206 *
K           -1.63813    0.47476  -3.450  0.00056 ***
U           -1.31547    0.73042  -1.801  0.07171 .
SECplus1     0.05764    0.20360   0.283  0.77711
DKCplus1     1.00287    0.42833   2.341  0.01921 *
G:SECplus1   0.54247    0.31728   1.710  0.08731 .
J:SECplus1  -1.29335    0.49119  -2.633  0.00846 **
F:DKCplus1  -1.07665    0.47217  -2.280  0.02260 *
H:DKCplus1   0.79326    0.41566   1.908  0.05634 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

21 References

Ageev, M., Guo, Q., Lagun, D., & Agichtein, E. (2011). Find it if you can: a game for modeling different types of web search success using interaction data. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 345–354). New York, NY, USA: ACM. doi:10.1145/2009916.2009965

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19–26). Seattle, Washington, USA: ACM. doi:10.1145/1148170.1148177

Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 3–10).

Allen, B. (1991). Topic knowledge and online catalog search formation. *Library Quarterly*, *61*(2), 188–213.

Ashkan, A., Clarke, C., Agichtein, E., & Guo, Q. (2009). Classifying and characterizing query intent. *Advances in Information Retrieval*, 578–586.

Aula, A., Khan, R. M., & Guan, Z. (2010). How does search behavior change as search becomes more difficult? In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 35–44).

Aula, A., Majaranta, P., & Räihä, K. J. (2005). Eye-tracking reveals the personal styles for search result evaluation. *Human-Computer Interaction*, 1058–1061.

Aula, A., & Nordhausen, K. (2006). Modeling successful performance in Web searching. *Journal of the American Society for Information Science and Technology*, *57*(12), 1678–1693.

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). New York, NY, USA: ACM. doi:10.1145/2484028.2484049

Barry, C. L. (1998). Document representations and clues to document relevance. *Journal of the American Society for Information Science*, *49*(14), 1293–1303.

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. Retrieved from http://CRAN.R-project.org/package=lme4

Bates, M. J. (1979). Information search tactics. *Journal of the American Society for Information Science*, *30*(4), 205–214.

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

Belkin, N. J. (2000). Helping people find what they don't know. *Communications of the ACM*, *43*(8), 58–61.

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, *38*(2), 61–71.

Bell, D. J., & Ruthven, I. (2004). Searches' assessments of task complexity for web searching. *Advances in Information Retrieval*, 57–71.

Bilal, D. (2001). Children's use of the Yahooligans! Web search engine: II. Cognitive and physical behaviors on research tasks. *Journal of the American Society for Information Science and Technology*, *52*(2), 118–136.

Bilal, D. (2002). Children's use of the yahooligans! web search engine. III. cognitive and physical behaviors on fully self-generated search tasks. *Journal of the American Society for Information Science and Technology*, *53*(13), 1170–1183.

Borgman, C. (1986). The users' mental model of an information-retrieval system – An experiment on a prototype online catalog. *International Journal of Man–Machine Studies*, *24*(1), 47–64.

Borlund, P. (2000a). *Evaluation of interactive information retrieval systems*. Abo Akademi University Press.

Borlund, P. (2000b). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *56*(1), 71–90.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, *53*(3), 225–250.

Borlund, P., & Schneider, J. (2010). Reconsideration of the simulated work task situation: a context instrument for evaluation of information retrieval interaction. In *Proceeding of the third symposium on Information interaction in context* (pp. 155–164). New Brunswick, New Jersey, USA: ACM.

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Computers in Human Behavior*, *21*(3), 487–508.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3–10.

Browne, G. J., Pitts, M. G., & Wetherbe, J. C. (2007). Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly*, *31*(1), 89–104.

Browner, W. S. (2006). *Publishing and Presenting Clinical Research*. Lippincott Williams & Wilkins.

