

HOW FAR WILL YOU GO? CHARACTERIZING ONLINE SEARCH STOPPING
BEHAVIORS USING INFORMATION SCENT AND NEED FOR COGNITION

Wan-Ching Wu

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in
partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of
Information and Library Science.

Chapel Hill
2014

Approved by:
Jaime Arguello
Nicholas Belkin
Robert Capra
Diane Kelly
Barbara Wildemuth

© 2014
Wan-Ching Wu
ALL RIGHTS RESERVED

ABSTRACT

Wan-Ching Wu: How far will you go? Characterizing online search stopping behaviors using Information Scent and Need for Cognition
(Under the direction of Diane Kelly)

This research sought to explain online searchers' stopping behaviors when interacting with search engine result pages (SERPs) using the theories of Information Scent and Need for Cognition (NFC). Specifically, the problems addressed were how: (1) information scent level, operationalized as the number of relevant documents on the first SERP, (2) information scent pattern, operationalized as the distribution of relevant and non-relevant results on the first SERP, and (3) NFC, a person's tendency to engage in and enjoy effortful cognitive activities measured by the Need for Cognition scale, impacted a person's search stopping behaviors. The two search stopping behaviors that were examined were query stopping, or the point at which a person decides to issue a new query, and task stopping, or the point at which a person decides to end the search task. A laboratory experiment was conducted with 48 participants, who were asked to gather information for six open-ended search tasks.

The results showed significant effects of Information Scent and NFC on search stopping behaviors. When there were more relevant results on the first SERP, participants examined more documents and explored deeper in the search results list; when relevant results were found at the top of the SERP, participants left the SERP after viewing only the first few results. Participants with lower NFC searched deeper but reformulated queries less frequently during a task. Moreover, the time participants with lower NFC spent evaluating search results was more variable depending on the number of relevant results displayed on the first SERP than the time

spent by higher NFC participants. Finally, participants reported that they tended to examine results beyond the first SERP when they conducted people, product, image and literature searches in daily life.

ACKNOWLEDGEMENTS

When I was asked to give a speech after I was announced a doctor my head went blank. I was full of excitement and had not decided what I should feel for being done. Honestly, two weeks after my defense I still dreamed of not passing my defense. Now it has sunk in but my feelings are still beyond words can describe.

My American experience was almost entirely derived from my experience at Chapel Hill, and my Chapel Hill experience was a story about life at Manning Hall. It was difficult at first, I remember pondering every word before I said it and being self-conscious in the presence of other people. Just when I am about to feel comfortable about being myself, the one who is fun, direct and sometimes silly, I am also going to leave UNC. I can't describe how thankful I am to my advisor, Dr. Diane Kelly, the most compatible mentor and colleague I can ever hope for. I look up to her in every respect and for a long time the first thing I woke up in the morning was to check whether I got emails from her. I am not the most confident person in the world, actually far from confident, which prevented me from realizing my strengths in research. Diane knew that I needed to be cued to answer questions which I had rehearsed in my mind a million times but just did not have the courage to share in public, and she never hesitated in giving me opportunities to shine and credits to cheer me up; of course, she never tolerated any slacking off at the same time, either. She understood my half-finished sentences and paper drafts so well that I sometimes joked that her supreme ability to comprehend suboptimal English prevented my English from improving. It will be challenging to find another successful, smart, sharp and

principled woman role model in another environment. Collaborating with Dr. Jaime Arguello and Dr. Robert Capra enriched my research experience and strengthened my mind. Their creative thinking forced me to think out of box, their extraordinary patience cultivated new skill sets in me, and their relentless encouragement helped me become a brave person in the face of challenges. I feel honored to have the opportunity to work on several projects with them and have them serve on my committee. I am also extremely grateful for Dr. Barbara Wildemuth and Dr. Nick Belkin's thought-provoking comments during my oral COMPs, proposal and defense. Their words had motivated me to rethink about my dissertation topic through different lenses and will continue to move me toward extending the current findings.

Despite my tendency to hide myself behind my laptop, friends at SILS manage to get me out to enjoy life like a real person. Thank you for being insistent, patient, supportive, diverse and fun. What you have done for you have way exceeded what I can ever be able to do for you. I am especially thankful for Angela Murillo, who has been a good friend from the beginning to the end, and who has provided unconditional support both in my academic and personal life. Moreover, this dissertation cannot possibly be completed without the help from Kathy Brennan, Rachael Clemens, Annie Chen, Anita Crescenzi, Ashlee Edwards, Oscar Guerra, Heejun Kim, Debbie Maron, Angela Murillo, Gabriele Paetsch, Leslie Thompson, Emily Vardell, Chris Wiesen and Chih-Da Wu, who played the role of document assessors, pilot subjects, statistics consultants and editors. Working with you was an enjoyment and I sincerely hope you felt the same. SILS and the Royster fellowship have generously provided me with financial support, without which I would not be able to sustain my life at UNC and would not have the opportunities to develop research projects with like-minded researchers.

Lastly, there are always two men behind a successful PhD student (Wu, 2014). My father has always stood by me for whatever choices I make in life. He is the silent force who keeps me up and the humble person I want to honor with my achievements. I hope with the attainment of my PhD degree he can finally feel he has fulfilled his responsibility as a father and begin enjoying his retirement life. The second man I would like to thank, my husband, has contributed to both my psychological well-being and the completion of this dissertation, even though I have only known him since SIGIR'12. Most newlyweds' first year of marriage is more or less an extended honeymoon while ours a marathon. Tears, sweat, frustration and long hours of work were symbolic of the past year, but I would not do any differently if I were asked to redo again. With his support and companion, I cannot wait to embark on the next marathon of my life.

TABLE OF CONTENTS

Chapter I. Introduction.....	1
Chapter II. Literature review	8
2.1 Search stopping behavior	8
2.1.1 Task stopping.....	8
2.1.1.1 Evaluation measures	9
2.1.1.2 Notions of enough information.....	11
2.1.1.3 Cognitive stopping rules	17
2.1.2 Query stopping	20
2.1.3 Summary.....	22
2.2 Information scent.....	23
2.2.1 Summary.....	30
2.3 Need for cognition.....	32
2.3.1 Consequences of NFC on information processing	37
2.3.1.1 Type of information sought or considered	37
2.3.1.2 Extent of information processing	40
2.3.2 Summary	45
Chapter III. Research questions	47

Chapter IV. Methods.....	51
4.1 Overview	51
4.2 Manipulation of information scent level.....	54
4.3 Manipulation of information scent pattern.....	55
4.4 Measurement of need for cognition.....	58
4.5 Tasks.....	59
4.5.1 Overview	59
4.5.2 Task development and evaluation from the previous work	60
4.5.3 Task selection	61
4.5.4 Task adaptation.....	64
4.6 Documents.....	66
4.7 Experimental search system	67
4.8 Search behavior measures	69
4.9 Recruitment	71
4.10 Procedure.....	71
4.11 Pilots.....	73
4.12 Data Analysis	73
4.12.1 Quantitative data analysis.....	73
4.12.2 Qualitative data analysis.....	74
Chapter V. Results	76

5.1 Participants	76
5.2 Tasks.....	76
5.3 Manipulation check.....	77
5.4 Overview	78
5.5 Query stopping	83
5.6 Task stopping	90
5.6.1 Effect of task length on task stopping	90
5.6.2 Effect of task length and order of treatments on task stopping	93
5.6.3 Effects of first treatment and NFC on task stopping	101
5.6.4 The last treatment prior to task stopping	107
5.6.5. Relationships among pre-task expectations, search behavior measures and post-task evaluations.	108
5.6.6 Predicting post-task evaluations.	109
5.6.7 Search stopping behavior patterns prior to task stopping.	111
5.7 Query and task stopping strategies.....	116
5.7.1 Query stopping strategies.	117
5.7.1.1 Properties of the search results	117
5.7.1.2 Properties of the queries	122
5.7.1.3 Properties of the search tasks.	123
5.7.1.4 Properties of the person.	125

5.7.1.5 Summary.	127
5.7.2 Task stopping strategies	128
5.7.2.1 Content	128
5.7.2.2 Goal.....	133
5.7.2.3 How they felt.	133
5.7.2.4 Study constraints.	134
5.7.2.5 Summary.	135
5.7.3 Pagination-prone searches	135
5.7.4 Search styles	135
5.8 Summary of results.....	137
Chapter VI. Discussion	141
6.1 Query stopping	141
6.1.1 First impression determined clicking	141
6.1.2 The higher the scent, the greater the number of interactions	142
6.1.3 Distribution of scent mattered, but not as much as ISL	143
6.1.4 Search forward vs. search deeper	144
6.1.5 Moderating effects of NFC on query stopping	145
6.1.6 The non-paginating behavior	148
6.2 Task stopping	149
6.2.1 The enduring effects of first impression	149

6.2.2 Predicting task stopping	152
6.2.3 Task stopping rules: old vs. new	154
Chapter VII. Conclusion and future work.....	157
Appendices.....	165
Appendix A. The first three queries	165
Appendix B. Entry questionnaire	167
Appendix C. Instruction sheet.....	169
Appendix D. Pre-task questionnaire.....	170
Appendix E. Post-task questionnaire	171
Appendix F. Exit questionnaire.....	172
Appendix G. Recruitment email.....	173
Appendix H. Example log file.....	174
Appendix I. Example query-level search behavior data file	175
Appendix J. Example task-level search behavior data file	176
Appendix K. Estimates of parameters in GEE models	177
Appendix L. Task-level search behavior measures for tasks with four or more query submissions	179
References	180

LIST OF FIGURES

Figure

1.	Illustration of information scent at the snippet level and at the SERP level.....	31
2.	Self-reported task difficulty of analyze and evaluate tasks. The labels indicate SD. (H=Health; S&T=Science & Technology; E=Entertainment).....	63
3.	An example search result summary generated by Bing API by submitting “Tatto art”.....	68
4.	A landing page of the experimental search system.....	69
5.	A search engine result page of the experimental search system.....	69
6.	Percentage of tasks where various numbers of queries were observed.....	78
7.	Reformulation, pagination and stopping by treatment (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting)	80
8.	Abandonment by treatment (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting).	81
9.	Interaction effect between ISL and NFC on time.	86
10.	Predicted probability of reformulation (top), pagination (medium) and stopping (bottom) by ISL and NFC.....	88
11.	Predicted probability of reformulation (top), pagination (medium) and stopping (bottom) by ISL and NFC.....	88
12.	Predicted probability of reformulation (top), pagination (medium) and stopping (bottom) by ISL and NFC.....	88
13.	Number of tasks by treatment when task length=1 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting)	95

14. Number of tasks by ISL treatment order (Left) and by ISP treatment order (Right) when task length=2 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting).....	96
15. Number of tasks by ISL treatment order (Left) and by ISP treatment order (Right) when task length=2 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting)	96
16. Number of task by ISL treatment order (Left) and by ISP treatment order (Right) when task length=3 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting)	97
17. Number of task by ISL treatment order (Left) and by ISP treatment order (Right) when task length=3 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting)	97
18. Interaction effect between the first ISL treatment and NFC on DeepestRankClick (Top), NumPred (Middle), and NumNonRele (Bottom) and DeepestRankClick.....	105
19. Interaction effect between the first ISL treatment and NFC on DeepestRankClick (Top), NumPred (Middle), and NumNonRele (Bottom) and DeepestRankClick	105
20. Interaction effect between the first ISL treatment and NFC on DeepestRankClick (Top), NumPred (Middle), and NumNonRele (Bottom) and DeepestRankClick.....	105
21. Interaction effects between first ISP treatment and NFC on NumPagination (Left) and DeepestRankHover (Right)	106
22. Interaction effects between first ISP treatment and NFC on NumPagination (Left) and DeepestRankHover (Right)	106
23. RP frequency by unique participant IDs (Each color represents a unique participant).....	115

LIST OF TABLES

Tables

1. Manipulation of information scent level and information scent pattern.....	51
2. Search result evaluation flow	53
3. All rotations for information scent level treatments.....	53
4. All rotations for information scent pattern treatments.....	54
5. Two examples of documents placement in relation to information scent level treatments.....	55
6. Two examples of documents placement in relation to information scent level treatments.....	55
7. Two examples of documents placement in relation to information scent pattern treatments.....	58
8. Two examples of documents placement in relation to information scent pattern treatments.....	58
9. Interaction signals of analyze and evaluate tasks.....	63
10. Study tasks.....	66
11. Latin square design of task order.....	67
12. Search behaviors at the task level.....	78
13. Search behavior measures (M, SD) by ISL (The highest mean for each measure is bolded to facilitate comparisons).....	82
14. Search behavior measures (M, SD) by ISP (The highest mean for each measure is bolded to facilitate comparisons)	82
15. Results for ISL treatments (Wald X^2 , significance)	84
16. Results for ISP treatments (Wald X^2 , significance)	85
17. Task-level search behavior measures by task length	

(ISL; the highest mean for each measure was bolded for comparison).....	91
18. Task-level search behavior measures by task length (ISP; the highest mean for each measure was bolded for comparison).....	92
19. Number of tasks by task length and order of ISL treatments.....	93
20. Number of tasks by task length and order of ISP treatments.....	95
21. One-query ISL tasks (the highest mean for each measure is bolded)	98
22. One-Query ISP Tasks (the highest mean for each measure is bolded)	99
23. Two-query ISL tasks (the highest mean for each measure is bolded)	100
24. Two-query ISP tasks (the highest mean for each measure is bolded).....	100
25. Three-query ISL tasks (the highest mean for each measure is bolded).....	101
26. Three- query ISP tasks (the highest mean for each measure is bolded).....	102
27. Search behavior measures at the task level by the first treatment in a task.....	103
28. Number of tasks by the last ISL treatment in a task and task length.....	107
29. Significant predictors of post-task evaluations in ISL tasks (“+” indicates positive relationship between a continuous predictor and a criterion; “-“ indicates negative relationship between a continuous predictor and a criterion)	110
30. Significant predictors of post-task evaluations in ISP tasks (“+” indicates positive relationship between a continuous predictor and a criterion; “-“ indicates negative relationship between a continuous predictor and a criterion)	110

31. One-SERP RPs.....	112
32. Two-SERP RPs.....	113
33. Three- or more-SERP RPs.....	114
34. Unique number of participants exhibiting each RP and average number of tasks exhibiting a RP by each unique participant.....	116

Chapter I. Introduction

During information search, a person needs to make a series of decisions. First, the person decides on an initial query to submit. After this, the person makes another decision about whether or not to click on search results returned by the system. Each time a search result is clicked and reviewed, the person further determines whether to continue reviewing other search results in the current returned set, or to refine the query to retrieve a different list of search results. At some point, the person decides to terminate searching.

For Web search tasks that require a single answer, such as navigation and fact-finding, people often stop when they have found an answer to their question. However, for tasks that do not have obvious end-points, it is unclear when and how people decide to stop searching. One thing that complicates this decision is that it is often not easy to assess if the retrieved results represent the best search outcome for a particular information need (Mansourian & Ford, 2007). As a result, over-acquiring information, obtaining excessive information, and under-acquiring information, not obtaining sufficient information, often occur (Connolly & Thorn, 1987). While over-acquiring and under-acquiring information can both be costly, under-acquiring, in particular, may lead to serious consequences for certain tasks; for instance, missing critical information during a health information search can potentially lead to mental and physical disadvantages.

In the context of online searching, searchers might also prematurely stop evaluating results for a query submission. Viewing only a few search results might result in missing important information, obtaining the same documents as everyone else, or biased decision

making. For example, a recent study showed that search engines favored different news outlets and advertisements (Alam & Downey, 2014). It was found that Google had a tendency to rank articles from smaller news outlets that were specialized or politically opinionated relatively higher in the results than Bing, while Bing ranked articles from larger US media and television news outlets higher than Google. The researchers also observed that both search engines ranked content containing their company's own advertisements higher than advertisements from competitors. These findings suggest that if searchers only view highly ranked search results, the information they gain can potentially be biased by the search engine.

Previous work on search stopping behavior has identified stopping rules searchers employ for determining when information is enough or when to declare search a failure (e.g., Nickles, Curley & Benson, 1995, Browne, Pitts & Wetherbe, 2007; Cooper, 1968; Cooper, 1973; Kraft & Lee, 1979) and summarized situational and individual differences influencing when stopping takes place (e.g., Prabha et al., 2007, Zach, 2005; Berryman, 2006; Dostert & Kelly, 2009). However, most of this past research has focused on *information seeking tasks*, in which information gathering involves interactions with systems and human beings, and relied on recall from past search experiences rather than studying *information search tasks*, where actions are restricted to interactions with information search systems as they take place (Saracevic, 2010). Therefore, even though criteria for stopping have been identified, they do not necessarily apply to online search tasks. Moreover, most of this past research has only focused on the search stopping behavior that occurs at the conclusion of a task, rather than the various stopping points that occur throughout the search, such as when a person decides to stop examining results for a query. It is unclear how online searchers decide when to issue a new query from this literature. Lastly, while several studies have highlighted the notion of "feeling good enough" as the main

reason why individuals stop searching (e.g., Zach, 2005), only a few of them have systematically quantified how much information is considered enough (e.g., Dostert & Kelly, 2009). This observation suggests a need to examine the topic of search stopping behavior with a different approach.

This dissertation research conceptualizes search stopping behavior in open-ended search tasks in two types: query stopping, the point at which a person stops interacting with current search results and issues another query, and task stopping, the point at which a person stops gathering information for a task. A theory-driven research framework is proposed to explain these two types of search stopping behaviors. Two theories, information scent (Pirulli & Card, 1999) and need for cognition (Cacioppo & Petty, 1982) are incorporated to guide the arguments and research design of this work.

Information scent is the subjective perception of the value and cost of information sources from proximal cues, such as search result snippets representing the page content (Chi et al., 2001). A central tenet of information scent is that navigational behaviors are guided by the information scent distributed by the immediately available proximal cues, which has been applied by researchers to investigate search result evaluation (e.g., Woodruff, Rosenholtz, Morrison, Faulring, and Pirulli, 2002; Loumakis et al., 2001; Card et al., 2001). For example, Loumakis et al. (2001) investigated how the information scent associated with images on SERPs impacted evaluation behavior. They found that when images were added to text snippets, regardless of image quality, participants were more confident they could find an answer. Kammerer et al. (2009) found that by adding source cues to search result snippets, searchers paid less attention to commercial search results and selected more results from authoritative sources than when source cues were not available. Researchers have also found that information scent

can be used to predict the amount of interaction searchers engage in at a website. Card et al. (2001) observed that if a participant started with a web page with high information scent, he or she would visit more web pages at the site. They also found that as the information scent of web pages declined, there was a tendency for the participant to leave the site or return to a previously visited page. These findings suggest that information scent can possibly be used to explain search stopping behaviors. In this work, it is proposed that the initial search result page can be viewed as a surrogate for the entire set of results retrieved for a query, and thus be used to examine how the characteristics of information scent of the initial SERP influence search stopping behaviors.

Need for Cognition (NFC), a personality trait that measures the extent to which a person enjoys cognitively effortful activities, is also investigated to examine if and how it impacts search stopping behaviors (Cacioppo & Petty, 1982). In general, research has found that high NFC is associated with higher motivation to seek information, increased information processing effort, and an increased ability to assess message quality. For example, students in Curseu's (2011) study who were high in NFC reported a greater tendency to actively seek advice from their teammates when they were asked to solve a complex problem regarding a group assignment than people low in NFC. Bailey (1997) found in a study during which managers were asked to evaluate job candidates, that high NFC managers evaluated candidates' information more thoroughly than low NFC managers. In another study where students were given editorials to evaluate, high NFC students performed better at discriminating between strong and weak arguments than low NFC students (Cacioppo, Petty & Morris, 1983). Given that search is a cognitive activity and that people with higher levels of NFC exert more effort processing

information, NFC may impact how deep a person goes in the search results list before they submit a new query and how many queries a person enters before deciding to stop searching.

The motivation for studying the relationship between Need for Cognition and search stopping behaviors is also informed by recent research about personality-based design in the human-computer interaction (HCI) research, which assumes that people with different personality traits will respond differently to design cues and interact in different ways with systems (Nov et al., 2013). While many studies have investigated the effect of individual differences such as cognitive styles, gender and age on search behavior (e.g., Ford, Miller & Moss, 2001), the relationship between personality and search behavior has received relatively less attention. Moreover, information retrieval (IR) evaluation measures such as rank-biased precision (RBP) (Moffat & Zobel, 2008) and discounted cumulative gain (DCG) (Järvelin & Kekäläinen, 2002) include parameters that can potentially be tuned based on individual user characteristics. However, no research has been done to investigate potential characteristics that might impact a person's persistence with respect to examining a search results list.

In this dissertation, the following research questions are addressed:

1. What is the relationship between the information scent level of the first SERP and search stopping behaviors?
2. What is the relationship between the information scent pattern of the first SERP and search stopping behaviors?
3. What is the relationship between NFC and search stopping behaviors?
4. How can we model task stopping using interaction signals?

The contributions of this dissertation research are summarized below:

This work conceptualizes search stopping behaviors by differentiating query stopping from task stopping in order to study multiple stopping points during iterative open-ended search tasks. The results reveal a variety of factors that affect when searchers reformulate, which have not been discussed in previous work, and factors affecting when searchers stop gathering information, including previously reported factors as well as newly discovered factors. Two distinct search stopping strategies are identified: *searching deeper*, examining search results ranked lower, and *searching forward*, issuing multiple queries. Even though pagination was not common in the study, this research discovers search scenarios that often lead to examining multiple search engine result pages for a single query submission. This research builds on existing knowledge of stopping behavior and contributes to the definition and characterization of search stopping behaviors specifically in the context of search engine result pages.