Burke, E. P., Fan, D., Wada, A., Malhotra, J., & Coe, B. (2008, March 13). Search Entry System with Query Log Autocomplete. Sunnyvale, CA. Retrieved from http://www.google.com/patents?hl=en&lr=&vid=USPATAPP11207675&id=_NWoAAAAEBAJ&oi=fnd&dq=google+search+autocomplete&printsec=abstract#v=onepage&q&f=false

Byström, K. (2002). Information and information sources in tasks of varying complexity. *Journal of the American Society for Information Science and Technology*, *53*(7), 581–591.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, *31*(2), 191–213.

Campbell, D. J. (1988). Task complexity: A review and analysis. *Academy of Management Review*, *13*(1), 40–52.

Capra, R., Marchionini, G., Oh, J. S., Stutzman, F., & Zhang, Y. (2007). Effects of structure and interaction style on distinct search tasks. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (p. 451).

Clarke, C., Agichtein, E., Dumais, S., & White, R. (2007). The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 135–142).

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cole, M., Gwizdka, J., Bierig, R., Belkin, N. J., Liu, J., Liu, C., & Zhang, X. (2010). Linking search tasks with low-level eye movement patterns. In *Proceedings of the 2010 European Conference on Cognitive Ergonomics*.

Cutrell, E., & Guan, Z. (2007). What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 407–416). San Jose, California, USA: ACM.

Dimitroff, A. (1992). Mental Models Theory and Search Outcome in a Bibliographic Retrieval System. *Library & Information Science Research*, *14*(2), 141 – 156.

Downey, D., Dumais, S., Liebling, D., & Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. In *Proceeding of the 17th ACM conference on Information and knowledge management* (pp. 449–458).

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Journal of Accounting Research*, *19*(1), 1–31.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. doi:10.1037/0033-295X.87.3.215

Fidel, R. (1985). Moves in online searching. *Online Information Review*, *9*(1), 61–74.

Fidel, R. (1987). What is missing in research about online searching behavior. *Canadian Journal of Information Science*, *12*(3/4), 54–61.

Ford, N., Miller, D., & Moss, N. (2005). Web search strategies and human individual differences: Cognitive and demographic factors, Internet attitudes, and approaches: Research Articles. *J. Am. Soc. Inf. Sci. Technol.*, *56*(7), 741–756.

Freelon, D. (2013). ReCal OIR: Ordinal, Interval, and Ratio Intercoder Reliability as a Web Service. *International Journal of Internet Science*, *8*(1), 10–16.

Google Autocomplete. (2012). *Google | Inside Search*. Retrieved June 9, 2012, from

Granka, L., 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. 25–29).

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 417–420).

Guo, Q., & Agichtein, E. (2008). Exploring mouse movements for inferring query intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 707–708).

Gwizdka, J., & Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–22.

Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's web use skills. *Journal of the American Society for Information Science and Technology*, *53*(14), 1239–1244.

Hargittai, E., Fullerton, L., Menchen-Trevino, E., & Thomas, K. (2010). Trust Online: Young Adults' Evaluation of Web Content. *International Journal of Communication*, *4*, 468–494.

Hargittai, E., & Hsieh, Y. P. (2012). Succinct Survey Measures of Web-Use Skills. *Social Science Computer Review*, *30*(1), 95–107.

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic Web session identification. *Information Processing & Management*, *38*(5), 727–742. doi:10.1016/S0306-4573(01)00060-7

Hembrooke, H., Granka, L., Gay, G., & Liddy, E. (2005). The effects of expertise and feedback on search term selection and subsequent learning: Research Articles. *J. Am. Soc. Inf. Sci. Technol.*, *56*(8), 861–871.

Holman, L. (2011). Millennial students' mental models of search: Implications for academic librarians and database developers. *The Journal of Academic Librarianship*, *37*(1), 19–27.

Hölscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks*, *33*(1-6), 337–346.

Hornof, A. J., & Halverson, T. (2003). Cognitive strategies and eye movements for searching hierarchical computer displays (pp. 249–256). Presented at the ACM CHI'03.