This work adopts a top-down, theoretical approach to explain search stopping behaviors and applies a controlled laboratory experimental design to examine causal relationships between theories and search behavior measures. The use of search behavior measures allow for quantification of “the feelings of enough” and enable systematic comparisons under different treatments, which is beneficial for translating the findings to practice.

This work extends the applicability of information scent from explaining Web navigation and relevance judgment to explaining search stopping behaviors. The findings of this dissertation show that information scent cannot only be operationalized at the individual search result snippet level, it can also be operationalized at the SERP level and that it influences how searchers interpret the quality of a search. Searchers interact with a search result list to a greater extent if the first SERP shows higher scent. In addition, searchers interact with a search result list to a lesser extent if scent discontinues. Online searchers can potentially be encouraged to search

deeper by manipulating the distribution of relevant results on the first SERP to maintain a continuous information scent.

The study shows how Need for Cognition, a personality trait, affects one's search stopping behaviors. A person's tendency to enjoy cognitive activities is related to the way he or she searches for information and therefore influences when stoppings take place. Unlike previous research which demonstrated a positive relationship between NFC and the extent of information processed, this research shows that NFC manifests itself in different search strategies: searching deeper and searching forward. A person with lower NFC tends to search deeper while a person with higher NFC prefers to search forward. These findings suggest that in the context of online searching, higher NFC is not associated with higher amount of information sought; instead, higher NFC is related to a lower persistence with search result exploration and a higher persistence with query reformulations. These results suggest different designs for search engine result pages according to online searchers' NFC tendencies and different tuning parameters in system evaluation measures based on potential users' NFC.

Chapter II. Literature Review

This chapter reviews previous work related to search stopping behaviors, information scent, and need for cognition. This review first examines studies relevant to search stopping behaviors, followed by studies that use information scent to study search behaviors. Lastly, research about the relationship between Need for Cognition and information processing is presented along with a discussion of its relationship to search stopping behaviors.

2.1 Search Stopping Behaviors

Stopping behavior is generally used to refer to when a person has enough information to complete his or her *task*. Terms such as search stopping behavior, search persistence and search termination have also been used in the literature. However, no formal definition has been given. While many external reasons contribute to the termination of a search task such as interruptions and technology breakdowns, this section primarily focuses on natural stopping points (Zach, 2005), or stopping resulting from the perception that an information need has been met. In this dissertation work, *search stopping behaviors* is used as an umbrella term to cover both task stopping and query stopping. Task stopping is the point at which a person stops gathering information for a task and query stopping is the point at which a person stops interacting with current search results and issues another query.

2.1.1 Task stopping. The discussion of past research about task stopping covers issues regarding when and how people terminate information gathering for both online search tasks and offline information seeking tasks.

2.1.1.1 Evaluation measures. Early task stopping research was closely related to research regarding IR evaluation measures and task stopping was mostly characterized through the quantity of relevant documents and/or non-relevant documents retrieved by searchers. As early as 1968, an IR measure called expected search length was proposed to evaluate system performance (Cooper, 1968). Expected search length was used to indicate the number of non-relevant documents a searcher was willing to look through to obtain a target number of relevant documents; a system with a shorter expected search length was considered more effective than another system with a longer expected search length. In addition, Cooper believed that for any type of search request, specific or exhaustive, there was always a corresponding desired quantity of relevant documents. That is to say, a searcher would grow satisfied with search output when a certain number of relevant results had been found, but this number was contingent on search request type. Similar to expected search length, Blair (1980) proposed the concept of the futility point, the maximum number of documents searchers were willing to examine to find useful documents. Even though neither study intended to investigate task stopping, the two measures provide ways to characterize search stopping points.

Cooper subsequently elaborated on two stopping rules for ranked output, which were the frustration-point stopping rule, where people stop when a certain number of negative-utility documents are encountered, and the satisfaction-point stopping rule, where people stop only when a certain number of positive-utility documents are obtained (Cooper, 1973). Kraft and Lee (1979) extended Cooper's work by modeling the effects of the satiation rule, the disgust rule, and the combination rule, on expected search length. The satiation and disgust rules were similar to the satisfaction and frustration point rules respectively, while the combination rule implies that a person stops searching because of a combination of the satiation and disgust rules. The

researchers demonstrated that expected search length could be approximated using each of the three stopping rules by considering the size of the retrieval set, the number of relevant documents in the set, the number of relevant documents a searcher wished to obtain, and the number of irrelevant documents a searcher would tolerate.

Using a somewhat different perspective, Brookes (1980) distinguished between the accretion of documents and accumulation of information. Brookes observed that relevant documents retrieved later were more likely to be redundant with previously viewed documents and thus the amount of information gained does not grow at the same rate as the number of relevant documents retrieved does, which suggests that it is the perceived amount of information gained rather than the objective number of relevant documents retrieved that affects one's stopping decision. However, all these studies assumed that people stopped searching when they felt they had reached a threshold, which was considered to be either a quantifiable or subjective sense of enough. This sense of enough could come from obtaining enough good information, tolerating too much bad information, or both.

In addition to considering the relationship between the amount of information encountered during the search process and stopping, Kantor (1987) also considered a searcher's resilience to repeated failures. Kantor suggested that a searcher is constantly monitoring the search process; encountering a relevant document increases the searcher's estimated probability of finding another relevant document. Moreover, a person who expects a 50% chance of retrieving a relevant document is more resistant to seeing an irrelevant document than a person who expects a 20% chance of success; the more resistant one is to non-relevant documents, the more likely one continues searching. In other words, the more likely a searcher believes he or she will

succeed in finding satisfactory documents, the more likely he or she is to tolerate non-relevant documents.

Contrary to the dominating view that people terminate searching after reaching a certain threshold, Bates (1984) argued that it was a fallacy to assume the existence of a high-quality n -item set to be retrieved regardless of topic; a searcher believing in this fallacy stops searching immediately once a desired number of results are obtained without considering whether these represent the best results. She maintained that rather than operating under this fallacy, the number of search results needed before stopping depended on task type. She suggested that searchers should distinguish among high recall search, high precision search and brief search tasks, where high recall searches call for everything relevant to a topic, high precision searches retrieve typical but not all documents on a topic, while brief searches aim to retrieve a few documents to approximate search before engaging in a more comprehensive search. That is to say, task type can be a critical factor when predicting when a searcher stops.

2.1.1.2 Notions of enough information. Many researchers have approached the task stopping problem using Herbert Simon's theories of bounded rationality and satisficing (Simon, 1955). Bounded rationality theory argues that because of time and cognitive constraints, it is impossible for human beings to consider all existing outcomes before making a choice. Rather than stopping after exhaustively considering all available information, the theory suggests that human beings will engage in satisficing or "a decision making process through which an individual decides when an alternative approach or solution is sufficient to meet the individuals' desired goals rather than the perfect approach" (Simon, 1971, p. 71). According to March's satisficing model of information seeking, a search is initiated when the information available fails to meet the needs of a person (March, 1994). Search continues until information goals are

met and ends when performance, or effort, exceeds needs. In other words, satisficing search is “thermostatic”; a searcher constantly monitors the search process in order to determine when the information obtained is just good enough (March, 1994).

Agosto (2002) explored young people’s decision making in the context of information seeking on the Web and bounded rationality and satisficing to understand the constraints and strategies she observed. Reduction and termination were two satisficing strategies participants used to address information overload while using the Web for homework purposes. Reduction meant that participants evaluated a subset of available sites on the Web until a satisfying outcome was found. For example, participants returned to known sites, relied on site synopses, and used indexing categories of search engines to filter out non-relevant Websites. They terminated searches when they found an acceptable site, felt physical discomfort, felt bored, faced time limits, or noticed repetition of information.

Mansourian and Ford (2007) analyzed search persistence and failure of academic staff, research staff and students from four Biology-related departments using the framework developed by Agosto (2002). The researchers identified the same satisficing behaviors as those identified by Agosto (2002). In addition, they categorized the behaviors they observed according to search impact and search depth.

Search impact was classified according to two dimensions: the perceived volume of information likely to be missed, and the perceived importance of information likely to be missed. In the *Inconsequential Zone*, the perceived volume of information missing was minimal and missing relevant information did not affect search performance considerably. The *Functional Zone* is more common in Web searching where even though the perceived volume of missing information was high, it did not affect search performance significantly because either alternative

resources were available or the amount of relevant information was good enough to satisfy information needs. In the *Damaging Zone* whether search performance would be adversely affected depended on the abundance of overall relevant information. If the number of relevant documents was high, missing a few did not matter much; yet if the overall number was low, missing a few might lead to search failure. Finally, the *Disastrous Zone* described situations where missing a large portion of highly relevant information led to loss one could not afford.

Search depth was classified according to two dimensions: the searcher's degree of search effort and the searcher's awareness of information likely to be missed. The *Perfunctory Search* described times when a searcher was not aware of missing information and he or she did not attempt to expend much effort searching for information. In the *Minimalist Search* category, a searcher was aware of missing information but he or she did not feel motivated to expend much effort to fill the gap because the easily retrieved information was good enough to meet the searcher's goal. *Nervous Search* happened when a searcher was not sure whether he or she had missed important information and thus was unsure about when to stop. *Extensive Search* took place when one knew that missing information would lead to disastrous consequences and searched thoroughly to avoid missing any information. Searches fell into the last zone mostly when working on literature reviews. While Mansourian and Ford (2007) did not necessarily address stopping rules in information seeking, their work provided insight into how effort and outcomes relate to task stopping.

Prabha et al. (2007) investigated the criteria applied by students and faculty members when they searched information to complete work in an academic setting. Student participants expressed that when they were preparing for presentations or writing reports, they stopped searching for information under the following conditions: (1) when they obtained the number of

citations required by instructors, (2) when they wrote the required number of pages, (3) when they finished answering all questions, or (4) when time was pressing. They also relied on qualitative criteria to determine when to stop searching, including having found accurate information, having obtained sufficient information, having understood the concept well enough to write the assigned report, and retrieving the same information across several sources. Similarly, faculty members referred to criteria such as searching every possible synonym and every combination of query terms, identifying representative or cutting-edge information, finding the same information repeatedly, searching exhaustively from all information sources, addressing colleagues' feedback or journal reviewers' comments, and meeting publishers' requirements to determine when to stop searching information for their research and teaching duties.

A longitudinal study conducted by Warwick, Rimmer, Blandford, Gow and Buchanan (2009) also revealed the application of satisficing strategies in fulfilling course requirements. When students were given the opportunity to choose research questions to work on, they often chose ones with which they were familiar. When they searched for information online, they usually applied familiar tactics such as using terms from assignment descriptions to conduct keyword searches. However, they often gave up if their initial searches did not work.

Historians are another group that have been found to apply satisficing strategies in their work flow. Dalton and Charnigo (2004) found that historians stopped searching when they had enough to write even when other information sources seemed promising to yield additional information. Some historians even tailored their research topics to avoid travel to potentially relevant information sources. When subject searches yielded too many search results, some historians satisficed by only reviewing items from the most respected journals or reviewing only the current work. Duff and Johnson (2002) also identified time and money to be the major

determinants of how much information historians could obtain. Because accessing information at archives was costly, historians made good use of finding aids to help them decide which collections would yield the greatest returns.

Similar criteria observed in academic environments have also been identified in professional environments. Zach (2005) investigated how senior art administrators in fine arts museums and symphony orchestras determined when to stop searching in their day-to-day jobs. It was found that feelings of enough, time constraints and task type, were three main reasons given for terminating information gathering with the feelings of enough being identified the most frequently. While many stopping rules have been identified previously in other studies, no administrator in this study applied any predetermined rules to make stopping decisions. They mostly stopped either because they felt satisfied with the information obtained or when they were forced to stop because of time constraints. Even though task type influenced when art administrators stopped exploring, art administrators essentially stopped when they felt they could complete the task, even if they knew they missed some information.

Berryman (2006) also investigated task stopping in professional environments. The researcher interviewed public sector policy workers about their assessments of enough information during information seeking in the workplace. The work of policy workers was characterized as complex, ambiguous, and “continuous work on persistent issues” (Considine, 1994, p.189), which ranged from the frequent and routine preparation of briefing papers for ministerial meetings to the rare development of legislation. Oftentimes the tasks appeared to be unstructured and unexpected, thus policy workers were constantly engaging in information seeking. One of the themes that emerged from participants’ responses was that they felt there was a lack of framework to determine when information was enough at the beginning of a task,

but once the structure was established, the assessments became easier. Nevertheless, the framework was fluid and changeable because information needs evolved over time. For example, during the drafting process they began to anticipate the reactions of their audience and used the anticipated reactions to determine if more information was needed. Toward the end of the task, they used their colleagues' feedback to determine whether there was potentially missing relevant information. Like many other studies, time constraints also appeared to be an important theme that affected when to stop seeking information. Participants admitted that they often had to cut corners to complete tasks on time. Ultimately, Berryman observed that assessments of enough information in the workplace are subject to the dynamic and complex environments in which people operate. Political climate, the interplay between multiple collaborators, the unfamiliarity of topics and the lack of task structure all increase the difficulty of determining when enough information has been obtained.

While the investigations of task stopping among academics and professionals have added to our understanding of reasons people stop searching at work, all of them were based on self-report data. So far only two studies have modeled or quantified online task stopping. Toms and Freund (2009) studied actions preceding end points in online information search to determine which behaviors could be used to predict stopping. Participants were asked to complete three assigned tasks from a set of 12 tasks developed for the INEX (Initiation for the Evaluation of XML Retrieval) 2007 Interactive Track on the Wikipedia corpus. Toms and Freund identified three major patterns of actions before participants ended their searches. The most prevalent stopping pattern included issuing a query, examining results and then viewing a page; the second most frequent stopping pattern involved a person viewing more results on the second and subsequent SERPs; and the third pattern ended with participants following a link on a page. They

observed that participants tended to revisit and reassess previously visited pages as a means to determine whether one had enough information, which is similar to Cooper's (1973) satiation rule discussed earlier.

Dostert and Kelly (2009) conducted a lab study about task stopping and investigated how accurately searchers were able to estimate recall rates of their own searches. They found that participants stopped searching most frequently based on intuition, which corresponds to the satisfice strategy of "feelings of good enough" reported by Zach (2005). Moreover, participants reported that they also stopped when they noticed a repetition of articles or a decrease in relevant articles. In terms of participants' ability to assess recall rates, Dostert and Kelly showed that when participants stopped they believed they had found 51-60% of relevant information, but in reality they had only identified 7.35%, which shows that searchers overestimate their ability to retrieve available information online. So far, this study has been the first attempt to quantify the feelings of good enough in a controlled laboratory experiment.

2.1.1.3 Cognitive Stopping Rules. While not many studies in information science explicitly address the topic of how much information is enough (Prabha et al., 2007), stopping rules have been investigated extensively in decision sciences and cognitive psychology, mostly in the context of choice tasks (Browne, Pitts and Wetherbe, 2007). This research has found that decision makers rely on cognitive stopping rules, or heuristics to make judgments of information sufficiency.

Cognitive stopping rules useful to information search were first discussed by Nickles, Curley and Benson (1995). They identified four rules decision makers used to terminate information search in two housing sales prediction tasks and two bank interest rate prediction tasks: (1) mental list: stop when a list of requirements are met; (2) magnitude threshold: stop

when the sufficiency of information reaches a predetermined level; (3) difference threshold: stop when not learning anything new; and (4) representational stability: stop when a stable mental model is formed.

Browne and Pitts (2004) and Pitts and Browne (2004) later investigated the use of these rules by systems analysts. System analysts were asked to gather requirements until they had enough information to draw diagrams to design an online grocery shopping system. During the task, analysts engaged in a think-aloud protocol. These data were coded according to the characteristics of the stopping rules identified in Nickles, Curley and Benson (1995). Browne and Pitts found that more experienced analysts tended to use the mental list and magnitude threshold rules, while less experienced analysts applied the difference threshold and representational stability rules. These results suggest that inexperienced problem solvers are more likely to adopt heuristics that had face validity and are easy to apply. The application of different stopping rules resulted in varying degrees of quantity, depth, and quality of information. Mental list and difference threshold rules led to the identification of more system requirements than the magnitude threshold rule and greater depth of requirements than the representational stability and magnitude threshold rules. Moreover, difference threshold rule elicited more quality requirements than the magnitude threshold rule. The analysts stopped after eliciting requirements from 57% of the categories considered important.

In a subsequent study, Browne, Pitts and Wetherbe (2007) explored the relationship between information search task type and use of cognitive stopping rules. A total of five cognitive stopping rules were investigated, including the four rules used in Browne and Pitts (2004) and Pitts and Browne (2004) and an additional rule: single criterion rule (stop searching once a person has gathered enough information about a single predetermined criterion). Results

showed that for well-structured tasks, more participants used the mental list and single criterion rules; while for poorly-structured tasks, more participants terminated search based on the magnitude threshold and representational stability rules. The researchers suggested that when tasks are more complex such as interpreting models, figures, and forms of artistic expressions, or when searchers are new to the tasks, searchers find information until they achieve the gist of the situation (magnitude threshold rule) or when their mental model of a situation is no longer changing (representational stability rule); however, when tasks have low complexity and are easy to decompose into discrete elements, the mental list stopping rule and single criterion rule play the major roles in triggering task stopping.

Two studies replicated a classic consumer choice experiment in the search engine environment to examine how retrieval size affected satisfaction. Oulasvirta, Hukkinen and Schwartz (2009) and Chiravirakul and Payne (2014) investigated how “the paradox of choice”, or how a large number of options leads to poorer choices and less satisfaction with the choice, applied to search results. While the two studies did not set out to investigate either task stopping or query stopping, their findings demonstrated relationships between the size of the returned result set and when participants stopped evaluating results. In Oulasvirta et al. (2009), participants were presented a search task, a search result list of six or twenty-four snippets on paper and were asked to choose the best result from the lists without referring to the landing pages within a 30 second time limit. It was found that the result list of size six led to greater satisfaction with the choice and greater confidence in the correctness. Moreover, when asked to choose an item among a six-snippet list, participants ended up selecting a result ranked higher than when asked to choose among a 24-item list, which shows that participants processed less information prior to making a decision in a smaller set than in a larger set. Chiravirakul and

Payne (2014) replicated the study and had participants conduct real searches on an experimental system with no time limits to increase the ecological validity of the original experiment. They found the opposite effect of retrieval size on satisfaction, albeit not-significant; participants preferred the 24-item list more than the 6-item list. In cases where participants were allowed to reformulate their own queries, participants reformulated more often in the smaller set than in the larger set. In the group where participants were allowed to reformulate, where query stoppings were observed, participants also viewed significantly more snippets before they made their final choice in the smaller set than in the larger set. While the second study improved the study design to align with real online search scenarios, it is still doubtful whether it is appropriate to conceptualize an information search problem as a choice problem. While in a choice task one has to compare and contrast among options in order to derive a final selection, in a search scenario one is often not required to decide on the best result among all documents.

2.1.2 Query stopping. Most previous work on search stopping behavior has focused on *information seeking tasks*, in which information gathering involves interactions with systems and human beings, rather than studying *information search tasks*, where actions are restricted to interactions with information search systems (Saracevic, 2010). Most past research has also constrained stopping behavior to that which occurs at the conclusion of a task, namely, task stopping, rather than investigating the various query stopping points that might occur throughout the search session.

Several common search behaviors extensively studied in the literature are closely related to query stopping and can be used, in part, to understand the nature of query stopping, such as query reformulation, pagination and search depth. For example, query reformulation means a searcher has decided to stop evaluating search results retrieved by his or her current query and

move on to a new query, while the action of clicking on the next button at the bottom of a search result page (referred to as pagination in this study), represents a desire to continue evaluating the results retrieved for a query. Search depth, or the rank of lowest search result a searcher examines, might also be used to indicate a searcher's persistence with respect to a query.

Query reformulation has been a popular search strategy in Web searching. Spink, Jansen, and Ozmultu (2000) analyzed Excite search logs and found that users typically entered 2.84 queries per session, and reformulated their queries in about two-thirds of the sessions. Jansen and Spink (2006) analyzed nine search engine logs from 1997 to 2002 and found that there was an increase in query reformulations. Jansen, Booth and Spink (2009) analyzed search logs collected from 2005 and found that about 40% of query occurrences were reformulations, yet it was unclear how these reformulations were connected to user sessions. Recently, Hassan, Shi, Craswell and Ramsey (2013) used clicks and query reformulations as indicators of search satisfaction. The researchers found query reformulations were a strong indicator of possible task difficulty and task failure. Moreover, queries associated with unsuccessful tasks were often more similar to one another than queries associated with successful tasks. These findings provide useful perspectives for using search behavior to understand query stopping.

Log analyses have also provided descriptive data about the extent to which searchers examine SERPs. Search logs from the Excite search engine showed that people typically examined 1.7 pages per query and for about 50% of the queries, searchers went to the next page before reformulating their queries (Spink et al., 2000). The researchers also found that search depth decreased over time; the percentage of queries leading to pagination decreased from 71% to 27% in US-based search engines from 1997 to 2002 (Jansen & Spink, 2006).

Recent studies were able to demonstrate search depth at a finer level of granularity by collecting mouse hover and eye tracking data. Cutrell and Guan's eye tracking study reported that people examined the first eight results before they re-issued another query (Cutrell & Guan, 2007). Lorigo et al. (2008) examined participants' scan paths as they carried out search tasks. They found that on average participants scanned only 3.2 distinct search results for each query with some participants revisiting the same results multiple times. Using cursor movements, Huang, White, and Dumais (2011) observed that people re-queried after inspecting only the top four results. Another experimental study motivated by Search Economy Theory found that search depth was affected by query cost (Azzopardi, Kelly, & Brennan, 2013). Participants who used an interface that required more time to enter a query, entered significantly fewer queries and went to greater depths in the search results list than participants who used a standard search interface. These results suggest that certain aspects of the search interface can impact search behavior and also provide a theoretical explanation for this behavior. Overall, this research regarding query reformulation, pagination and search depth indicates that online searchers rarely evaluate results beyond the first SERP when using standard search interfaces.

2.1.3 Summary. Early research on IR measures attempted to estimate the number of documents a searcher needed with regard to a query submission. Some work developed stopping rules and simulated stopping points mathematically to address the issues. However, since the studies equate query stopping for a single query submission with task stopping, the rules found in this literature are more useful for understanding task stopping than query stopping.

Studies from the past decade have seen progress in accumulating criteria for task stopping in both online and offline search scenarios, extending the scope of task stopping to the discussion of its associated effort and consequences. A few studies showed success in obtaining

empirical search data from laboratory experimental studies, marking the initial accomplishments in systematically predicting and identifying task stopping. Among studies where cognitive stopping rules were investigated, only three studies attempted to discover stopping rules for search tasks and validate these rules in online search tasks (Brown & Pitts, 2004; Pitts & Brown 2004; Browne, Pitts, & Wetherbe, 2007). These studies provided useful criteria for task stopping. Yet research is still needed to investigate how these cognitive stopping rules are applied while interacting with search engines.

Laboratory and log-based studies shed light into search depth, pagination and query reformulation, which described the evolution and status quo of query stopping. While these studies did not set out to investigate query stopping, they revealed insights into the profiles of modern search behaviors and indicated that most online searchers only viewed results on the first SERP. However, these studies did not explain why most searchers did not examine results beyond the first ten results.

This dissertation work sought to address the gap in our understanding of search stopping behaviors by distinguishing between query stopping and task stopping, by identifying the use of existing stopping rules and exploring other stopping rules in the SERP environment, and by explaining search stopping behaviors with theories of information scent and need for cognition.

2.2 Information Scent

What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate the attention efficiently among the overabundance of information resources that enrich the information people process.

(By H. A. Simon; as quoted in Pirolli & Card, 1999, p. 643)

Determining where to allocate attention during information seeking is a tough decision, especially in a world where information is made readily available by the Internet. To address the constant information overload experienced in modern life, understanding the interactions between information seekers and information systems has great value in facilitating effective information search. During the past two decades, Information Foraging Theory (Pirolli & Card, 1999) has been used to explain how information seekers select, navigate and switch between information sources. The core of the theory is made by an analogy between information seeking behavior and animal foraging behavior: while animals use environmental cues to identify the most fruitful places to forage, human beings are guided by *information scent*, or “the imperfect, subjective perception of the value and cost of information sources from proximal cues” (Chi et al., 2001, p. 491). Proximal cues are “imperfect information at intermediate locations” (Pirolli & Card, 1999, p. 646) “whose trail leads to information of interest” (Pirolli, 1997, p.3). Search result snippets, anchor text, Uniform Resource Locators (URLs) and thumbnails are some examples of proximal cues from which information scent can be obtained.

Optimal Foraging Theory, the origin of information foraging theory, explains animals’ adaptation behaviors to environmental constraints when foraging for food. When a forager seeks prey, it encounters different types of habitat that promise different amounts of net energy. The costs of obtaining food differ depending on difficulties associated with access or navigation. The ultimate goal of a forager is therefore to find the best solution to maximize the net energy by taking into consideration the associated costs. Since oftentimes food is located in patches such as in berry bushes, a forager spends some time foraging within a bush and when the point is reached in which the amount of food available is below a certain threshold, the forager leaves the current bush to seek other bushes.

An analogous scenario in Information Foraging Theory to the food-seeking scenario in Optimal Foraging Theory may be observed when a searcher looks for information in a specific subject domain. Information relevant to the searcher's interest area can be found in file drawers, libraries, online bibliographic databases and in other online sources. The searcher needs to navigate between information patches, or information sources, to find information; for example, he or she may start by browsing personal bookshelves, searching online databases, consulting a physical library. Just like an animal forages for food, an information seeker forages for information; therefore an information seeker is called an *informavore* in Information Foraging Theory. But how does an *informavore* decide whether to stay within a single patch or move between patches? How does he or she determine when to switch to another information resource?

Informavores make foraging decisions based on comparing the profitability associated with each information patch to another. In order to determine the profitability of each information patch, *informavores* base their predictions on "information scent" before taking navigation actions. Such scent-based assessments are also carried out constantly during the search process in order for *informavores* to decide whether to stay in one patch or switch to another. For example, when a searcher issues a query to a search engine, he or she is presented with a list of search results that are predicted to be relevant to what the searcher is looking for. Each search result is represented by three information scent cues: its title, a URL, and a summary snippet. As the searcher continues examining the returned set, he or she makes a cost-benefit analysis about whether or not more effort should be expended evaluating the current list. If the quality of search results turns out promising, the searcher is likely to stick to reviewing more results; on the contrary, if the effort expended viewing search results is greater than the

information gained, the searcher may switch to another patch, by, for example, reformulating his or her query.

In order to understand how information scent guides Web surfing, Pirolli and his colleagues first conducted a survey to find out what activities people engage in on the Web (Morrison, Pirolli, & Card, 2001) by inserting a survey question into the Gvu Tenth WWW Searcher Survey (Kehoe, Pitkow, Sutton, & Agrawal, 1999). They used the Critical Incident Technique and presented participants with the following statement: “Please try to recall a recent instance in which you found important information on the World Wide Web; information that led to a significant action or decision. Please describe that incident in enough detail so we can visualize the situation.” The analysis of the collected responses showed that a major purpose of using the Web was to sense-make through finding specific pieces of information. Based on the Web activity descriptions derived from the study, Card et al. (2001) developed two representative search tasks - City and Antz - to study Web navigation in the laboratory environment. In the City task participants were asked to imagine themselves being the Chair of Comedic events for a university; they were asked to figure out a date of an event that was going to take place on campus and look for a photograph of the event to put on the advertisement. In the other task, participants were asked to find a Website where one could purchase the set of four “Antz” movie posters depicting the princess, the hero, the best friend, and the general. During the study, participants were instructed to think aloud while they searched, and the audio recordings were later transcribed and drawn into “Web Behavior Graphs” (WBG), which allowed researchers to easily visualize the navigation paths. Each page visited by participants was later ranked by three independent judges according to potential utility (information scent): 0=no scent, 1=low scent, 2=medium scent and 3=high scent.

Card et al. (2001) found that participants spent more time searching and exploring on the Antz than on the City task. Moreover, the average information scent of pages visited for the Antz task was lower than that for the City task, suggesting that under conditions of strong information scent, searchers moved directly to the target information.

Card et al. (2001) also found a higher frequency of within-site than between-site transitions for both search tasks. From the WBGs they mapped, it was observed that as the information scent of Web pages encountered at a site declined, there was a tendency for participants to leave the site or return to a previously visited page. More importantly, they identified a relationship between the information scent first encountered on a site and the number of pages visited at the site; if a person started with a high information scent Web page, he or she visited more Web pages at the site.

Using the theory of information scent, Chi and his colleagues (2000, 2001 & 2003) developed an algorithm and infrastructure to predict online surfers' destinations. The resulting infrastructure - Bloodhound Simulator - allowed researchers and practitioners to test alternative interface designs. Bloodhound Simulator was evaluated using four sites - help.yahoo.com (the help system of Yahoo!), www.rei.com (an outdoor online store), hivinsite.ucsf.edu (AIDS and HIV information site), and parcweb.parc.com (the intranet of a company). Chi et al. (2003) invited a diverse group of 244 subjects to conduct eight search tasks. The simulator predicted how many times a searcher should have visited each page and compared the value to observed frequencies from actual searcher data. It was shown that Bloodhound's predictions strongly correlated with real searcher data in one third of the cases and moderately correlated with searcher data in roughly another two thirds of the cases. While very little theory in Human-Computer Interaction has been able to guide usability studies, these results demonstrated the

applicability of the theory of information scent to predict navigation behavior and stopping in the context of usability evaluations.

Information scent has also been used to understand relevance assessment behavior. Loumakis, Stumpf, and Grayson (2011) investigated the relationship between the information scent of images on SERPs and people's relevance assessments. Three types of search snippets were compared: text-only, image-only, and text + image snippets. Overall, text + image snippets led to the highest information scent. The results showed that images had their own distinct scent. When images were added to text snippets, participants were more confident they could find an answer regardless of image quality, which suggested that images contributed positively to the overall information scent of the SERP. However, adding images to text snippets did not actually enhance search effectiveness and search efficiency.

The same conclusion was made in another study where the snippets were manipulated in a similar way. Woodruff, Rosenholtz, Morrison, Faulring, and Pirolli (2002) compared navigation behavior when participants used three snippet types - enhanced thumbnails (a snippet which combined both plain thumbnails and text summaries), plain thumbnails alone, and text summaries alone - to search for different types of information. They found that for some categories of tasks, thumbnails outperformed text summaries; in some other categories of tasks, text summaries outperformed thumbnails; but enhanced thumbnails outperformed both on all measures. They argued that for some tasks, text summaries provided lower scent while for others, text summaries were perceived as embodying higher scent than thumbnails; however, in all tasks the enhanced thumbnail view was consistently perceived as scent-enriching, suggesting that offering both text summaries and thumbnails have the most benefits for searchers regardless of task differences.

In addition to images, Kammerer et al. (2009) found that by adding source cues to search result snippets, such as the label “Science/Institutions” to signify the authoritativeness of the content, searchers paid less attention to commercial search results and selected more results from authoritative sources than when source cues were not available. However, participants were provided 30 results to “select” from rather than to “search;” therefore, the effect of information scent on search behaviors such as reformulation and pagination remains unclear.

Information scent appears to affect the perceived credibility of news articles as well. Sundar, Knobloch-Westerwick, and Hastall (2007) investigated the impact of three types of information scent cues – source credibility, recency and number-of-related-articles – on perceived message credibility. Each subject was presented with 12 news items differing in high and low source credibility, six levels of number of related articles, and three levels of recency and asked to evaluate the news site after reviewing all items. The effects of these cues on perceived credibility were not straightforward: when a news article was attributed to a high credibility source, how recently the article was published or how many related articles there were did not matter; however, if a news article was distributed by a low credibility source, it received the highest message credibility ratings when the other two cues were at their highest.

Katz and Byrne (2003) applied information scent to explain a searcher’s choice between a search function and a menu browsing function while locating a product on a Website. It was hypothesized that when menus had poor information scent, operationalized by poor labeling, searchers would prefer to use a search function over browsing. Subjects were instructed to decide how to locate specific products on sixteen customized commercial sites that differed in information scent (high and low) and their first actions for retrieving the products were recorded. It was found that higher information scent led to less searching. Moreover, the observation that

subjects browsed higher scent menus even in the presence of a search function suggests that providing a search function should not preclude careful labeling of product categories. Their results may also imply that in the context of search engine usage, a SERP with high information scent can potentially attract more interactions with results (similar to Website browsing) while a SERP with low information scent causes searchers to reformulate (similar to using the search function on a Website).

Information scent has also been applied to model navigation paths on specific software applications. Lawrence, Bellamy, Burnett, and Rector (2008) presented a model to predict how programmers navigated code for software maintenance tasks. Programmers were predicted to visit the source code revealing the highest scent. Their model turned out to predict debugging behavior close to expert programmers' real debugging behavior, and most importantly, the scent-modeled navigation paths were indistinguishable from historical navigation paths. The researchers suggested that scent-based indicators should be added to existing software tools to enhance the usage of navigation history to discover source code relationships.

2.2.1 Summary. Past research investigating the effect of information scent on search result evaluation and navigation behavior mostly manipulated information scent at the snippet level. Findings of Card et al. (2001) are especially useful for formulating the argument of using information scent to predict online search stopping behavior. First, information scent in Card et al. (2001) was measured by how likely a Webpage would lead to useful information, which is similar to the granularity of a SERP. Secondly, the results demonstrated that the information scent of the initial Web page encountered was related to the level of search persistence within a site, which signals its usefulness in predicting query stopping within a returned search result set. Considering the idea that search results returned by a search query to a search engine can be

regarded as a type of information patch (Pirolli, 2007, p. 52), it is reasonable to hypothesize that the initial search results one encounters on the first SERP can be used to model query stopping. Information scent can perhaps be treated as an attribute of the search results in one's immediate virtual environment, the first SERP. Moreover, by the same analogy that an information scent of a single proximal cue provides estimation of the quality of a single distal information object as shown in the relevance behavior studies, information scent obtained from collective proximal cues may suggest the quality of collective distal information objects as well (Figure 1).

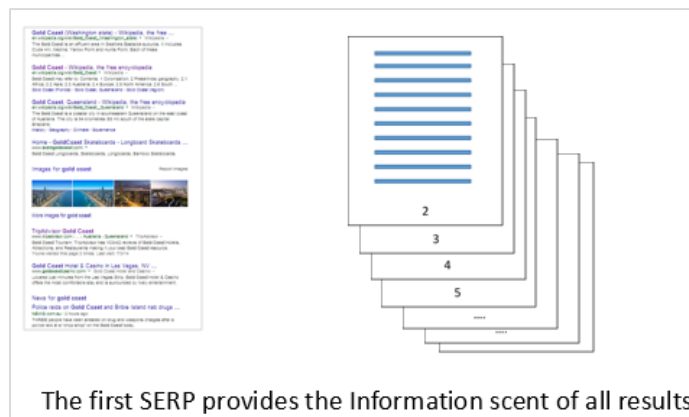
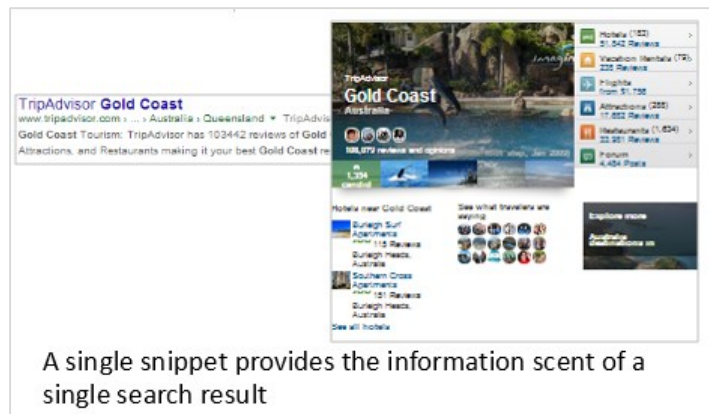


Figure 1. Illustration of information scent at the snippet level and at the SERP level

2.3 Need for Cognition

There are a number of theories in the psychological research about how human beings process information. While most research has focused on situational factors that determine when people engage in effortful information processing and when they think heuristically (Kahneman, Slovic, & Tversky, 1982), a common source of variance has been attributed to individual differences. Many studies in information science have investigated the effect of individual differences such as demographic variables on search behavior (e.g., Ford, Miller, & Moss, 2001; Morahan-Martin, 1998; Borgman, 1989; Borgman, Hirsh, & Walter, 1995). Others discussed the differences in search strategies between experts and novices (Rieh, 2002). Still others examined the effect of cognitive styles (e.g. Field dependent vs. field independent, Palmquist & Kim, 2000; imager vs. verbalizer, Ford, Miller, & Moss, 2001; wholist vs. analytic, Wang, Hawk, & Tenopir, 2000) on search behavior and search result evaluation styles (Economical vs. exhaustive, Aula, Majaranta & Rähä, 2005; breadth-first vs. depth-first, Klöckner et al., 2004). However, empirical work investigating the effect of personality differences on search behavior is uncommon.

Personality, a “relatively stable emotional, motivational, interpersonal, and attitudinal characteristics of the individuals distinguished from abilities” (Pocius, 1991), has been studied extensively in behavioral sciences and social sciences but has not yet been discussed in depth in the search behavior literature. When it comes to the issue of how stable personality traits influence behavior, the psychology community is divided into three camps. The personality approach scholars believe personality traits are the prime predictor of behavior; the situational approach emphasizes the characteristics of the situation in which behavior takes place as the

main predictor of behavior; and the interactionist approach emphasizes the joint contribution of personality and situation in determining behavior.

In HCI and IR research, personality has been treated as a moderator of the relationship between the situational factors and the behavioral measures of interest in several studies, reflecting the interactionist approach (e.g. Nov et al., 2013; Jozsa et al., 2012). According to Mischel and Shoda's (1995) theory, this means personality accounts for behavior in predictive patterns across situations (e.g., in situation A one does X, in situation B one does Y). For example, Nov et al. (2013) found that users' extroversion level moderates the relationship between the user interface of audience size and user contribution, and users' emotional stability moderates the relationship between the effectiveness of social anchors and user contribution. Jozsa et al. (2012) studied the interaction between Myers-Briggs personality preference and language (native vs. foreign) on search behavior. They found the Thinking-Feeling dimension was significantly associated with the success rate on the foreign language task but not the native language task; Feelers tended to achieve better results than Thinkers by spending more time and effort to find suitable Webpages.

There are two IR measures, Discounted Cumulative Gain (DCG) (Järvelin & Kekäläinen, 2002) and Rank-Biased Precision (RBP) (Moffat & Zobel, 2008) that have parameters that can be tuned based on characteristics of the searcher, although no research has been conducted to determine what types of characteristics might be represented by this parameter or what appropriate values might be for this parameter. For example, with DCG it is believed that searchers examine search results linearly from top to bottom and the lower the ranked position of a relevant document, the less likely it will be assessed by a searcher. As such, documents ranked lower are associated with less value (i.e., their value is discounted) than documents ranked

higher. DCG has a parameter to represent characteristics of the searcher, which Järvelin and Kekäläinen (2002) call searcher persistence. A small value of the persistence parameter ($b=2$) is used to represent an impatient searcher for whom the lower ranks of documents are not likely to be assessed; and the persistent parameter discounts late documents to a greater extent than earlier documents. Whereas for patient searchers who are more likely to access lower ranks of documents, a large value, $b>100$, is applied so that later documents are weighted almost equally to earlier relevant documents. While formal guidelines do not exist about the choice of b values, the authors suggested 2 as the value for b and many experiments have applied $b=2$ by default. Nevertheless, the authors admitted that the choice of the base is arbitrary. They believed “either the evaluation scenario should advise the evaluator of the base or a range of bases could be tried” (Järvelin & Kekäläinen, 2002, pp. 439-440).

Similarly, in another measure, RBP, different p values represent different levels of searcher persistence, and thus different ways in which ranked lists can be used. Values closer to 1 indicate highly persistent searchers who examine many documents before terminating searching. Values smaller than 0.5 represent less persistent searchers who only evaluate the first few documents on the SERP. Searchers with $p=0.0$ implies that the searchers will only examine the first document and not look further. The choice of p can be determined based on searcher or task characteristics (Moffat & Zobel, 2008). If assumptions about searcher persistence can be made on the targeted searcher population, researchers have a better idea of the range in which p should be chosen. For general-purpose search systems, a range of p values can be used to report RBP results. For tasks that require high recall, a relatively high p should be used; for most Web search tasks, a smaller p is appropriate. So far, only one study has examined the distribution of the search persistence parameter by analyzing logs from a commercial search engine (Park &

Zhang, 2007), but no controlled laboratory user studies have been conducted to investigate the relationship between real search stopping behavior and these persistence parameters.

A literature search in personality psychology suggests that *Need for Cognition (NFC)*, a personality trait that describes a person's tendency to enjoy and engage in the process of information processing, bears a close relationship with search stopping behaviors and can possibly be used to inform the persistence parameters. Awareness that individuals differ in their motivation to engage in effortful thinking can be traced back to early history in personality and social psychology, when the first empirical studies about Need for Cognition were conducted by Cohen, Stotland, and Wolfe (1955) and Cohen (1957). Cohen et al. (1955) described Need for Cognition (NFC) as “a need to understand and make reasonable the experiential world” (p. 291). Cohen and his colleagues (1955, 1957) conducted two empirical studies in which it was found that individuals who had high NFC preferred structured information to ambiguous information (Cohen et al., 1955), and they were generally more likely to organize, elaborate, and evaluate information they were exposed to than individuals who had low NFC (Cohen, 1957). Yet during this time there were no instruments to produce NFC scores; participants were classified into high or low NFC only by their reactions to a hypothetical situation and ordering of statements about five needs by importance.

Cohen and his colleagues (1955) argued that “stronger needs lead people to see a situation as ambiguous even if it is relatively structured, indicating that higher standards for cognitive clarity are associated with greater need for cognition” (p. 292). Their characterization of NFC emphasized ambiguity intolerance and tension reduction, which seemed to be similar to need for closure (Webster & Kruglanski, 1994), tolerance of ambiguity (Shaffer & Hendrick, 1974), and need for structure (Neuberg & Newsom, 1993). Cacioppo and Petty (1982)

conceptualized NFC somewhat differently. They proposed NFC as “a stable individual difference in people’s tendency to engage in and enjoy effortful cognitive activity” (Cacioppo, Petty, Feinstein, & Jarvis, 1996, p. 198).