Howard, H. (1982). Measures that discriminate among online searchers with different training and experience. *Online Information Review*, *6*(4), 315–327.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*, *44*(3), 161–174.

Huang, J., & Efthimiadis, E. (2009). Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 77–86). New York, NY, USA: ACM. doi:10.1145/1645953.1645966

Jansen, B. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, *28*(3), 407–432.

Jansen, B., Booth, D., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management*, *44*(3), 1251–1266. doi:10.1016/j.ipm.2007.07.015

Jansen, B., Booth, D., & Spink, A. (2009a). Patterns of Query Reformulation During Web Searching. *Journal of the Americal Society for Information Science and Technology*, *60*(7), 1358–1371.

Jansen, B., Booth, D., & Spink, A. (2009b). Predicting query reformulation during web searching. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems* (pp. 3907–3912). Boston, MA, USA: ACM.

Jansen, B., & 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.

Jansen, B., Spink, A., & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, *56*(6), 559–570.

Jarvenpaa, S. L., Dickson, G. W., & DeSanctis, G. (1985). Methodological issues in experimental IS research: experiences and recommendations. *MIS Quarterly*, *9*(2), 141–156.

Jenkins, C., Corritore, C., & Wiedenbeck, S. (2003). Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. *IT & Society*, *1*(3), 64–89.

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*, *25*(2), 7.

Kelly, D. (2005). Implicit Feedback: using behavior to infer relevance. In A. Spink & C. Cole (Eds.), *New Directions in Cognitive Information Retrieval* (Vol. 19). Springer Netherlands.

Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval*, *3*(1—2), 1–224.

Kelly, D., & Belkin, N. J. (2004). Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 377–384).

Kerr, S. (1990). Wayfinding in an Electronic Database: the Relative Importance of Navigational Cues vs. Mental Models. *Information Processing & Management*, *26*(4), 511–523.

Kim, J., & Can, A. (2012). Characterizing Queries in Different Search Tasks. In *2012 45th Hawaii International Conference on System Science (HICSS)* (pp. 1697–1706). doi:10.1109/HICSS.2012.150

Kim, K. S. (2001). Information-seeking on the Web: Effects of user and task variables. *Library & Information Science Research*, *23*(3), 233–255.

Kim, K. S., & Allen, B. (2002). Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology*, *53*(2), 109–119.

Koenemann, J., & Belkin, N. J. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground* (pp. 205–212).

Krippendorff, K. (2004). Reliability in Content Analysis. *Human Communication Research*, *30*(3), 411–433. doi:10.1111/j.1468-2958.2004.tb00738.x

Lancaster, F. W. (1969). MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation*, *20*(2), 119–142.

Lau, T., & Horvitz, E. (1999). Patterns of search: analyzing and modeling Web query refinement. *Courses and Lectures - International Centre for Mechanical Sciences*, 119–128.

Lazonder, A. W., Biemans, H. J. ., & Wopereis, I. (2000). Differences between novice and experienced users in searching information on the World Wide Web. *Journal of the American Society for Information Science*, *51*(6), 576–581.

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, *44*(6), 1822–1837.

Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N. J. (2010). Analysis and Evaluation of Query Reformulations in Different Task Types. Presented at the ASIST 2010.

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., … Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, *59*(7), 1041–1052.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using Google. *Information Processing & Management*, *42*(4), 1123–1131.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, *40*(1), 54–66.

Meister, D., & Sullivan, D. J. (1967). *Evaluation of user reactions to a prototype on-line information retrieval system* (No. NASA CR-918). Oak Brook, IL: Bunker Ramo Corporation: For sale by the Clearinghouse for Federal Scientific and Technical Information.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, Calif.: Sage Publications.

Norman, D. (1983). Some Observations on Mental Models. In D. Genter & A. L. Stevens (Eds.), *Mental models* (pp. 7–14). Hillsdale, NJ: Lawrence Erlbaum.

O'Day, V., & Jeffries, R. (1993). Orienteering in an information landscape: how information seekers get from here to there. In *CHI '93: Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems* (pp. 438–445). ACM.