Cacioppo and Petty (1982) characterized individuals with low NFC as chronic cognitive misers, and individuals with high NFC as chronic organizers. Both chronic cognitive misers and chronic organizers have a need to make sense of the world, but they exhibit variations in terms of attitudes and behaviors. Chronic organizers tend to actively seek, think about, and reflect on information to make sense of the stimuli and relationships among stimuli in their world while chronic cognitive misers tend to rely on others’ opinions (e.g., experts and celebrities) and heuristics to arrive at mental models of the world. Moreover, compared to chronic cognitive misers, chronic organizers often exhibit more positive attitudes toward tasks or stimuli that require effortful thinking (e.g., tests and reading) than non-intellectual stimuli (e.g., sports and pets). In terms of problem solving strategies, chronic organizers are more likely to apply technologies that involve effortful thinking. Because of the greater extent of effortful information processing, chronic organizers tend to accumulate a wider array of topical knowledge than chronic cognitive misers.

Although individuals with different levels of NFC are prone to engage in and enjoy effortful thinking to different degrees, the literature confirms that the relationship can be moderated by situational factors such as personal relevance of an event (Cacioppo et al., 1996). For events with high personal relevance, nearly everyone will exert greater cognitive effort (Axsom, Yates, & Chaiken, 1987). In such cases, differences between low and high NFC individuals should not be as evident. Differences in cognitive effort are most evident when situational demand is neither very high nor very low.

2.3.1 Consequences of NFC on information processing. Considerable research about NFC has prevailed in fields of social psychology, personality psychology, behavioral medicine, education, journalism, marketing and law (Cacioppo et al., 1996). Much of the literature was developed based on the hypothesis that individuals who differ in terms of NFC also differ in terms of their tendency to seek detailed information (Cacioppo et al., 1996). Studies found that individuals with high NFC showed higher motivation to process information, recalled more information, paid more attention to argument quality than peripheral cues, generated more task-related and thoughtful responses, possessed more knowledge, performed better on cognitive tasks, and reacted more positively to complex rules (Cacioppo et al., 1996). Studies investigating how NFC affects information seeking, processing and evaluation in both online and offline contexts, especially with an emphasis on the type and amount of information sought or considered are discussed in the following sections.

2.3.1.1 Type of information sought or considered. As part of the Elaboration Likelihood Model (ELM), Petty and Cacioppo (1981, 1986) proposed that persuasion can be characterized by way of two routes: central and peripheral. The former involves careful scrutiny of argument content, while the latter influences decision making process through external and irrelevant cues or mental shortcuts. When individuals have both the ability and motivation to evaluate message content thoroughly, they are more likely to follow the central route. In contrast, when individuals lack sufficient knowledge or motivation to scrutinize messages carefully, they tend to take the peripheral route. The authors believe that a high NFC person is prone to engage in deeper thinking, and are thus more likely to pay more attention to message arguments, which is part of the central route of information processing. On the other hand, a low NFC person avoids

extensive thinking, which makes adopting heuristics appealing to them; therefore, the peripheral route of information processing in ELM is more likely to be favored by low NFC individuals. To test their predictions, Cacioppo, Petty, & Morris (1983) examined the effect of argument quality on message evaluations and source impressions between individuals high and low in NFC. Participants were told to evaluate editorials written by journalism students. During the experiment participants were presented with either an editorial with a strong or a weak argument. Results showed that participants with high NFC were better at discriminating strong from weak arguments than participants with low NFC, which suggested message quality had a greater impact on individuals with high NFC. Self-report data also showed that participants scoring high in NFC expended more effort in thinking and also recalled more arguments in the editorials than participants scoring low in NFC, indicating that individuals high and low in NFC processed messages to different extents.

In contrast to Cacioppo et al. (1983), Chaiken, Axsom, Hicks, Yates, and Wilson (cited in Chaiken, 1987) investigated how peripheral information cues rather than argument quality affected message effectiveness when NFC varied. Two audio tapes which contained the same speech were played to participants except that in one of them the speaker started by stating he would discuss two points while in the other version the same speaker stated that ten points would be discussed. Low NFC participants agreed more when ten instead of two arguments were claimed to be presented, whereas no significant difference was found among high NFC participants. The finding suggested that high NFC participants tended to pay more attention to the arguments, while low NFC participants relied on peripheral cues such as number of arguments to evaluate persuasive messages, which corresponded to the findings in Cacioppo et al. (1983).

Haugtvedt, Petty, and Cacioppo (1992) conducted a study similar to Cacioppo et al. (1983) and Chaiken et al. (1987) in an advertisement context. In one experiment, participants with high and low NFC were presented with strong and weak advertisements for the same product. While both groups expressed more favorable attitudes towards the product when the arguments were strong rather than weak, low NFC participants performed worse than high NFC participants in discriminating strong from weak advertisements. In another experiment, advertisement quality was held constant while advertisement attractiveness was manipulated by assigning an attractive endorsement to one condition and an unattractive endorsement to another condition. Consistent with Chaiken et al. (1987), the findings again showed that peripheral cues such as endorsement attractiveness affected low NFC participants more than high NFC participants.

See, Petty, & Evans (2009) built on the hypothesis that individuals with low NFC tend to be affected more by peripheral cues, namely, cues requiring minimal processing effort than individuals with high NFC, and tested whether messages corresponding to one's NFC level motivated individuals to process messages to a greater extent than messages inconsistent with one's NFC level. It was shown that individuals with high NFC were more motivated to process messages that were labeled as complex rather than simple, while individuals with low NFC were motivated to process messages that were labeled as simple rather than complex, even though the content did not differ in complexity.

Lin, Lee, and Horng (2011) also examined the choice of routes between high and low NFC consumers when they were simultaneously exposed to both the central and peripheral routes: the peripheral route was operationalized by manipulating online review quantity while the central route differed by message quality. While review quantity and review quality both had a

significant main effect on purchase intention, there was also an interaction effect. Participants high in NFC expressed more positive attitudes after exposure to the strong argument quality version than after exposure to the weak argument quality version but participants low in NFC did not differ from one version to another. Participants low in NFC expressed more positive attitudes after exposure to large quantity condition than after exposure to the small quantity condition, yet no difference was found for high NFC participants. Lin and Wu (2006) further compared the likelihood of accepting online recommendations between high and low NFC participants and found low NFC participants were more likely to accept recommended alternatives, which provided additional support for the finding that peripheral cues such as heuristics are more influential on those with low NFC.

2.3.1.2 Extent of information processing. Differences in the amount of information processed among people varying in NFC can be found in both online and offline search settings, in educational settings, purchase settings and other decision making contexts. For example, Curseu (2011) explored the effect of NFC on active information seeking in the classroom. The extent of advice seeking in small student groups was investigated with 213 master students who were instructed to form groups of 3 to 7 members. It was shown that NFC was an important asset for active information seeking. People high in NFC reported a higher tendency to actively seek advice from their teammates when they were asked to solve a complex problem regarding a group assignment than people low in NFC.

In consumer behavior research, external information search, which is defined as “the attention, perception, and effort directed toward obtaining environmental data or information related to the specific purchase under consideration” (Verplanken, Pieter, Hazenberg, & Palenewen, 1992, p. 85), can also vary as NFC increases. Verplanken et al. (1992) presented

participants with an information board that displayed a 3 (brands) X 10 (attributes) matrix and asked them to indicate the brand and attributes about which they wished to receive information (An information board is an instrument commonly used in consumer research where brands and attributes of the brands are listed. The content of the attributes are covered and can only be revealed when participants ask to see them). They found that participants scoring low in NFC expressed a desire for fewer attributes than participants who scored high in NFC. Furthermore, participants with high NFC generated more thoughts relevant to the task (e.g., mentioning an attribute, a brand and search strategy) than participants with low NFC, which suggested that more cognitive effort was expended by those with high NFC.

Building on the findings of Verplanken et al. (1992), Verplanken (1993) examined the interactive effect between NFC and time pressure on external information search preceding a purchase decision. They found that regardless of the absence or presence of time pressure, high NFC participants generated more task-related thoughts and they also reported expending more effort than low NFC participants. However, high NFC participants did not actually search for more information than low NFC participants even though in Verplanken et al. (1992) high NFC participants expressed a desire for more information. This indicates that even though high NFC participants did not appear to search for more information than low NFC participants, they could have processed information more intensively by reading it more carefully than low NFC participants.

Bailey (1997) investigated whether or not NFC affected decision strategy in the context of employee hiring. Managers low, medium and high in NFC were asked to engage in *judge* and *choice* response modes: in the former condition participants were asked to evaluate job candidates and in the latter they were asked to choose one candidate based on candidate features

they received on an information board. Bailey (1997) found that high NFC participants conducted more thorough searches (asked to see content of more attributes) than low NFC participants. Even though the *choice* response mode by nature demanded fewer searches, high NFC participants still produced thorough searches.

The relationship between NFC and complex problem solving in management was explored in Nair and Ramnarayan (2000). The authors defined complex problems as problems that were not routine and did not have well-defined solutions. To solve such problems an individual should have a tendency to actively and persistently engage in thinking, thus a high NFC individual was predicted to gather a greater amount of information, gather more diverse information, and be more effective in solving complex problems. Participants in the study were presented with a case description of a company and asked to manage all affairs of the firm as its CEO. The findings showed that as NFC increased, the diversity of information sought about the company increased, but the amount of information sought did not significantly correlate with NFC, which once again shows that NFC may manifest itself in other characteristics of information processing rather than just the amount of information sought.

The positive relationship between NFC and the amount of information processed has been observed in Web search. Das, Echambadi, McCardle, and Luckett (2001) showed that individuals with higher NFC tended to use the Web for information seeking to a greater extent than those with lower NFC based on self-report data collected from a questionnaire. Similarly, Tuten and Bosnjak (2001) found that NFC was positively correlated with Web usage in gathering product information, current events and news, and in learning and education. Both studies demonstrated that a higher orientation to thinking motivated Web search. Different amounts of information processing were also distinguished in studies of Web design. Sicilia, Ruiz, and Jose

(2005) investigated the effect of Website interactivity on information processing. Based on the connection between ELM and NFC, the authors hypothesized that low NFC searchers would rely more on Website interactivity (the peripheral route) rather than Website content (the central route) for determining navigation paths. Therefore, low NFC participants would increase information processing when using an interactive site to a greater extent than high NFC participants. To measure the amount of information processing, participants were instructed to write down all the thoughts that occurred to them after exposure to the sites. While both groups produced more product- and Website-related thoughts when exposed to an interactive than a non-interactive Website, the amount of increase was only significant for low NFC participants.

Amichai-Hamburger, Kaynar, and Fine (2007) examined the effects of Website interactivity and time pressure on Website preferences between high and low NFC individuals. Participants were asked to determine whether or not they were willing to spend \$10 downloading a software program on a commercial site in four conditions: interactive Website with time pressure, interactive Website with no time pressure, flat Website with time pressure, and flat Website with no time pressure. Results showed that low NFC participants were more likely to choose the site when using an interactive Website while Website interactivity had no significant effect on the Website preference among high NFC participants. It was also shown that participants high in NFC spent more time surfing than participants low in NFC. Yet high and low NFC participants did not differ in the number of hyperlinks clicked. The findings provide further support that high NFC individuals are prone to spend more time gathering information and are less likely to be influenced by peripheral information cues present in interface design.

Kaynar and Amichai-Hamburger (2008) broadened the scope by examining the effect of NFC on the use of thirty different Internet services that were divided into three major types:

professional, social and leisure. They found that NFC was positively associated with the use of professional services such as using email or real-time messaging for work related purposes and acquiring information for study purposes, but not with the other two uses. NFC also was found to be positively correlated with perceived importance of information in creating a persuasive site and negatively correlated with perceived importance of environmental characteristics (i.e., graphical searcher interface and technological advancements) in creating a successful site. Not only was the finding consistent with previous studies in that high NFC people valued more central than peripheral cues, the finding that people high in NFC tended to expend cognitive effort only in work-related activities might also suggest that individuals with high NFC will be more likely to devote more effort searching on work-related search tasks than non-work-related tasks.

While Amichai-Hamburger et al. (2007) examined only one peripheral cue, Website interactivity, on Website usage and found no relationship with high NFC individuals, Crystal and Kalyanaraman (2005) distinguished between two Website features and found that they had different effects on high NFC individuals. The moderating role of NFC on the relationships between two usability guidelines, informative feedback and descriptive labeling, and search performance were investigated in an online health information seeking context. Participants were presented with a health-related Website in four conditions: feedback and labeling, no feedback and labeling, feedback and no labeling, and no feedback and no labeling. During the experiment participants were instructed to look for information on the assigned Website in order to answer five multiple choices and five free-response questions. High NFC participants generally answered more questions correctly than low NFC participants. The lack of labeling did not seem to deteriorate the search performance of high NFC participants but it significantly affected the

performance of low NFC participants. The absence or presence of feedback influenced high NFC participants' attitudes toward the Website but not low NFC participants'. These findings suggest that high and low NFC people may react to usability problems differently. Carenini (2001) also demonstrated that not all interface features were treated as peripheral cues by high NFC individuals. They found a positive relationship between NFC and usage of dynamic querying, an interactive technique for database querying, during a real-estate search task. The finding suggests that high NFC potentially increases a person's willingness to use complex interface features to accomplish a demanding search task.

Scholer et al. (2013) investigated whether NFC mediated the extent to which a list of documents, with varying densities of relevant and non-relevant documents, impacted people's relevance judgments of those documents. While the researchers did not find that NFC mediated this relationship, their study participants did not differ greatly with respect to NFC, so lack of variance on this measure might have prevented them from observing an impact.

2.3.2 Summary. Most studies have demonstrated that high NFC leads to increased information processing by way of accessing more information, accessing a greater diversity of information, processing information more thoroughly, or spending more time in search. While it may be challenging to generalize findings from some of these studies to search, findings from other studies are more applicable to the present study. Several studies showed that argument or content quality influenced high NFC individuals more while interface features affected low NFC individuals to a greater extent, and ELM provides a way to explain why high NFC and low NFC searchers process information with different strategies given different SERP characteristics. These studies provide useful perspectives to predict and explain if and how people with different NFC exhibit differences in query stopping and task stopping, for instance, some people may

search deeper than others under certain SERP characteristics. Investigating the interplay between NFC and SERP characteristics can perhaps shed light into the role NFC plays in search stopping behaviors while interacting with SERPs.

Chapter III. Research Questions

While the literature revealed many factors that can affect search stopping behaviors, this work focuses on the effects of two factors, SERP characteristics and personality, and incorporates two theories to establish the arguments. These two theories are: NFC, one that has received little attention in information science but has demonstrated much potential through its impact on information processing from multiple disciplines, and the other, information scent, a theory which has received ample attention in studies of Web navigation and relevance assessment yet has mostly been operationalized at a the surrogate level for a single information resource. Based on these theories, four research questions are formulated:

RQ1: What is the relationship between the information scent level of the first SERP and search stopping behaviors?

Searchers often have to issue several queries before they obtain a sufficient amount of information for open-ended search tasks. It is proposed that the first SERP can be viewed as a surrogate for the entire set of results returned in response to a query. Just as searchers can be made aware of the potential value of a single search result by the amount of information scent of the snippet (Kammerer et al., 2009; Loumakis et al., 2011) and the potential usefulness of an entire website based on its homepage (Card et al., 2001), arguably they may also attempt to predict the potential value of the entire set of results retrieved for a query based on the quality of the initial SERP. Based on the same analogy, the information scent of the first SERP can possibly be used to predict to what extent a searcher will evaluate a set of search results for a

single query. If the *number* of relevant results is higher on the first page, this might increase the interactions with the result set compared to when the first SERP has fewer relevant results.

RQ2: What is the relationship between the information scent pattern of the first SERP and search stopping behaviors?

Modern search engines are convenient to searchers in that they reduce the cost of information access by ranking results by algorithmic relevance. However, searchers might be more likely to believe they have seen all the relevant results once the algorithmically relevant search results no longer satisfy their information needs. In other words, once searchers perceive that information scent is discontinued, they might be more likely to believe that the information patch does not contain more relevant results. This argument can be further supported by Card et al. (2001), who demonstrated that when information scent declined on a web page, searchers tended to leave the website. Their finding suggests that the *distribution* of relevant results, or the sequence of relevant and non-relevant results a searcher encounters on the first SERP, might impact to what extent a searcher interacts with a set of results. Therefore, it is proposed here when relevant search results are scattered across the first SERP, a certain degree of information scent is maintained throughout the first SERP. This in turn, might induce the searcher to interact with the result set to a greater extent.

RQ3: What is the relationship between NFC and search stopping behaviors?

Individuals with high NFC have been found to exert more effort during information processing. This has two possible implications for information search. First, people with high NFC may examine more information for a given query. Since high NFC searchers enjoy the

thinking process, their higher motivation to process information may allow them to be more resilient to non-relevant search results, thus lowering their motivation to reformulate. Second, it is also likely that people with high NFC may exert more effort on query reformulation. Since people with high NFC enjoy cognitive activities and query reformulation is a cognitive task, effort may manifest in more frequent query reformulations rather than prolonged engagement with search results. Support for the latter hypothesis also comes from research that has demonstrated that high NFC people make more accurate judgments about message quality (Cacioppo et al., 1983). If high NFC people are more capable of discriminating high quality from low quality content, they may reformulate as soon as they encounter a bad document in search for higher quality information.

RQ4: How can we model task stopping using interaction signals?

While the literature has accumulated many factors that can affect when a searcher stops looking for information at the task level, most of the studies relied on interview data, thus evidence for supporting causal relationship between the factors and stopping is limited. Besides the theories of NFC and information scent, early IR literature suggests some useful starting points for predicting when a searcher terminates information search for a task. First, Cooper (1968, 1973) discussed the issue of how much information is enough in terms of either number of relevant documents needed or number of non-relevant search results to be tolerated, and Kraft and Lee (1979) also proposed stopping rules that took into account of the quantity of relevant and non-relevant results. These findings suggest that the number of relevant and non-relevant search results can perhaps influence when searchers decide when they have enough information to stop their search task.

IR measures such as DCG and RBP treat query stopping as the rank of the last assessed result in the returned results, suggesting another interaction signal to consider for predicting task stopping. Furthermore, according to Pirolli (2007), search results can be regarded as patches with diminishing returns: “the likelihood more relevant search results will be found with more foraging effort diminishes as a function of how many search results have already been scanned” (p. 52); that is to say, number of search results scanned or evaluated may also affect one’s decision as to when to stop looking. Common search behavior measures such as time and abandonment are also incorporated in this research to leverage the existing understanding of which search behaviors can potentially be used to model task stopping.

Chapter IV. Methods

4.1 Overview

A laboratory experiment was conducted with three independent variables: information scent level (ISL), information scent pattern (ISP) and need for cognition (NFC). The first two independent variables were within-subject variables, while the third was a between-subjects variable. Information scent level (ISL) was defined as the number of relevant documents on the first SERP and was operationalized with three levels: high, medium and low (Table 1). Information scent pattern (ISP) was defined as the distribution of relevant documents on the first SERP and was operationalized with three levels: persistent, disrupted and bursting, each of which differed according to the distribution of four relevant search results on the first SERP (Table 1).

Table 1
Manipulation of Information scent level and information scent pattern

	Tasks 1-3			Tasks 4-6		
	Information Scent Level (ISL)			Information Scent Pattern (ISP)		
Rank	Low	Medium	High	Persistent	Disrupted	Bursting
1	R	R	R	R	R	NR
2	NR	R	R	R	R	NR
3	NR	R	R	NR	R	NR
4	NR	NR	R	NR	R	R
5	NR	NR	R	R	NR	R
6	NR	NR	NR	NR	NR	R
7	NR	NR	NR	NR	NR	R
8	NR	NR	NR	R	NR	NR
9	NR	NR	NR	NR	NR	NR
10	NR	NR	NR	NR	NR	NR
System Effectiveness	0.1111	0.33	0.555556	0.88	1	0.53

Note. System effectiveness is represented by recall @10 for ISL and nDCG @10 for ISP.


During the experiment, a total of six search tasks were assigned to each participant. The study tasks were rotated according to a Latin Square design. In three of the six tasks, ISL was manipulated while in the other three tasks ISP was manipulated (Table 1). Participants were made to believe they were using a custom search engine and were asked to enter self-generated queries to complete the tasks. However, no matter what queries they issued for their first three query submissions for a given task, they received a preselected search result set of 100 results where the first ten results reflected a specific ISL or ISP treatment.

The flow of the search process can be summarized using a particular experimental condition experienced by a participant in the study (Table 2). From left to right, the first three columns indicate the three treatments experienced for the first, the second, and the third query submissions for all six tasks. For example, in task 2, the participant was exposed to medium ISL after submitting the first query, high ISL after submitting the second query, and low ISL after submitting the third query. The order of these treatments was rotated to balance order and sequence effects. More details about the rotations are discussed later. Results presented at the 11th to 100th positions for the first three sets of results were also preselected. Only the twelfth, fifteenth and eighteenth results on the second SERPs were relevant; more relevant documents were not provided on this page because participants' transitions from the first SERP was of primary interest. However, some relevant documents were present to prevent participants who went to these subsequent pages to learn from their interactions that paginating to the second page always ended up futile. Participants were not required to view all SERPs or enter any pre-specified number of queries. If a participant submitted more than three queries, starting from the fourth query submission, the Bing search API ¹ was used to fetch results. To control for the

¹ <http://datamarket.azure.com/dataset/bing/search>

experience from interacting with each search result, participants were asked to not follow links on landing pages and not open multiple tabs.

Table 2
Search Result Evaluation Flow

	1st Query	2 nd Query	3 rd Query	From the 4 th query....
Task 2	Medium	High	Low	
Task 3	High	Low	Medium	
Task 4	Persistent	Disrupted	Bursting	
Task 5	Disrupted	Bursting	Persistent	
Task 6	Bursting	Persistent	Disrupted	
Task 1	Low	Medium	High	

Note. Tasks and treatments are rotated.

As mentioned in the previous paragraph, treatments were displayed on the first SERPs of the three preselected search result sets in each task. The order in which treatments were displayed for the first query, the second query and the third query was rotated in every possible combination to balance sequence effects (effects resulted from preceding treatments) and order effects (effects resulted from the positions of treatments among other treatments). These can be seen in Table 3 for information scent level and Table 4 for information scent pattern. The order of the six study tasks were rotated using a Latin Square design to balance order effects (Table 12). Detailed information about the study tasks can be found in “4.5 Tasks”.

Table 3
All Rotations for Information Scent Level Treatments

1st Query	2 nd Query	3rd Query
High	Med	Low
High	Low	Med
Med	High	Low
Med	Low	High
Low	High	Med
Low	Med	High

Table 4

All Rotations for Information Scent Pattern Treatments

1 st Query	2 nd Query	3 rd Query
Persistent	Disrupted	Bursting
Persistent	Bursting	Disrupted
Disrupted	Persistent	Bursting
Disrupted	Bursting	Persistent
Bursting	Persistent	Disrupted
Bursting	Disrupted	Persistent

4.2 Manipulation of Information Scent Level

The manipulation of information scent level was limited to the first SERP of each preselected search result set, namely, the first ten search results. Specifically, the high, medium and low information scent levels were distinguished by placing different numbers of relevant search results on the SERP. One relevant document, three relevant documents, and five relevant documents were placed on the SERP to reflect low, medium and high levels, respectively. To control for the influence of result positioning, relevant results were always placed consecutively starting from the first result and also represented the best possible ranking for each level (Table 1). The best possible rankings were selected because modern search engines attempt to place documents by descending algorithmic relevance. The objective system effectiveness for each treatment was computed by recall @ 10, assuming the total number of relevant documents was 9 (from adding all relevant documents on all three treatments: 1+3+5).

There was a one-to-one relationship between the treatments and the documents shown. Tables 5 and 6 show two example rotations for a task where information scent levels were manipulated. Bold DocIDs are relevant documents while non-bold DocIDs are non-relevant. Documents are fixed to treatments. For example, while the low treatment takes place after the

first query in Table 5 and after the second query in Table 6, the ten documents in the low treatment are exactly the same in both rotations.

Table 5 (Left) & Table 6 (Right)
Two Examples of Document Placement in Relation to Information Scent Level Treatments

Rank	1 st Search Result Set (Low)	2 nd Search Result Set (Med)	3 rd Search Result Set (High)
1	Doc1	Doc2	Doc6
2	Doc13	Doc3	Doc7
3	Doc14	Doc4	Doc8
4	Doc15	Doc5	Doc9
5	Doc16	Doc22	Doc10
6	Doc17	Doc23	Doc11
7	Doc18	Doc24	Doc12
8	Doc19	Doc25	Doc28
9	Doc20	Doc26	Doc29
10	Doc21	Doc27	Doc30

Rank	1 st Search Result Set (Med)	2 nd Search Result Set (Low)	3 rd Search Result Set (High)
1	Doc2	Doc1	Doc6
2	Doc3	Doc13	Doc7
3	Doc4	Doc14	Doc8
4	Doc5	Doc15	Doc9
5	Doc22	Doc16	Doc10
6	Doc23	Doc17	Doc11
7	Doc24	Doc18	Doc12
8	Doc25	Doc19	Doc28
9	Doc26	Doc20	Doc29
10	Doc27	Doc21	Doc30

Note. Even though information scent level treatments were presented in two different rotations, the specific documents displayed for each treatment did not change by rotation.

For the 100 preselected results preselected for each query submission, results from the second to the tenth SERP were fixed to positions regardless of treatment. The 2nd, 5th and 8th result on the second SERPs were relevant while all other results from the second to the tenth SERP were non-relevant to the task.

4.3 Manipulation of Information Scent Pattern

The manipulation of information scent patterns was also limited to the first SERP of each preselected search result set. Three information scent patterns were examined, differing in terms of the distribution of relevant search results on the first SERP (Table 1). Four relevant search

results were included on the SERP to create three information scent patterns: persistent, disrupted, and bursting. Four relevant documents were chosen instead of other numbers because of a concern that with fewer documents it would have been difficult to construct patterns, and with more documents participants could have felt that they had found sufficient information on the first SERP. System effective for each treatment was computed by nDCG @10.

While information scent level is concerned with the “amount” of scent, information scent pattern is related to the “sequence” of scent searchers experience, assuming that searchers process search results linearly from the first rank (c.f. Joachim et al., 2005). As the concept of information scent pattern is not widely researched in the literature, detailed explanations as to why the information scent pattern treatments in the current study were chosen and what they mean in real life are provided below.

The *persistent* information scent pattern simulates a scenario where results relevant to a topic are dispersed on the SERP. When coming across a SERP with persistent scent, a searcher does not receive all relevant search results in the first few ranks; rather, he or she finds relevant search results one after another every one or two ranks apart. The persistent pattern initially has a strong scent, with relevant documents in the first two positions, followed by two more relevant documents at a consistent interval. This information scent pattern is used to represent a scenario where scent continues along the trail.

It is possible to imagine the case where all relevant search results are found at the top ranks of the SERP. This pattern is simulated in the current study by placing all four relevant search results at the first to the fourth ranked positions on the SERP. From the search engine’s perspective, this information scent pattern is what would be regarded as the best output ranking. However, from the information scent perspective, searchers may not continue to explore results

beyond the first SERP because there is no scent toward the end of the first SERP, even though in reality they may encounter more relevant results beyond the first SERP. Therefore, this information scent pattern is called the *disrupted* information scent pattern to represent the loss of scent from the mid-region to the end of the SERP.

Still there are cases when searchers do not immediately receive relevant search results, but encounter them as they continue exploring. To simulate this information scent pattern, four relevant search results are placed from the fourth to the seventh rank on the SERP. This treatment is called the *bursting* information scent pattern, in which the scent is not initially present, but then appears strong and steady in the middle of the list before extinguishing, giving searchers the impression that all relevant search results have been observed.

The search results on the first SERP were rearranged within the SERP to reflect different treatments in the study. Two example rotations are shown in Tables 7 and 8; relevant results are bold while non-relevant results are not bold. As the ten first results from the first SERP in Table 7 are compared to Table 8, it can be seen that Doc 181 to Doc 184 are always fixed to the first search result set, yet they are rearranged to accommodate to the specific treatment. This approach ensured that the effect of information scent pattern would not be confounded by any single idiosyncratic set of four results selected. Ideally the same approach should be taken when it came to manipulating information scent levels, yet it would have entailed moving around relevant and non-relevant documents among the three treatments, which could have introduced additional confounding variables.

Table 7 (Left) & Table 8 (Right)

Two Examples of Documents Placement in Relation to Information Scent Pattern Treatments

Rank	1 st Search Result Set (Persistent)	2 nd Search Result Set (Disrupted)	3 rd Search Result Set (Bursting)	Rank	1 st Search Result Set (Disrupted)	2 nd Search Result Set (Persistent)	3 rd Search Result Set (Bursting)
1	Doc181	Doc185	Doc205	1	Doc181	Doc185	Doc205
2	Doc182	Doc186	Doc206	2	Doc182	Doc186	Doc206
3	Doc193	Doc187	Doc207	3	Doc183	Doc199	Doc207
4	Doc194	Doc188	Doc189	4	Doc184	Doc200	Doc189
5	Doc183	Doc199	Doc190	5	Doc193	Doc187	Doc190
6	Doc195	Doc200	Doc191	6	Doc194	Doc201	Doc191
7	Doc196	Doc201	Doc192	7	Doc195	Doc202	Doc192
8	Doc184	Doc202	Doc208	8	Doc196	Doc188	Doc208
9	Doc197	Doc203	Doc209	9	Doc197	Doc203	Doc209
10	Doc198	Doc204	Doc210	10	Doc198	Doc204	Doc210

Note. Documents are fixed to search result sets regardless of rotations. The same ten documents on each SERP are rearranged to reflect different information scent pattern treatments.

Results from the second to the tenth SERP in all search result sets were fixed to positions regardless of experimental conditions, too. The 2nd, 5th and 8th result on the second SERPs were relevant results while all other results from the second to the tenth SERP were non-relevant.

4.4 Measurement of Need for Cognition

The 18-item NFC Scale developed by Cacioppo, Petty, and Kao (1984) was used to measure participants' NFC orientation. The NFC scale items were measured on a 5-point scale, where 5 = "extremely characteristic of me" and 1 = "extremely uncharacteristic of me." A NFC score for each participant was derived by averaging a participant's ratings of the 18 items. Participants were not grouped according to their NFC scores because the richness of continuous data can be lost when continuous data is transformed into discrete data, which has been criticized by researchers for two reasons: (1) Dichotomizing continuous independent variables reduces the statistical power available to test hypotheses; and (2) inappropriate dichotomizing of continuous

data can also create spurious significant results when the independent variables are correlated (Fitzsimons, 2008). In addition, as no hypothesis was proposed about the relationship between NFC and search stopping behaviors, the analysis of its potential effect on stopping behavior was exploratory in nature.

4.5 Tasks

4.5.1 Overview. The goal of the present study is to investigate search stopping behaviors when searchers are engaged with tasks that do not have obvious end points and that require multiple documents to satisfy their information needs. In order to render search stopping behaviors comparable under different treatments while at the same time retain the realism of the searches, simulated work tasks were used (Borlund, 2003). While a common approach in Interactive Information Retrieval research is for researchers to design ad-hoc tasks to satisfy their specific research goals, there was a concern that newly developed tasks that had not been systematically evaluated would exhibit inconsistent qualities that could impact result interpretation. For example, some tasks might appear to be more difficult than others. In this study, a set of tasks that had been used in a previous project was used so that more evidence was available about how people would search. One specific benefit of using these prior tasks was that the interaction signals already collected from the previous study provided evidence about the range and types of queries people might submit. This information was extremely important to avoid selecting tasks where participants would enter many different queries, as preselecting search results relevant for many queries would be difficult.

During the task selection process, not only were empirical data generated from the previous work considered, but also task descriptions were revised to address some shortcomings of the original descriptions to ensure all study tasks had the same task description format. These

efforts resulted in a selection of six study tasks for the present study. Since the research in which these study tasks were developed and evaluated has not been published, background information of these study tasks is described in more detail in the following. Preliminary results of this study can be found in Wu et al. (2012).

4.5.2 Task development and evaluation from the previous work. My colleagues and I worked on a project with a goal of developing a set of search tasks to be used by the Interactive Information Retrieval community. In light of the trend that many search tasks were developed arbitrarily, we decided to construct search tasks with a systematic approach and we picked the concept of cognitive complexity as the framework. The concept of cognitive complexity arises from the cognitive process dimension from Anderson and Krathwohl's Taxonomy of Learning (Anderson & Krathwohl, 2001), a well-known resource for creating educational exercises and exams. In their taxonomy, six types of cognitive processes are identified: *remember*, *understand*, *apply*, *analyze*, *evaluate* and *create*, with each type requiring increasing amounts of cognitive processing. We chose five types, *remember*, *understand*, *analyze*, *evaluate* and *create*, and four domains, *commerce*, *science and technology*, *health* and *entertainment*, and created a total of 20 search tasks. *Apply* was not selected because we were unable to generate appropriate search tasks meeting the criteria.

Forty-eight undergraduate UNC students were recruited to conduct searches on the 20 tasks. Each of them completed five tasks of varying cognitive complexity level in one domain. During the search their interactions with the system were logged. They were also asked to evaluate how difficult the tasks were at the end of the study. It was found that as the cognitive complexity of a task increased, participants spent more time searching, submitted more queries, and visited more Webpages. The findings provided empirical evidence that tasks of the same

cognitive complexity level are more similar in terms of the behavior exhibited to address information needs than tasks of a different complexity level. This means that using tasks of the same cognitive complexity for my dissertation research to some extent should help eliminate potential task effects on search stopping behaviors.

4.5.3 Task selection. Among the tasks with five different levels of cognitive complexity, evaluate tasks were the most appropriate for my study goal. During *evaluate* tasks one had to make judgments based on criteria and standards through checking and critiquing, which was similar to the scenario where Web users often explore topics and gather multiple documents in order to support real life decisions. An examination of the queries submitted when participants were conducting the evaluate tasks showed that during the evaluate tasks of health, science and technology, and entertainment, participants generally submitted generic query terms which were directly extracted from the task description stems (example queries can be found in Appendix A). However, in the evaluate task of commerce, participants submitted product names (i.e., Toyota, Honda, & Jeep Liberty) interchangeably in unpredicted orders to retrieve relevant information. Such variability rendered preselecting search results almost impossible and risky; therefore, the commerce task was eliminated from further consideration.

In order to construct a total of six tasks for the present study, three analyze tasks (Health, Science and Technology, and Entertainment) were compared to evaluate tasks to examine whether they were similar in nature. If analyze tasks exhibited no significant differences from evaluate tasks, they could be used as the study tasks, too. Analyze tasks were considered because they were different from evaluate tasks by only one level of cognitive complexity. Even though Create tasks were also only one cognitive complexity level different from evaluate tasks,

completion of these tasks required more personal creativity, which posed great challenges in preselecting search results.

The range of participants' self-reported task difficulty on the six tasks is shown in Figure 2. An ANOVA was applied to test the differences among them and it was shown that the six tasks yielded similar self-reported task difficulty ($F(5, 71)=1.39, p=.238$). In addition, Table 9 provides the descriptive statistics of the interaction signals for the six tasks, including time to complete a task, number of queries submitted during a task, and number of URLs visited during a task. ANOVA results show that these six tasks did not differ significantly from one another in terms of time spent on search ($F(5, 71)=.371, p=.867$). However, significant differences were found in terms of number of queries submitted ($F(5, 71)=3.997, p=.003$) and number of URLs visited ($F(5, 71)=2.833, p=.002$) among these tasks. Bonferonni post-hoc tests show that significantly more queries were submitted to the analyze task in the Entertainment domain than the evaluate task in the Entertainment domain and the analyze task in the Science and Technology domain. The analyze task in the Entertainment domain also had significantly more URL visits than the evaluate task in Entertainment. These findings suggest the analyze task in the Entertainment domain was different in nature from most of the other tasks.

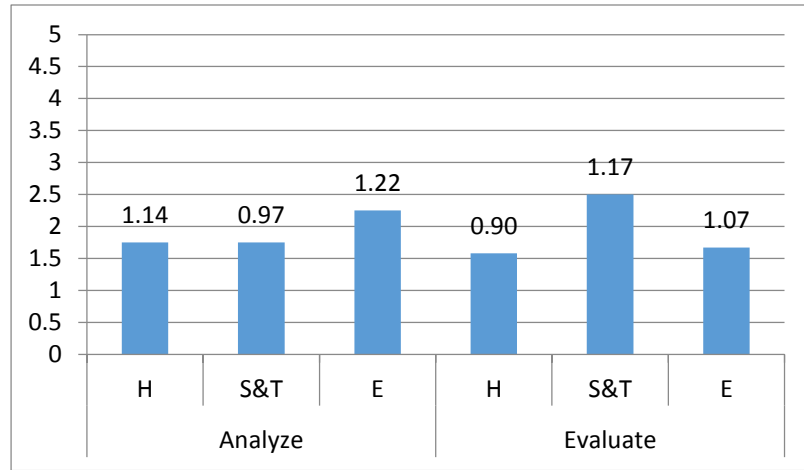


Figure 2. Self-reported task difficulty of analyze and evaluate tasks. The labels indicate SD. (H=Health; S&T=Science & Technology; E=Entertainment)

Table 9

Interaction Signals of Analyze and Evaluate Tasks

Task Type	Domain	Time to Completion	Number of Queries	Number of URLs Visited
Analyze	Health	532.67	2.83	5.50
	Science & Technology	460.00	2.00	3.67
	Entertainment	472.50	4.00	7.92
Average		488.39	2.94	5.69
Evaluate	Health	520.67	2.33	5.92
	Science & Technology	571.00	2.33	5.17
	Entertainment	438.00	1.25	3.33
Average		509.89	1.97	4.81

An examination of the queries submitted for the three analyze tasks showed that in the health and science and technology tasks many query terms were actually the keywords appearing in the task descriptions (Appendix A), which suggested that it would be reasonable to assume

that future participants' queries would follow the same trend. It also meant that it would be likely that preselected results would match self-generated queries. However, in the entertainment task it was observed that some participants used specific extreme sport names as queries, which meant the situation where search results did not align with self-generated queries could take place for this task. This finding again suggested that the analyze task in the entertainment domain needed revision.

4.5.4 Task adaptations. The statistical results described in the previous section showed that some analyze tasks were more similar to evaluate tasks while the Entertainment task required more work in adaptation. In addition, the analyze tasks in general also needed adaptations to reach the same level of task requirements as the evaluate tasks. The following explains the steps taken to adapt analyze tasks to evaluate tasks and details additional strategies taken to control for the range of self-generated queries in the dissertation research.

The first step to adapt analyze tasks into evaluate tasks involved comparing the similarities and differences in task descriptions between the two task types. In both task types, the original task descriptions required participants to identify and describe options, yet it was only in evaluate tasks participants were asked to compare the options and make decisions. In order to increase the cognitive complexity level of analyze tasks and keep the task descriptions uniform, one question was added at the end of the original task description for analyze tasks to urge decision making, such as “which one would you recommend?” or “which method do you think is the best?”

After the above mentioned adjustment, all six study tasks conformed to the same task description format: first, a scenario was presented to motivate information need; second, all tasks asked participants to find information in order to identify options and describe the differences

among the options; lastly, participants were asked to make a judgment or to indicate a preference based on the differences among these options.

From the “Task Selection” section it was demonstrated that during the search for the analyze task in the Entertainment domain significantly more queries were submitted than other tasks to address the information need, and many of them involved specific extreme sport names. Such differences from other tasks can potentially be resolved by narrowing down the task requirement. Instead of asking participants to evaluate “extreme sports”, the requirement was changed to evaluate “extreme sports on the water”. There were fewer extreme water sport names to submit, and even if participants chose to submit specific water extreme sport names to the system and the search results were about another water extreme sport, the results would not appear to be as non-relevant as compared to when the results were about miscellaneous types of extreme sports.

Lastly, since participants tended to use terms directly from task descriptions as search queries, in most task description the keywords were rephrased in at least two different ways so that participants would have more word choices. For example, in the extreme sports task, “extreme sports on the water” and “action water sports” were used interchangeably in the task description. The goal of providing more word choices was to decrease the likelihood of a participant submitting idiosyncratic query terms, which could result in a mismatch between queries and preselected search results and suspicion on the part of the participant.

The resulting six study tasks included two health tasks, two science and technology tasks, and two entertainment tasks (Table 10). Tasks #1, #2, and #3 were originally analyze tasks and tasks #4, #5, and #6 were evaluate tasks to begin with. Tasks #1, #2, and #3 were used to investigate the effect of information scent level and tasks #4, #5, and #6 were used to investigate

the effect of information scent pattern; this decision was made so that participants would experience a wider range of SERP characteristics during the experiment as opposed to only ISL or ISP treatments. The order of the tasks was rotated using a Latin Square design to balance order effects, as shown in Table 11.

Table 10

Study Tasks

	Information Scent Level Tasks		Information Scent Pattern Tasks	
	Task ID	Task Description	Task ID	Task Description
Health	1	Having heard some of the recent reports on risks of natural tanning, it seems like a better idea to sport an artificial tan this summer. What are some of the different types of artificial tanning methods? How risky are they? Which one would you recommend?	4	One of your siblings got a spur of the moment tattoo, and now regrets it. What are the current available methods for tattoo removal, and how effective are they? Which method do you think is best?
Science & Technology	2	You recently became involved with a conservation group that picks-up trash from local waterways. One of the group members told you that your work was important because it helps keep pollution out of the ocean. What are some of the different types of marine pollutants? What environmental risks are associated with each pollutant? Which one seems to be the most harmful to the environment?	5	Many people believe that social media has many benefits to our life, but your sibling argues that people are losing their ability to communicate face-to-face and that social media makes people lonelier and more isolated. What are the positive and negative consequences of using social media? Do you believe that social media is detrimental to the development of personal relationships?
Entertainment	3	Your sister is turning 25 next month and wants to do something exciting for her birthday. She is considering some type of extreme sport on the water. What are some different types of action water sports in which amateurs can participate? What are the risks involved with each one? Which one would you recommend?	6	For his 16th birthday, your nephew has asked you for a video game that is rated "M" for mature audiences because it contains intense violence. You are unsure about whether to purchase this game because you recently overheard two people discussing the effects of violent video games on teenagers. What are some positive and negative effects of playing violent video games on teenagers? Would you buy a video game rated "M" for him?