Oddy, R. N. (1977). Information Retrieval Through Man-Machine Dialogue. *Journal of Documentation*, *33*(1), 1–14. doi:10.1108/eb026631

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.

Park, T. K. (1993). The Nature of Relevance in Information Retrieval: An Empirical Study. *The Library Quarterly*, *63*(3), 318–351.

Penniman, W. D. (1975). A stochastic process analysis of on-line user behavior. In *Proceedings of the annual meeting of the American Society for Information Science* (Vol. 12, pp. 147–48).

Quiroga, L. M., & Mostafa, J. (2002). An experiment in building profiles in information filtering: the role of context of user relevance feedback. *Information Processing & Management*, *38*(5), 671–694.

Rieh, S. Y., & Xie, I. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, *42*(3), 751–768. doi:10.1016/j.ipm.2005.05.005

Ruthven, I., Lalmas, M., & Van Rijsbergen, K. (2003). Incorporating user search behavior into relevance feedback. *Journal of the American Society for Information Science and Technology*, *54*(6), 529–549.

Sahami, M., & Heilman, T. D. (2011, September 6). Generating Query Suggestions Using Contextual Information. Mountain View, CA. Retrieved from http://www.google.com/patents?hl=en&lr=&vid=USPAT8015199&id=Gl3tAQAAEBAJ&oi=fnd&dq=google+search+query+suggestion&printsec=abstract#v=onepage&q=google%20search%20query%20suggestion&f=false

Saito, H., & Miwa, K. (2001). A cognitive study of information seeking processes in the WWW: the effects of searcher's knowledge and experience. In *Proceedings of the Second International Conference on Web Information Systems Engineering (WISE'01)* (Vol. 1, pp. 321–327).

Saracevic, T. (1969). Comparative effects of titles, abstracts and full texts on relevance judgments. *Proceedings of the American Society for Information Science*, *6*(1), 293–299.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343.

Saracevic, T. (1996a). Interactive models in information retrieval (IR): A review and proposal. In *Proceedings of the 59th annual meeting of the American Society for Information Science* (Vol. 33, pp. 3–9).

Saracevic, T. (1996b). Relevance reconsidered. In *Proceedings of the 2nd Conference on Conceptions of Library and Information Science* (pp. 201–218).

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–1933.

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science*, *39*(3), 197–216.

Saracevic, T., Kantor, P., Chamis, A., & Trivison, D. (1988a). A study of information seeking and retrieving. I. Background and methodology. II. Users, questions and effectiveness. III. Searchers, searches and overlap. *Journal of the American Society for Information Science*, *39*(3), 161–216.

Saracevic, T., Kantor, P., Chamis, A. Y., & Trivison, D. (1988b). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, *39*(3), 161–176.

Sareen, G., Kumar, G., Yu, Y., & Wang, J. (2008, May 8). Presenting Predetermined Search Results with Query Suggestions. Retrieved from

http://www.google.com/patents?hl=en&lr=&vid=USPATAPP11531119&id=1KqqAAAAEBAJ&oi=fnd&dq=google+search+autocomplete&printsec=abstract#v=onepage&q&f=false

Sharit, J., Hernández, M. A., Czaja, S. J., & Pirolli, P. (2008). Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the web. *ACM Transactions on Computer-Human Interaction*, *15*(1), 1–25.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. In *ACM SIGIR Forum* (Vol. 33, pp. 6–12).

Smith, C. (2009). Searcher adaptation: A response to topic difficulty. In *Proceedings of the American Society for Information Science and Technology* (Vol. 45, pp. 1–10).

Smith, C. (2010). *Adaptive search behavior: a response to query failure*. Rutgers University, New Brunswick, New Jersey.

Smith, C., & Kantor, P. (2008). User adaptation: good results from poor systems. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 147–154). Singapore, Singapore: ACM.

Spink, A. (1997). Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science*, *48*(5), 382–394.