Table 11

Latin Square Design of Task Order

1 st	2 nd	3 rd	4 th	5 th	6 th
1	2	3	4	5	6
2	3	4	5	6	1
3	4	5	6	1	2
4	5	6	1	2	3
5	6	1	2	3	4
6	1	2	3	4	5

4.6 Documents

Special care was taken to ensure that participants would experience the preselected search results in accordance to the manipulated information scent levels and patterns. Relevant search results were selected from the clicked webpages gathered by participants in the previous project (except relevant results for task #3 which were gathered by submitting queries to a popular search engine instead). Non-relevant search results were identified by submitting queries composed of a keyword from the task description and some terms unrelated to the task to a popular search engine. For example, non-relevant search results for one task which was about methods of tattoo removal were gathered by submitting the queries *tattoo designs* and *tattoo mistakes*. All relevant and non-relevant pages presented on the first two SERPs were reviewed by the researcher and two other assessors. Only Webpages that were judged by all three people as relevant or non-relevant were used. In addition, the search results displayed on the third and fourth SERPs were verified by the researcher to make sure they were non-relevant to the study tasks. Note that search results presented on the third through tenth SERP did not repeat any non-relevant search results appearing on the first two SERPs.

It was also important to ensure that the result summaries clearly reflected whether a landing page was relevant or not so that participants would experience the intended information

scents. Search result summaries were generated using Bing API; each summary included a title, a display URL, and a snippet (Figure 3). For the relevant documents reused from the previous project, two to three sentences were manually selected from the content as the search result summaries. All search result summaries on the first and second SERPs were evaluated by three assessors as well. It is worth noting that the summary generation approach applied here is likely to decrease the ecological validity of the study since search result summaries were manipulated to match the task rather than to match any query, which is the norm.



The image shows a search result summary enclosed in a blue border. The title is "Tattoo Pictures & Designs: Every Tattoo Magazine" in blue text. Below the title is the URL "www.everytattoo.com" in green text. The snippet text is "Tattoos and tattoo pictures of: Dragons, Angels, Stars. Tattoo conventions and tattoo parlor listings. Many tattoo flash designs, both real tattoos and fake tattoos" in blue text.

Figure 3. An example search result summary generated by Bing API by submitting “Tattoo art”.

4.7 Experimental Search System

An experimental search system was built to carry out the experiments. At the start of the task, the task description was shown at the top along with a query box. After participants submitted their initial queries, a page of ten results was displayed. When a participant clicked on a search result, the landing page was presented in a separate tab and participants were asked whether they wanted to save the page (Figure 4). Once participants submitted a response, the tab automatically closed and participants were taken back to the SERP. If participants attempted to close the tab without answering the question, a warning message appeared. Participants clicked “Done” in the upper right corner of the SERP when they finished the task.

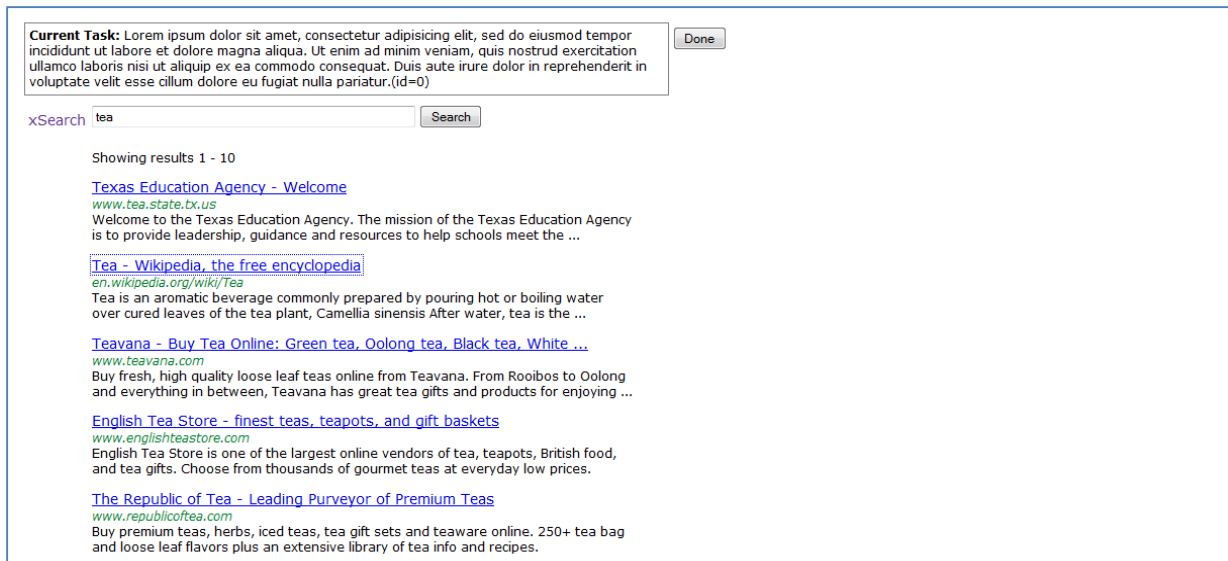


Figure 4. A search engine result page of the experimental search system.

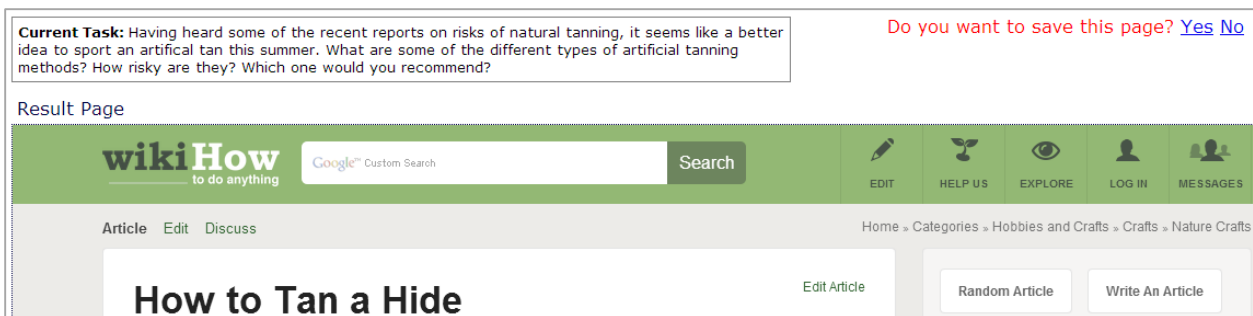


Figure 5. A landing page of the experimental search system.

4.8 Search Behaviors Measures

The dependent measures consisted of search behaviors as shown below, which were recorded in search logs:

QueryAction: a categorical measure of the outcome after a query submission. The measure has three values: “*Reformulation*”, or query reformulation on the first SERP; “*Pagination*”, or advancing beyond the first SERP; “*Stopping*”, or clicking on “Done” on the first SERP.

Abandonment: not clicking on a SERP after a query submission.

NumPagination: number of SERPs visited within a search result set.

NumQuery: number of query submissions during a task.

Time: amount of time spent examining a search result set.

DeepestRankClick: the deepest rank of a clicked result. For instance, if the result of the deepest rank clicked in a result set was ranked at the 5th position, “5” was recorded as the stopping point, whereas if the deepest clicked result was the 3rd on the second SERP, “13” was recorded.

DeepestRankHover: the deepest rank where a mouse hover was observed.

DeepestRankHover was used to capture the amount of attention invested to searching for useful information but did not transfer into clicking. Previous work has shown that gaze and cursor movement are correlated (Huang et al., 2012); about 33% of participants in the study moved mouse cursors around while they examined the page and another 2.5% of participants used mouse cursors to follow the text as they read the page. Rodden et al., (2007) and Guo and Agichtein (2010) also found that the distance between gaze and cursor positions was smaller along the y-axis than the x-axis on SERPs; moreover, Rodden et al., (2007) found gaze and mouse cursor were in the same region 42.2% of the time. These two studies further demonstrate the potential usefulness of using mouse hover to capture the deepest rank of result a participant attended to. While mouse hover does not necessarily align with attention with 100% accuracy, this measure was not used to replace mouse click so it should be treated as an additional signal for understanding search depth. *NumPagination*, *DeepestRankClick* and *DeepestRankHover* were all used to characterize the depth of search in a search result set.

NumExamined: number of documents examined for each search result set.

NumPred, *NumRele* and *NumNonRele*: A *predictive* judgment of relevance is made when a searcher views only a search result snippet, while an *evaluative* judgment of relevance is made after a searcher examines the content (Rieh, 2002). *NumPred* was the number of documents participants clicked on but decided not to save; *NumRele* was the number of documents that were clicked and saved; *NumNonRele* refers to the number of documents that were ranked before *DeepestRankClick* but not clicked (only the snippets were examined).

Participants' searches were also captured with *Morae* and at the end of the study they were interviewed using stimulated recall with video recordings of three of their searches.

4.9 Recruitment

In a previous study where NFC scores were obtained from UNC undergraduate students (Scholer et al., 2013), it was found that a majority of students scored within one standard deviation above or below the average ($M=3.16$, $SD=.56$), which suggested that undergraduate students from a typical four-year university could be homogeneous on the NFC scale. Therefore, participants were recruited by sending an email to the staff mailing list at UNC instead (Appendix G). To qualify, participants needed to be at least 18 years old, be proficient English speakers and have at least two years of Web search experience.

4.10 Procedure

The 1 to 1.5 hour protocol was administered individually in the Interactive Information Retrieval Lab in Manning Hall on campus and consisted of two stages: (1) online searching and (2) a post-task interview. In the first stage, participants filled out an entry questionnaire, which contained questions related to their Web search experience and search-self efficacy (Niu &

Kelly, 2014) (Appendix B). An instruction sheet (Appendix C) was handed to participants after the entry questionnaire. Once participants finished reading the instructions, they were directed to conduct a practice task using the experimental system. If participants had no questions regarding the procedure, they proceeded to the first study task. Before each search task, participants filled out a Pre-Task Questionnaire (Appendix D). During the search process, participants were allowed to submit as many queries and view as many search results as they desired just like using a regular search engine, except that when a search result was clicked, participants were asked to answer whether they wanted to save the page. After a task was completed, participants were directed to fill out the Post-Task Questionnaire (Appendix E). The same steps were repeated for five other study tasks. When all six tasks were completed, participants were asked to fill out the Exit Questionnaire (Appendix F), which contained the NFC scale questionnaire items. The NFC scale was shown at the end of the study so as not to influence search behavior.

After participants completed all six tasks, a semi-structured, stimulated recall interview was conducted to understand their search strategies for three of the study tasks. Half of the participants were interviewed about the three information scent level manipulated tasks while the other half about the three information scent pattern manipulated tasks. The tasks were chosen in a way that every search task had an equal chance of being at each task order; the goal was to balance the potential effects of memory loss. Participants were shown the Morae video and were asked questions to understand what they were thinking about as they searched. The purpose of playing the video was to facilitate recall. At the beginning of each play-back for each task, the experimenter started by asking the participants what their first impression was at the sight of seeing the task description. Then the experimenter played the videos until a pagination or reformulation took place on the screen and then asked participants to explain why they

paginated/reformulated; this question was intended to collect information about factors contributing to query stopping. Periodically, participants were also asked why they clicked on certain results. The question about reformulation and pagination was asked repeatedly until the end of each task. Then, participants were asked how they decided it was the time to stop searching. At the end of the interview they were asked to share a real life search experience during which they paginated multiple times to fulfill their information needs.

4.11 Pilots

Eight pilot tests were administered prior to the official study to examine the amount of time needed to complete the six study tasks. Pilot participants included PhD students from the school of Information and Library Science, the school of Journalism and Mass Communication, the department of Biostatistics, and a staff working for the School of Public Health at UNC. The pilot tests also served to test whether participants could understand the study instructions and whether the experimental system was self-explanatory. Moreover, the pilot tests were conducted to ensure that all the manipulations worked and did not cause suspicion. Lastly, the pilot tests checked whether people submitted unexpected queries to the system, which could cause obvious mismatches between self-generated queries and preselected documents. Wordings of the instruction sheet and of the system were revised based on pilot participants' feedback.

4.12 Data Analysis

4.12.1 Quantitative data analysis. Search behavior data was extracted from the logs captured on the server side (Appendix H). Search behavior data aggregated at the query level (each row represents data points to a unique experimental treatment) and at the task level (each row represents data points to a unique task) were saved in two separate Excel files for the statistical analyses of query stopping and task stopping, respectively, which can be found in

Appendix I and Appendix J. SPSS and SAS were used to conduct descriptive and inferential statistical analyses. Specific statistical methods employed in the analyses are discussed in detail in the Results section.

4.12.2 Qualitative data analysis. The audio recordings of the interviews were listened to first to prepare the materials for the qualitative data analysis. The interviews were not completely transcribed, but passages relevant to the research questions were transcribed to enable subsequent coding. Passages from each individual participant were organized into one unique row in an Excel spreadsheet and within each row passages were further categorized into three columns corresponding to (1) query stopping (2) task stopping (3) past search experience involving pagination. In the second round, an additional column was added next to each existing column to enter codes that applied to the quotes. In the third round the codes were compared and refined.

A combination of deductive and inductive coding was used. Two codes based on the theory of information scent were used initially (deductive): “Level” and “Pattern”, which were associated with passages where the number of relevant documents and the distribution of relevant documents on the first page of results were said to influence query stopping. These two codes were chosen because it was hoped that evidence that could help explain how information scent level and information scent pattern each influenced participants’ stopping decisions could be identified. In addition to these codes, other codes were inductively identified during the coding process.

It is worth noting that the search results experienced by participants in the study were subject to manipulations, which may have caused responses that were different from average Web search scenarios. Nevertheless, the manipulations allowed the present study to better

understand the rationales behind participants' stopping decisions under a wide range of search result characteristics. The goal of the qualitative data analysis was not to differentiate behavioral variances by treatments but to describe factors leading to query stopping and task stopping, more generally.

Chapter V. Results

The results section is divided into seven subsections: first, the demographics of the research participants are provided. Secondly, participants' pre-task expectations and post-task evaluations regarding the study tasks are summarized. Thirdly, results of the manipulation check for the treatments is reported. An overview of the findings from the first stage of the study is then provided. Following this, a section is dedicated to query stopping and another section to describing and predicting tasks stopping. Lastly, the observations from the interviews regarding participants' query and task stopping strategies are reported and discussed. Information scent level and information scent pattern are respectively abbreviated below as ISL and ISP for the ease of reading.

5.1 Participants

Forty-eight people participated, but only data from 47 participants were included because of a logging failure. Seventeen participants were male (36%) and 30 were female (64%). Participants' average age was 38.29 (range: 19-65). Their job titles included Web developer, HR specialist, financial aid counselor, administrative assistant, librarian, lab manager, instructor, research assistant, play writer, fire department technician, and sales manager. Participants scored an average of 7.81 ($SD=1.34$) on a 10-point search self-efficacy scale, showing a medium to high level of confidence in web search skills. Participants were paid \$20 cash for their participation.

5.2 Tasks

Overall, 78.42% of participants had never attempted searches on the topics of the search

tasks (Search_Experience) and participants did not know much about the topics before searching (Knowledge) ($M=2.18$, $SD=0.95$). Participants were moderately interested in the search topics (Interest) ($M=3.22$, $SD=1.20$) and expected the search tasks to be moderately easy before search (Pre-Task Difficulty) ($M=2.72$, $SD=1.01$). After the search tasks, participants still found the tasks to be moderately easy (Post-Task Difficulty) ($M=2.41$, $SD=1.13$) and found it moderately easy to determine when they had enough information (Difficulty_Enough) ($M=2.68$, $SD=1.14$). When asked how successful they were at solving the tasks (Success_Self) and how successfully the system was at finding relevant information (Success_System), it was shown that participants believed both they and the system were relatively successful ($M=3.98$, $SD=0.85$; $M=3.87$, $SD=1.02$).

5.3 Manipulation Check

During each search task, participants were shown preselected search result sets for their first three queries. Among the documents participants clicked on for their first three queries, 98.22% were judged relevant by the assessors and among those they saved, 99.22% were judged relevant by the assessors, which shows the manipulation of document relevance was successful. To examine whether the pre-selected search results caused any suspicion, participants were asked at the end of the experiment to comment on the quality of the search results. Most participants reported the quality was good. Some commented there were many non-relevant results on the first page, but explained this by the popularity of certain webpages or advertisements or attributed this to their own ambiguous queries. No participant indicated they suspected manipulation.

5.4 Overview

Forty-seven participants completed a total of 282 tasks. Participants were able to enter as many queries as desired. Figure 6 shows the distribution of the number of queries submitted per task. In about 60% of the tasks, participants submitted 1-3 queries.

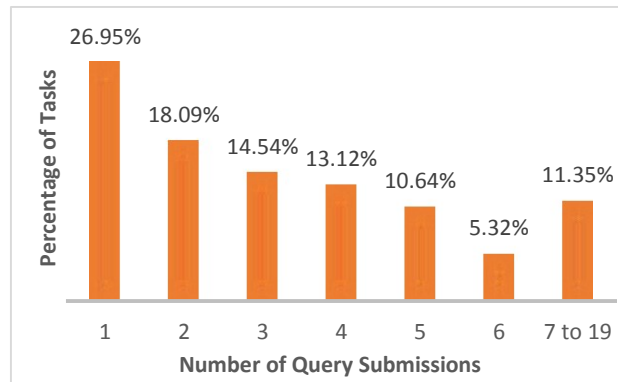


Figure 6. Percentage of tasks where various numbers of queries were observed.

Descriptive statistics for continuous search behavior measures aggregated at the task level can be seen in Table 12. “n” next to the mode represents the frequency of tasks for the most common action. On average, participants issued 3.47 queries per task, and for 76 tasks only one query was issued. Forty-six of the 76 one-query tasks were ISP tasks (four relevant results were presented on the first SERPs), and in another 19 tasks, participants encountered the high ISL (five relevant results on the first SERPs). In addition, 32 out of the 76 tasks involved at least one pagination. This breakdown of one-query tasks suggests that encountering more relevant documents resulted in satisfaction without reformulation. Participants paginated an average of 1.47 times per task; however, in more than half of the tasks they never paginated, and as many as twelve participants never paginated during the entire experiment. Tasks lasted 5.6 minutes on average with a large range: the minimum time was 76 seconds and the maximum, 19 minutes. Participants examined an average of 6.7 results and saved 4.82 results per task. They also clicked

on 1.76 non-relevant results and examined 10.59 snippets of non-relevant results before they terminated searching. DeepestRankClick and DeepestRankHover indicate the accumulative depth of mouse click and mouse hover aggregated from all query submissions in a task. For a search task, participants' average accumulative depth of click was 17.28 and the accumulative depth of mouse hover was 33.22. Table 12 also shows that the median of every measure was smaller than the average, indicating the distributions of the search behavior measures were skewed to the right.

Table 12
Search behaviors at the task level

Measures	Mean	Median	Mode	SD
Time (sec)	332	291	--	427
NumQuery	3.47	3	1 (n=76)	2.67
NumPagination	1.47	0	0 (n=167)	2.97
NumExamined	6.70	6	4 (n=56)	3.96
DeepestRankClick	17.28	14	12 (n=17)	14.36
DeepestRankHover	33.22	19	19 (n=24)	35.15
NumPred	1.76	1	0 (n=98)	2.28
NumRele	4.82	4	4 (n=71)	2.70
NumNonRele	10.59	8	0, 1, 5 (n=66)	11.79

During the three search tasks where ISL was manipulated, participants were exposed to 105 low, 109 medium, and 126 high result manipulations, while in the other three tasks where ISP was manipulated, participants were exposed to 98 persistent, 101 disrupted, 104 bursting result manipulations. The outcomes of participants' first three query submissions are presented in Figure 7. Each bar represents the percentage of total manipulations in each treatment leading to reformulation on the first SERP, pagination to the second SERP or beyond, or stopping a task on the first SERP. From low to high ISL, the percentage of reformulations decreased and stopping increased. When comparing ISP, bursting appeared to lead to the highest percentage of

reformulations and the lowest percentage of stoppings. The distributions of behaviors for persistent and disrupted were similar. Pagination remained relatively constant across all ISL and ISP treatments.

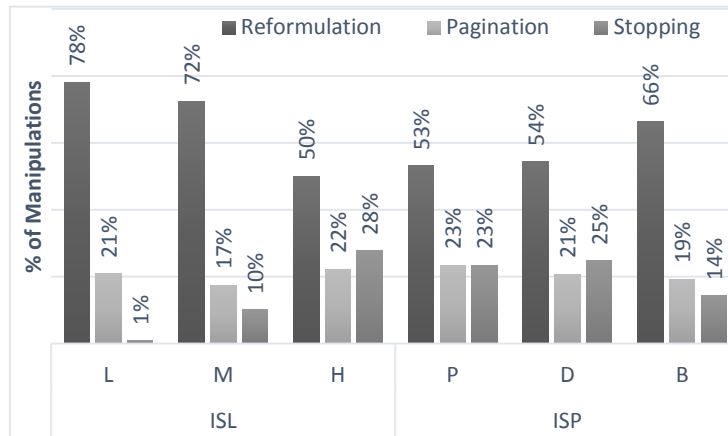


Figure 7. Reformulation, pagination and stopping rate by treatment (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting).

Query abandonment was also examined to understand how ISL and ISP affected participants' reactions to result snippets. Each bar in Figure 8 represents the percentage of total manipulations in each treatment leading to abandonment. When participants were presented with a SERP with low ISL, they chose to leave without examining any document around 42% of the time, while the abandonment rate for high ISL was only 1.6%. The differences among ISP treatments were not as dramatic, but abandonment for the bursting treatment happened 10% more than in the persistent treatment, which is interesting since these treatments had the exact same number of relevant documents. Also interestingly, this abandonment rate was higher than that of medium ISL, which had one less relevant result.

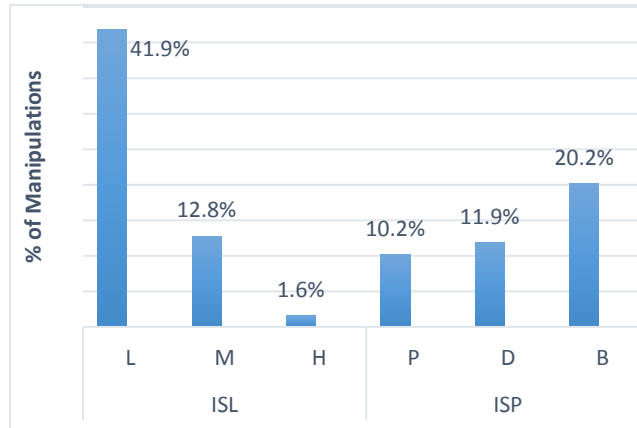


Figure 8. Abandonment by treatment (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting).

Descriptive statistics for each continuous search behavior measure given each ISL and ISP treatment are reported in Tables 13 and 14. It is worth noting that the search behavior measures were mostly correlated with one another. The purpose of presenting them all is an attempt to characterize query stopping in greater detail. From Table 13, one can see as ISL increased from low to high, participants spent more time searching in the search result set, examined more results, went to greater depths both in terms of clicks and mouse hovers, examined more relevant documents, examined more non-relevant documents, and examined more snippets of non-relevant results. For ISP tasks, the persistent ISP led to the greatest amount of interaction in most measures. One can see that when ISP was persistent participants spent the most time evaluating results, examined the most documents, went to the greatest depth, clicked on the most non-relevant results, and examined the most snippets of non-relevant results; however, the bursting ISP led to the greatest pagination frequency and the greatest depth of mouse hover. Both the persistent and bursting ISPs led to the greatest number of saved results (Table 14).

Table 13
Search Behavior Measures (M, SD) by ISL (The Highest Mean for Each Measure is Bolded to Facilitate Comparisons)

Measures	Low	Medium	High
Time	61.85 (63.12)	101.86 (75.35)	128.05 (79.28)
NumPagination	0.48 (1.27)	0.47 (1.38)	.61 (1.58)
NumExamined	0.78 (0.81)	2.02 (1.33)	2.82 (1.37)
DeepestRankClick	2.71 (5.45)	4.10 (4.70)	6.49 (10.03)
DeepestRankHover	9.23 (14.57)	9.85 (16.24)	12.02 (18.51)
NumPred	0.21 (0.47)	0.48 (0.70)	.65 (.78)
NumRele	0.57 (0.55)	1.47 (1.14)	2.13 (1.28)
NumNonRele	1.94 (4.87)	2.11 (4.09)	3.67 (9.22)

Table 14
Search Behavior Measures (M, SD) by ISP (The Highest Mean for Each Measure is Bolded to Facilitate Comparisons)

Measures	Persistent	Disrupted	Bursting
Time	121.17 (102.11)	112.54 (76.39)	97.48 (77.16)
NumPagination	0.47 (1.21)	0.48 (1.45)	0.53 (1.45)
NumExamined	2.35 (1.69)	2.31 (1.62)	1.95 (1.59)
DeepestRankClick	6.33 (5.71)	4.82 (5.21)	5.72 (4.34)
DeepestRankHover	10.74 (13.21)	10.05 (15.86)	11.56 (16.79)
NumPred	0.40 (0.69)	0.37 (0.58)	0.40 (0.62)
NumRele	1.94 (1.61)	1.91 (1.58)	1.49 (1.39)
NumNonRele	3.98 (4.37)	2.51 (4.33)	3.77 (3.26)

Participants' NFC scores ranged from 2.22 to 4.83 on a five-point scale. Participants scored an average of 3.75 ($SD=0.55$) and a median of 3.72 and the distribution of NFC scores was normal (Shapiro-Wilk Test, $p=.454$). Since there is no established norm of NFC, whether a score is high or low was considered relatively. NFC scores were correlated with search behavior measures aggregated at the task level to examine whether there were any relationships. It was found that higher NFC was negatively related to the frequency of pagination ($r=-.33$, $p=.023$, $N=47$).

5.5 Query Stopping

To examine whether ISL, ISP and NFC affected query stopping statistically significantly, Generalized Estimating Equations (GEE) (Vittinghoff, 2012) were applied to model the effect of ISL, ISP and NFC on search behavior measures. GEEs overcome the assumptions of normality and independence, which is suitable for this data set in which the data points were repeated measurements, continuous search behavior measures were skewed, and non-continuous search behaviors did not follow a normal distribution. GEEs were used to conduct linear regression analysis (for continuous measures) and logistic regression analysis (for categorical measures) for repeated measurements. For each search behavior measure, ISL, NFC and their interaction term were entered in one model, and ISP, NFC and their interaction term were entered in another model. It is worth noting that NFC was treated as a continuous variable rather than a categorical variable in these models. Results of GEEs are reported in Tables 15 and 16 and are discussed in detail in the subsequent sections according to research question. Estimates for both models can be found in Appendix K.

The first research question of this study was: what is the relationship between ISL and search stopping behaviors? The results from Table 15 indicate that ISL significantly influenced Time, QueryAction, Abandonment, NumExamined, NumPred, NumRele, NumNonRele and DeepestRankClick prior to query stopping, Follow-up contrasts were conducted to compare whether the differences between any two treatments were significant. The results show that high ISL led to the greatest NumExamined, DeepestRankClick, NumPred, and NumRele, followed by medium and then low (all contrasts: $p < .05$); high ISL led to more NumNonRele than low ISL ($p < .05$), but no significant difference was found between high and medium ISL for NumNonRele. Participants abandoned their queries more often when the ISL was low, followed by medium and

then high (all contrasts: $p < .01$). Since there were interaction effects between ISL and NFC on Time and QueryAction, the main effects of ISL on time and QueryAction were not further analyzed.

Table 15
Results for ISL Treatments (Wald X^2 , significance)

Measures	ISL	NFC	Interaction
Time	62.94****	9.15**	6.01*
NumPagination	0.74	3.99*	2.33
QueryAction	17.69**	0.20	13.17*
Abandonment	46.42****	43.56	0.44
NumExamined	219.26****	0.45	0.12
NumPred	33.90****	0.00	0.61
NumRele	110.85****	0.31	0.24
NumNonRele	6.47*	1.41	0.96
DeepestRankClick	30.33****	1.28	0.89
DeepestRankHover	2.02	4.78*	2.77

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$

Next, since in most tasks participants did not paginate to the second page, only the cases where participants did not paginate were included to examine how ISL influenced the amount of interaction on the first SERPs. One more significant effect of ISL was found: DeepestRankHover ($X^2=20.32$, $p < .0001$). When the first SERP had high ISL ($M=4.68$, $SD=0.20$) participants hovered to lower ranks than medium ($M=4.30$, $SD=0.20$) and low ($M=3.39$, $SD=0.30$) ($L < M$, $L < H$; $p < .0001$).

The second research question addressed the relationship between ISP and search behaviors. Results in Table 16 show ISP had a significant effect on Abandonment, NumRele and NumNonRele. Follow-up contrasts indicated NumRele was significantly greater in the persistent and disrupted treatments than in bursting ($p < .05$), but no difference was found between persistent and disrupted. NumNonRele was significantly higher in persistent and bursting than in disrupted ($p < .05$), but no significant difference was found between persistent and bursting. Results also

revealed that abandonment for persistent and disrupted were significantly lower than bursting ($p<.05$).

Table 16

GEE Results for ISP Treatments (Wald X^2 , significance)

Measures	ISP	NFC	Interaction
Time	3.80	0.50	0.59
NumPagination	0.16	2.24	1.51
QueryAction	2.74	1.81	3.16
Abandonment	6.33*	0.58	3.34
NumExamined	5.52	0.42	1.50
NumPred	0.33	0.50	2.43
NumRele	8.57*	0.83	2.57
NumNonRele	9.98**	0.85	2.18
DeepestRankClick	5.07	0.80	2.54
DeepestRankHover	1.60	2.54	1.78

*Note. * $p<.05$, ** $p<.01$*

Next, the queries that led to paginations were excluded and the search behavior measures within the range of the first page were examined. It was found that there was a significant effect of ISP on DeepestRankClick ($X^2=79.30$, $p<.01$). While in both persistent and bursting ISPs DeepestRankClick were similar ($P=4.51$ and $B=4.61$), when disrupted ISP was encountered, participants did not click beyond rank 2.76 (contrasts: $D<P$, $D<B$, $p<.0001$). There were also significant effects of ISP on DeepestRankHover ($X^2= 13.95$, $p=<.001$) ($P=B=5.45$, $D=4.51$; contrasts: $P=B>D$, $p<.0001$), and NumNonRele ($X^2=172.73$, $p<.0001$) ($P=2.23$, $B=2.73$, $D=0.68$; $P>D$, $B>D$, $p<.05$).

The last research question examined the relationship between NFC and search stopping behaviors. The effects of NFC according to ISL and ISP are reported in Tables 15 and 16. There was a main effect of NFC on Time, NumPagination, and DeepestRankHover in ISL treatments but not in ISP treatments. People with higher NFC scores paginated less and stopped hovering at higher ranks.

An interaction effect between ISL and NFC was found for Time. The relationships are plotted in Figure 9 at NFC=10th, 50th and 90th percentile for better understanding. While overall ISL was related positively to time and NFC was related negatively to time, the effect of ISL on time was moderated by NFC the least when NFC was low. As NFC increased from the 10th percentile to the 90th, the difference in time between ISL became less obvious; in other words, participants with higher NFC did not increase search time as much as participants with lower NFC as ISL shifted from low to high. In addition, when ISL was low, the time spent in a search result set was most similar across NFC scores at different percentiles than ISL was medium or high.

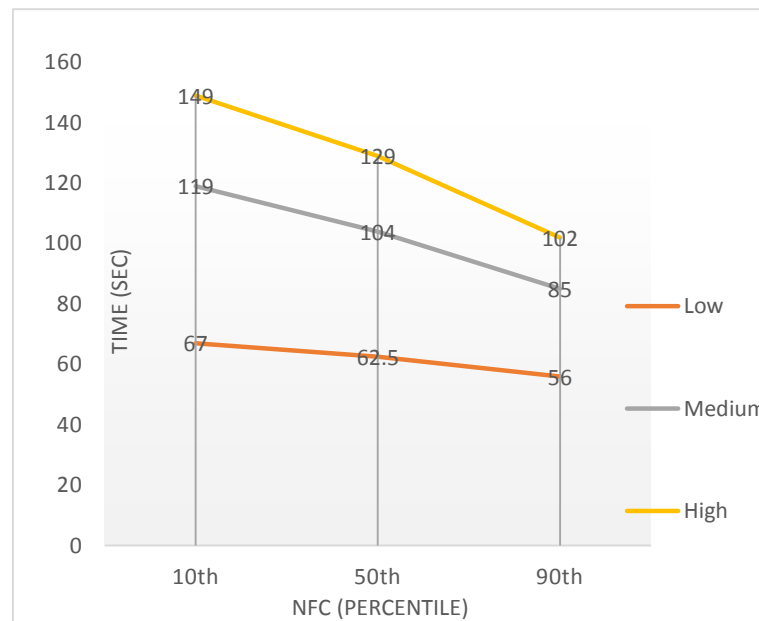


Figure 9. Interaction effect between ISL and NFC on time.

Another interaction effect between ISL and NFC was found on QueryAction ($X^2=17.69$, $p<.01$) and is shown in Figure 10, 11 and 12. An interaction between ISL and NFC on the three outcomes of QueryAction means the relationship between NFC and the predicted probability for each QueryAction outcome depended on ISL, and the nature of the effect of ISL varied with the outcomes. To understand a significant interaction effect in a multi-category logistic regression it is often best to examine results graphically (Azen & Walker, 2011). The predicted probabilities for reformulation, pagination and stopping are plotted in Figures 10, 11, and 12, respectively by NFC with each ISL treatment as a separate line. Figure 10 shows that the relationship between NFC and the probability of reformulation was positive for medium ISL, but negative for low and high ISLs, indicating that higher NFC was related to a higher probabilities of reformulation when participants encountered medium ISL but lower probabilities when they encountered low and high ISLs. The slope is also the steepest for medium, which means the effect of NFC was the strongest for medium ISL; the probability of reformulation varied to a greater extent according to NFC when it was medium ISL than low and high ISLs. The relationship between NFC and pagination was negative and stronger for medium, but positive and weaker for low and high ISLs. Lastly, the relationship between NFC and stopping was weak and negative for medium, but weak and positive for low and high.

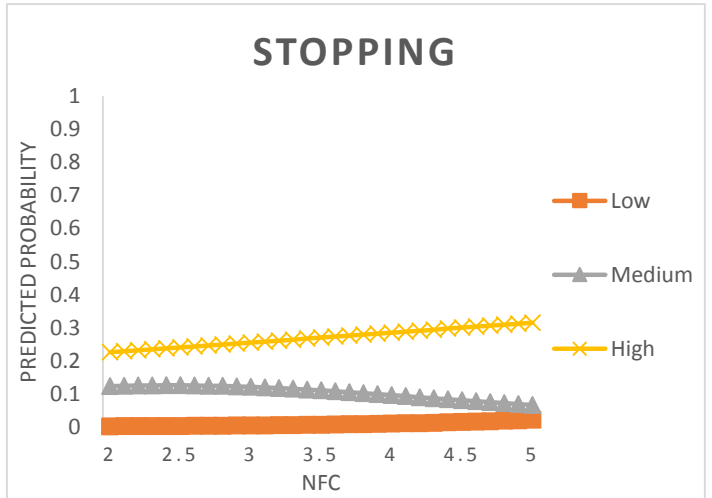
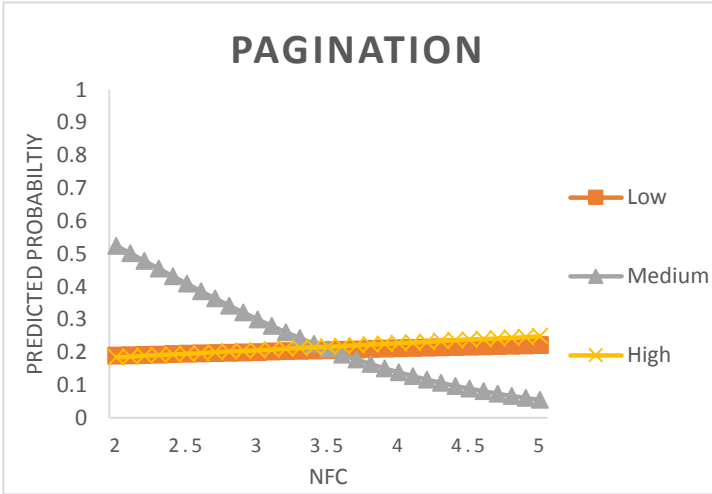
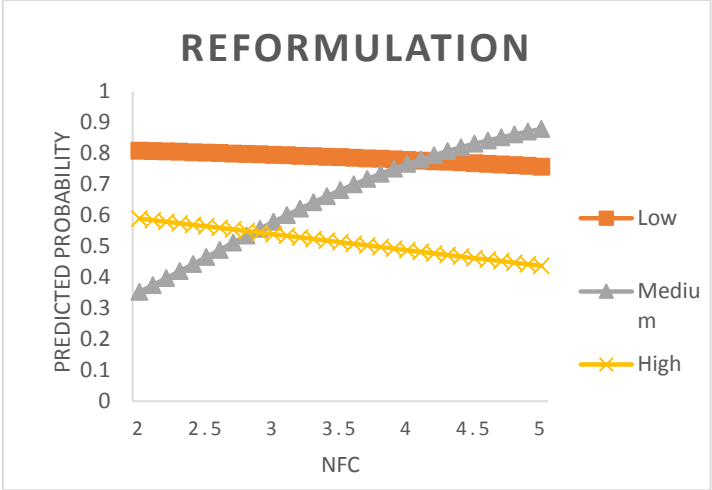


Figure 10, 11 & 12. Predicted probability of reformulation, pagination and stopping by ISL and NFC

5.6 Task Stopping

The last section answered the first three research questions about query stopping. The relationships between ISL and query stopping, between ISP and query stopping, and between NFC and query stopping were examined. This section addresses the first three research questions for task stopping and the last research question.

5.6.1 Effect of task length on task stopping. Since ISL and ISP were within-subject variables and participants experienced varied numbers of ISL or ISP treatments during the search tasks, task stopping was subject to the potential influence of both number of queries submitted in a task and order of treatments experienced in a task. Before the effect of treatment order can be analyzed on task stopping, task-level statistics of the search behavior measures were aggregated by task length, defined as the number of queries submitted in a task. The effect of task length on task stopping was then investigated. A task with more query submissions is described as a longer task than another task with fewer query submissions. Because starting from the fourth query submission participants were exposed to search results from the open Web without control, only tasks with one-query, two-query and three-query submissions are included in the following analyses. Search behavior measures at the task level were computed by adding the values of search behavior measures at the query level. Search behavior measures at the task level showcase the amount of effort expended to search prior to task stopping in an accumulative fashion. Performance on the same search behavior measures for tasks with four or more than four queries can be found in Appendix L.

As shown in Table 17, in ISL tasks, longer tasks generally led to higher amounts of interaction. In three-query tasks, participants took more time searching, paginated more SERPs, examined more results, clicked deeper and hovered deeper on SERPs, found more relevant

documents, examined more non-relevant document, and examined more snippets of non-relevant documents. Significant differences from ANOVA tests were found for Time, NumExamined, NumRele and NumPred for ISL tasks. In other words, participants who submitted one, two, or three queries tended to spend significantly different amounts of time searching, examined significantly different numbers of results, and examined significantly different numbers of relevant results and numbers of non-relevant results.

Post-hoc multiple comparisons were applied to find out between which task lengths variances occurred (Table 17). The results indicate that participants who submitted two queries and three queries spent significantly more time searching than participants who submitted one query; participants who submitted three queries evaluated the most search results, followed by participants who submitted two queries, and last by one query. Similarly, participants who submitted three queries saved significantly more relevant results than those who submitted only one query. Likewise, participants who issued two or three queries examined more non-relevant result snippets than participants who issued just one query in a task. No significant differences were found in NumPagination, DeepestRankClick and DeepestRankHover, which means that task length did not have an impact on search depths prior to task stopping (note: These accumulative task-level measures were computed by summing the values of all query-level search behavior measures prior to query stoppings during a task).

Table 17

Task-Level Search Behavior Measures by Task Length (ISL; the Highest Mean for Each Measure was Bolded for Comparison)

Measures	1-Query (n=30)	2-Query (n=23)	3-Query (n=20)	ANOVA <i>F</i> (2, 70)	Multiple comparisons
Time	200.80	297.88	324.30	7.29 **	1<2; 1<3
NumPagination	1.07	1.35	2.10	0.98	--
NumExamined	3.27	4.91	6.35	21.71 ***	1<2<3
DeepestRankClick	7.23	15.39	16.95	2.61	--
DeepestRankHover	18.50	24.30	36.95	2.16	--
NumRele	2.87	3.74	4.50	9.18 ***	1<3
NumPred	.33	1.13	1.75	11.04 ***	1<2; 1<3
NumNonRele	3.97	10.48	10.75	1.64	--

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$

A different story can be told about ISP tasks. While overall three-query ISL tasks were associated with a higher amount of interaction than one-query and two-query ISL tasks, two-query ISP tasks rather than three-query tasks resulted in the highest values in all measures except NumPred. ANOVA tests further show that task length resulted in significant differences in Time, NumExamined, DeepestRankClick, NumRele, NumPred and NumNonRele (Table 18). Follow-up multiple comparisons indicated that one-query tasks were significantly shorter in time than two- and three-query tasks; more documents were examined, results ranked deeper were clicked, and more relevant documents were found in two-query tasks than one-query tasks. Besides, participants examined more non-relevant snippets in two-query tasks than one-query and three-query tasks. Similar to ISL tasks, it was shown that despite different task lengths, NumPagination and DeepestRankHover did not differ significantly, either.