Spink, A., & Jansen, B. (2004). *Web Search: Public Searching of the Web* (Softcover reprint of hardcover 1st ed. 2004.). Springer.

Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: selection and effectiveness of search terms. *Journal of the American Society for Information Science*, *48*(8), 741–761.

Spink, A., & Saracevic, T. (1998). Human-computer interaction in information retrieval: Nature and manifestations of feedback. *Interacting with Computers*, *10*(3), 249–267.

Strijbos, J.-W., & Stahl, G. (2007). Methodological issues in developing a multi-dimensional coding procedure for small-group chat communication. *Learning and Instruction*, *17*(4), 394–404. doi:10.1016/j.learninstruc.2007.03.005

Taghavi, M., Patel, A., Schmidt, N., Wills, C., & Tew, Y. (2012). An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces*, *34*(1), 162–170. doi:10.1016/j.csi.2011.07.001

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 415–422). Vienna, Austria: ACM. doi:10.1145/985692.985745

Toms, E. G., O'Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., … MacNutt, A. (2008). Task Effects on Interactive Search: The Query Factor. In N. Fuhr, J. Kamps, M. Lalmas, & A. Trotman (Eds.), *Focused Access to XML Documents* (Vol. 4862, pp. 359–372). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://www.springerlink.com.libezproxy2.syr.edu/content/1020251806153342/

TREC ad-hoc track 2001 topics. (2001). Retrieved July 3, 2011, from http://trec.nist.gov/data/topics_eng/topics.501-550.txt

TREC filtering 2002 topics. (2002). Retrieved July 3, 2011, from http://trec.nist.gov/data/filtering/T11filter_T2002-filt-topics.txt

Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, *39*(3), 445–463.

Wen, L., Ruthven, I., & Borlund, P. (2006). The effects on topic familiarity on online search behaviour and use of relevance criteria. In *Advances in information retrieval: 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006* (pp. 456–459).

White, R., Dumais, S., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 132–141).

White, R., & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 255–262).

White, R., Ruthven, I., & Jose, J. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 35–42).

Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246–258.

Wildemuth, B., & Freund, L. (2009). Search tasks and their role in studies of search behaviors. Presented at the HCIR 2009: Bridging Human-Computer Interaction and Information Retrieval, Washington, DC.

Wildemuth, B., & Freund, L. (2012). Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors. In *Proceedings of the Symposium on Human-Computer Interaction*

and Information Retrieval (pp. 4:1–4:10). New York, NY, USA: ACM. doi:10.1145/2391224.2391228

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. Retrieved from http://arxiv.org/pdf/1308.5499.pdf

Xie, I. (2000). Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. *Journal of the American Society for Information Science*, *51*(9), 841–857.

Xie, I. (2009). Dimensions of tasks: influences on information-seeking and retrieving process. *Journal of Documentation*, *65*(3), 339–366.

22 Vita

NAME OF AUTHOR: Veronica Maidel

PLACE OF BIRTH: Minsk, Belarus

DATE OF BITH: April 28, 1979

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

Tel Aviv University, Tel Aviv, Israel

Ben-Gurion University, Beer-Sheva, Israel

DEGREES AWARDED:

Master of Science in Information Systems Engineering, 2008, Ben-Gurion University, Israel

Bachelor of Science in Industrial Engineering, 2001, Tel-Aviv University, Israel

AWARDS AND HONORS:

The Jeffrey Katzer Doctoral Student Scholarship awarded by the Syracuse University iSchool to a doctoral student with outstanding academic performance.

PROFESSIONAL EXPERIENCE:

Instructor, School of Information Studies, Syracuse University, 2013-2014

Graduate Assistant, School of Information Studies, Syracuse University, 2008-2013

Junior Researcher, Deutsche Telekom Lab, Ben-Gurion University, Israel, 2006-2008

Teaching Assistant, Ben-Gurion University, Israel, 2006-2007

Systems Analyst, Israel Defense Forces, Intelligence Corps, Israel, 2001-2006

Consultant, Optimum Inc., Israel, 2000-2001