Table 18

Task-Level Search Behavior Measures by Task Length (ISP; the Highest Mean for Each Measure was Bolded for Comparison)

Measures	1-Query (n=46)	2-Query (n=28)	3-Query (n=21)	ANOVA <i>F</i> (2, 92)	Multiple comparisons
Time	193.34	340.86	323.42	16.61***	1<2; 1<3
NumPagination	1.17	1.54	0.43	1.76	--
NumExamined	4.04	5.68	5.24	6.26**	1<2
DeepestRankClick	9.63	17.29	12.19	9.67***	1<2
DeepestRankHover	19.54	29.07	18.86	1.89	--
NumRele	3.50	4.79	4.10	3.97*	1<2
NumPred	0.53	0.86	1.05	2.84*	--
NumNonRele	5.59	11.61	6.95	9.14***	1<2; 3<2

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$

5.6.2 Effect of task length and treatment order on task stopping. In the previous section, the impact of task length on task stopping in terms of search behavior measures was examined and it was found that while task length had a positive effect on some measures, on other measures no differences were found. One may argue that treatment order might have contributed to such differences rather than search length alone. In this section, the effect of both task length and treatment order on task-level search behaviors are examined. To prevent uncontrolled search results after the fourth query submissions from confounding the interpretations of results, the only tasks included in this analysis are one-query, two-query and three-query tasks.

Search tasks were classified based on task length (1-, 2- or 3-query) and order of treatments experienced (e.g. High->Medium->Low) for ISL and ISP tasks, respectively. Each cell in Table 19 and Table 20 demonstrates the number of tasks in each category of the classification (also in Figures 13-17). As the sizes of some categories are very small, no inferential statistical procedures are applied here. Note that when participants submitted only one query during an entire task, order of treatment refers to the only treatment presented to them.

When comparing among one-query tasks, it can be seen from Table 19 that the highest number of one-query tasks were observed when the high ISL was experienced, while the lowest number of tasks were where participants were shown low ISL SERPs first. This result indicates that participants who encountered more relevant results on the first SERP became satisfied earlier, thus ended tasks earlier than when encountering fewer relevant documents on the first SERP. Among two-query tasks, the highest count of tasks fell in the category of LH and MH; in other words, more participants stopped searching when they experienced either low or medium ISL first followed by the high ISL than other treatment orders. This result demonstrates that encountering a SERP with more relevant results were more likely to lead to the end of a task. Similar trends can be observed in three-query tasks, too. The most frequently observed treatment order was MLH, followed by LMH, both ended with high ISL.

Table 19
Number of Tasks by Task Length and Order of ISL Treatments

Task Length (Number of Queries)	Order of Treatments	Number of Tasks
1	L	3
	M	8
	H	19
2	LM	3
	LH	7
	ML	1
	MH	7
	HL	3
	HM	2
3	LMH	5
	LHM	0
	MLH	9
	MHL	2
	HLM	3
	HML	1

The differences in the number of tasks from one category to another were not as pronounced among one-query ISP tasks. First, it can be seen from Table 20 that regardless of the treatment experienced, the numbers of tasks observed were about the same; in addition, there were many more one-query tasks in ISP tasks than ISL tasks. Whereas in two-query tasks, tasks starting with bursting ISP outnumbered tasks of other treatment orders, which shows that seeing results ranked lower at the start of a task motivated participants to reformulate to a greater extent; moreover, participants ended tasks after seeing bursting ISP less often than any other treatment, which also implied that seeing the bursting ISP often led to reformulations. Among three-query tasks, bursting ISP played its role in encouraging reformulation again when it was presented as the second treatment in a task. Treatment order PBD and DBP together accounted for 62% of the total number of three-query tasks. In other words, regardless of the first treatment participants received, when they experienced a burst of relevant results for their second queries there was a higher likelihood they reformulated again than when the second treatment was otherwise.

Table 20

Number of tasks by task length and order of ISP treatments

Task Length	Order of Treatments	Number of Tasks
1	P	13
	D	16
	B	17
2	PD	4
	PB	2
	DP	4
	DB	2
	BP	8
	BD	8
3	PDB	0
	PBD	6
	DPB	4
	DBP	7
	BPD	2
	BDP	2

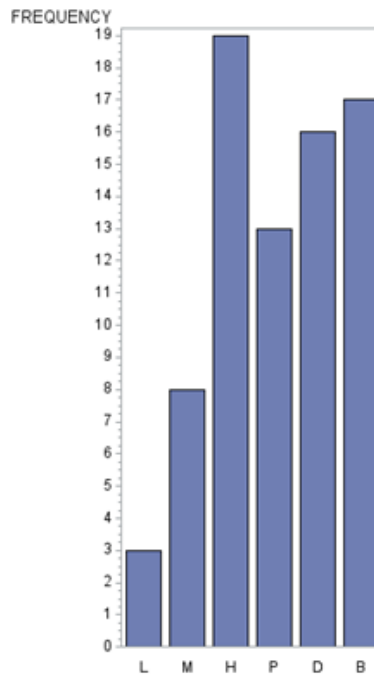


Figure 13. Number of tasks by treatment when task length=1 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting).

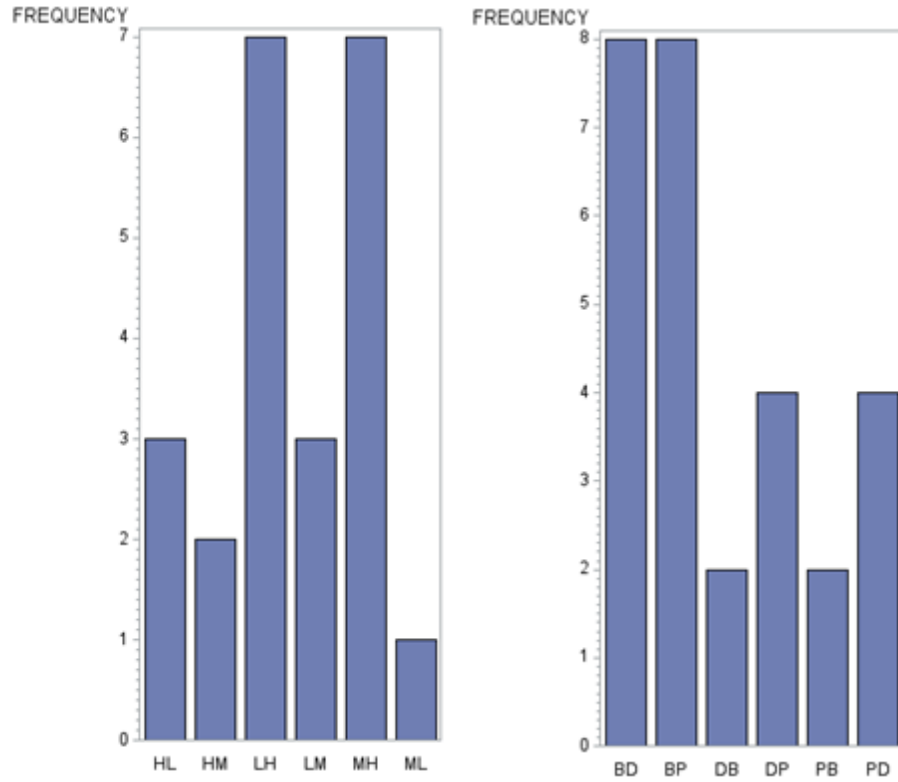


Figure 14 & Figure 15. Number of tasks by ISL treatment order (Left) and by ISP treatment order (Right) when task length=2 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting).

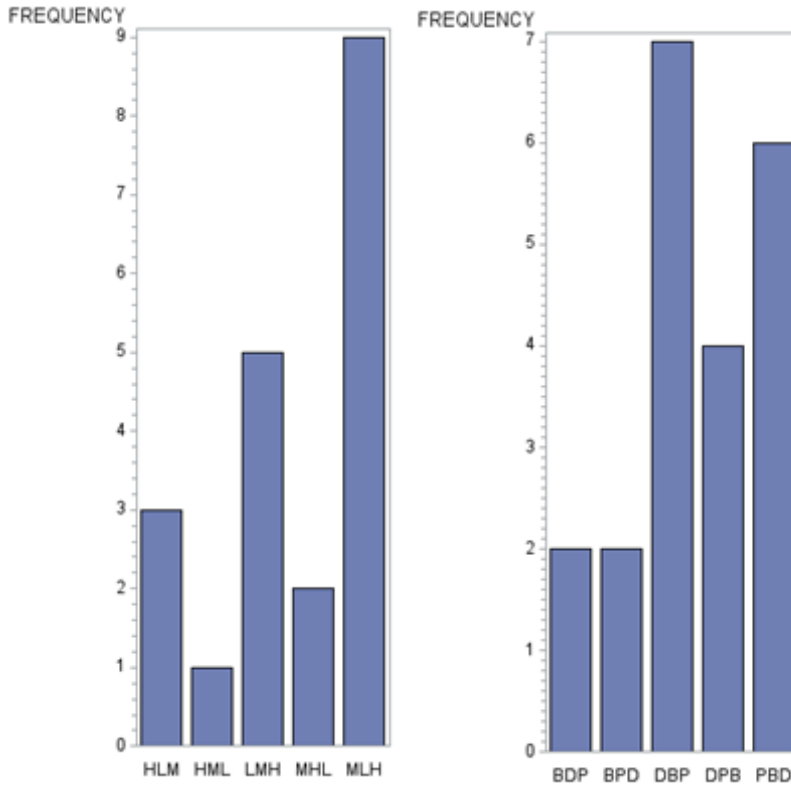


Figure 16 & Figure 17. Number of task by ISL treatment order (Left) and by ISP treatment order (Right) when task length=3 (L=Low; M=Medium; H=High; P=Persistent; D=Disrupted; B=Bursting).

Search behavior measures were then compared by considering two factors: task length and order of treatment in a task. A combination of these two factors resulted in a total of 15 categories for ISL and ISP tasks, respectively, including three categories of one-query tasks, six categories of 2-query tasks and six categories of 3-query tasks. Because the numbers of tasks in some categories are very small, the analysis in this section is descriptive in nature.

Among one-query tasks, the mean of every search behavior measure was the highest in the Medium ISL category except for NumNonRele and DeepestRankClick (Table 21). This means that when participants submitted only one query in a task, those who experienced the medium ISL spent more time, paginated more SERPs, examined more results, hovered lower,

clicked on more non-relevant documents and more relevant documents than those who experienced the low and high ISLs.

Table 21

One-Query ISL Tasks (the Highest Mean for Each Measure is Bolded)

<i>M</i> (<i>SD</i>)	Low (n=3)	Medium (n=8)	High (n=19)
Time (sec)	188 (110)	213 (105)	198 (81)
NumPagination	1 (0)	1.38 (2.39)	0.95 (1.99)
NumExamined	2 (0)	3.63 (1.30)	3.32 (1.34)
DeepestRankClick	14 (3.46)	8.62 (7.76)	5.58 (3.95)
DeepestRankHover	17.67 (2.31)	21.25 (2.77)	17.47 (26.84)
NumRele	1.67 (0.58)	3.00 (1.41)	3.00 (1.25)
NumPred	0.33 (0.57)	0.38 (0.52)	0.32 (0.48)
NumNonRele	12 (3.46)	5 (6.65)	2.26 (3.03)

Performance on the same behavioral measures for ISP tasks where participants submitted only one query can be seen in Table 22. Participants who encountered the persistent ISP spent the most time searching, examined the most documents, examined more non-relevant results and the depths of clicking was the greatest before they ended the tasks. However, the bursting ISP led to the most frequent paginations and the deepest mouse hover. Disrupted ISP resulted in saving the most relevant results even though this did not differ much from the other two ISPs.

Table 22

One-Query ISP Tasks (the Highest Mean for Each Measure is Bolded)

<i>M</i> (<i>SD</i>)	Persistent (n=13)	Disrupted (n=16)	Bursting (n=17)
Time	208 (166.21)	198.27 (62.27)	177.53 (78.97)
NumPagination	0.69 (1.11)	0.94 (2.21)	1.76 (3.05)
NumExamined	4.08 (1.38)	4.06 (1.06)	4.00 (0.94)
DeepestRankClick	11.23 (4.44)	8.50 (5.70)	9.47 (4.20)
DeepestRankHover	14.85 (11.42)	16.00 (23.39)	26.47 (32.31)
NumRele	3.38 (1.56)	3.63 (1.31)	3.47 (0.72)
NumPred	0.69 (1.03)	0.38 (0.62)	0.53 (0.72)
NumNonRele	7.15 (3.80)	4.44 (4.91)	5.47 (3.87)

Among two-query ISL tasks, the order HL led to the most interaction before task stopping, while PB dominated all measures among two-query ISP tasks (Table 23 and Table 24).

Table 23

Two-Query ISL Tasks (the Highest Mean for Each Measure is Bolded)

<i>M</i> (<i>SD</i>)	HL (n=3)	HM (n=2)	LH (n=7)	LM (n=3)	MH (n=7)	ML (n=1)
Time	507.96 (64.28)	267.32 (101.45)	232.34 (84.41)	205.13 (98.15)	316.32 (139.40)	336.77 (NA)
NumPagination	7 (5)	0 (0)	1 (1)	0 (0)	0.29 (0.76)	1 (NA)
NumExamined	8 (2.65)	3.50 (0.71)	4.57 (0.98)	2.67 (1.53)	5.29 (0.76)	5 (NA)
DeepestRankClick	62.67 (56.90)	6.00 (1.41)	9.29 (5.25)	3.67 (0.58)	9.00 (5.34)	15.00 (NA)
DeepestRankHover	84.67 (55.08)	7.00 (2.83)	21.43 (18.30)	8.67 (3.06)	13.00 (9.06)	24 (NA)
NumRele	4 (2.65)	3.00 (0)	3.57 (0.08)	2.67 (1.53)	4.43 (0.79)	4.00 (NA)
NumPred	4 (1)	0.5 (0.71)	0.86 (1.46)	0 (0)	0.86 (0.90)	1 (NA)
NumNonRele	54.67 (54.26)	2.5 (0.71)	4.71 (4.72)	1.00 (1.00)	3.71 (5.12)	10.00 (NA)

Table 24

Two-Query ISP Tasks (the Highest Mean for Each Measure is Bolded)

	BD (n=8)	BP (n=8)	DB (n=2)	DP (n=4)	PB (n=2)	PD (n=4)
Time	335.77 (135.81)	317.89 (164.83)	358.97 (177.34)	278.89 (108.40)	418.87 (17.05)	410.85 (75.14)
NumPagination	1.75 (2.76)	1.63 (1.77)	2.50 (3.54)	0.75 (1.50)	3.50 (0.71)	0.25 (0.50)
NumExamined	4.88 (3.14)	4.88 (2.10)	4.00 (1.41)	7.00 (2.94)	9.5 (0.71)	6.5 (2.38)
DeepestRankClick	15.25 (11.46)	16.00 (8.01)	21.50 (20.50)	16.25 (13.28)	33.00 (0)	15 (7.39)
DeepestRankHover	30.38 (31.09)	30.79 (16.24)	36.5 (41.72)	21.50 (17.67)	52.00 (5.66)	15.50 (7.19)
NumRele	3.88 (2.64)	4.13 (2.10)	3.00 (1.41)	6.25 (2.87)	7.50 (0.71)	6.00 (2.31)
NumPred	1.00 (0.93)	0.63 (0.74)	1.00 (0)	0.75 (0.96)	2.00 (1.41)	0.50 (0.58)
NumNonRele	10.38 (8.80)	11.13 (6.08)	17.5 (19.09)	9.25 (10.59)	23.50 (0.71)	8.5 (5.80)

For three-query ISL tasks HML resulted in the most interaction prior to task stopping, while for ISP tasks, BPD resulted in the most interaction (Table 25 and Table 26).

Table 25
Three-Query ISL Tasks (the Highest Mean for Each Measure is Bolded)

<i>M</i> (<i>SD</i>)	HLM (n=3)	HML (n=1)	LHM (n=0)	LMH (n=5)	MHL (n=2)	MLH (n=9)
Time	390.94 (324.34)	313.03 (NA)	NA	281.21 (135.91)	268.42 (0.39)	339.70 (125.65)
NumPagination	2.67 (4.62)	6.00 (NA)	NA	0.80 (1.30)	3.50 (4.95)	1.89 (3.02)
NumExamined	5.33 (2.08)	10.00 (NA)	NA	5.4 (0.89)	7.00 (0)	6.67 (1.80)
DeepestRankClick	10.33 (8.08)	40.00 (NA)	NA	15.08 (10.28)	28.50 (27.58)	14.67 (12.23)
DeepestRankHover	41.67 (52.54)	85.00 (NA)	NA	24.80 (18.89)	54.00 (57.98)	33.00 (35.99)
NumRele	4.33 (1.53)	9.00 (NA)	NA	4.00 (0.71)	4.00 (0)	4.44 (1.24)
NumPred	1.00 (1.00)	1.00 (NA)	NA	1.40 (0.55)	3.00 (0)	2.00 (1.22)
NumNonRele	5.00 (7.00)	30.00 (NA)	NA	10.40 (9.91)	21.50 (27.58)	8.33 (10.79)

Table 26

Three-Query ISP Tasks (the Highest Mean for Each Measure is Bolded)

<i>M</i> (<i>SD</i>)	BDP (n=2)	BPD (n=2)	DBP (n=7)	DPB (n=4)	PDB (n=6)	PDB (n=0)
Time	319.02 (40.21)	443.04 (43.49)	297.06 (138.41)	320.99 (133.16)	317.37 (168.13)	NA
NumPagination	0 (0)	1.5 (2.12)	0.71 (1.89)	0 (0)	0.17 (0.41)	NA
NumExamined	4.50 (2.12)	8.00 (4.24)	5.43 (3.46)	4.50 (1.00)	4.83 (1.47)	NA
DeepestRankClick	10.00 (2.83)	18.50 (13.44)	13.71 (6.44)	10.25 (3.20)	10.33 (4.59)	NA
DeepestRankHover	14.50 (0.71)	29.50 (27.58)	22.14 (22.37)	13.00 (1.41)	16.83 (3.54)	NA
NumRele	3.50 (2.12)	6.50 (3.54)	4.14 (3.02)	3.00 (1.41)	4.17 (1.33)	NA
NumPred	1.00 (0)	1.50 (0.71)	1.29 (1.89)	1.00 (0.82)	0.67 (0.52)	NA
NumNonRele	5.50 (0.71)	10.50 (9.19)	8.29 (5.25)	5.75 (2.22)	5.50 (3.39)	NA

5.6.3 Effect of first treatment and NFC on task stopping. The comparisons above describe how task length and order of treatments together affected task stopping. However, due to low counts in many categories, conclusions cannot be drawn. In order to examine the impact of order of treatment on task stopping, all search tasks (N=282) were first collapsed by the first treatment presented at the start of each search task. The mean and standard deviation for each measure by the first ISL and ISP treatment can be seen in Table 27. The first ISL treatment in a task, NFC and their interaction term were entered into a GEE model, while the first ISP treatment in a task, NFC, and their interaction term were entered into another model. GEE results show there was a significant effect of NFC on NumQuery ($X^2=4.17$, $p<.05$); regardless of the first ISL treatment in a task, the higher the NFC, the more queries participants issued.

Table 27

Search Behavior Measures at the Task Level by the First Treatment in a Task

First Treatment	Low (n=47)	Medium (n=47)	High (n=47)	Persistent (n=47)	Disrupted (n=47)	Bursting (n=47)
Time	348.62 (199.55)	349.89 (153.18)	323.9339 (176.96)	359.07 (204.03)	318.16 (177.54)	298.02 (164.72)
NumQuery	4.25 (2.53)	4.11 (3.53)	3.45 (3.01)	3.55 (2.28)	3.00 (2.25)	2.43 (1.77)
NumPagination	1.26 (1.88)	1.98 (4.63)	1.74 (2.97)	0.91 (1.46)	1.09 (2.13)	1.83 (3.50)
NumExamined	7.17 (3.74)	7.02 (3.82)	6.72 (4.45)	6.87 (4.05)	6.53 (4.20)	5.87 (3.51)
DeepestRankClick	18.53 (12.40)	16.94 (14.19)	18.89 (22.68)	17.53 (10.77)	17.00 (13.67)	14.77 (8.74)
DeepestRankHover	33.77 (26.58)	39.38 (51.18)	36.66 (37.98)	28.21 (20.67)	27.60 (27.03)	33.68 (38.85)
NumPred	2.15 (2.40)	1.96 (2.05)	2.04 (2.94)	1.57 (1.83)	1.72 (2.60)	1.09 (1.49)
NumRele	4.81 (2.13)	4.96 (2.72)	4.62 (2.95)	5.21 (2.65)	4.64 (2.71)	4.68 (3.02)
NumNonRele	11.36 (9.90)	9.98 (11.84)	12.17 (19.91)	10.66 (7.32)	10.47 (10.75)	8.89 (6.23)

Interaction effects were also found for NumPred ($X^2=7.84$, $p<.05$), NumNonRele ($X^2=7.00$, $p<.05$) and DeepestRankClick ($X^2=7.18$, $p<.05$). The means of these search behavior measures for each treatment under NFC=10th, 50th and 90th percentile are plotted in Figures 18, 19 and 20 to show the relationships. Figures 18 and 20 show that when the first SERP encountered in a search task displayed low ISL, as NFC increased, participants' depth of click and number of non-relevant snippets they examined did not change. Yet when the first SERP displayed high ISL, as NFC increased, the depth of click and the number of non-relevant snippets examined increased; when ISL was medium, as NFC increased, the depth of click and the number of non-relevant snippets examined decreased. Figure 19 indicates that overall, as NFC increased, participants clicked on fewer non-relevant documents except when NFC was low and ISL was high, participants clicked on the fewest non-relevant documents. An examination of the tasks starting with high ISL conducted by participants who scored the 10th percentile on NFC shows that they

only issued one query during the tasks, which was probably the reason why they clicked on the fewest number of non-relevant documents.

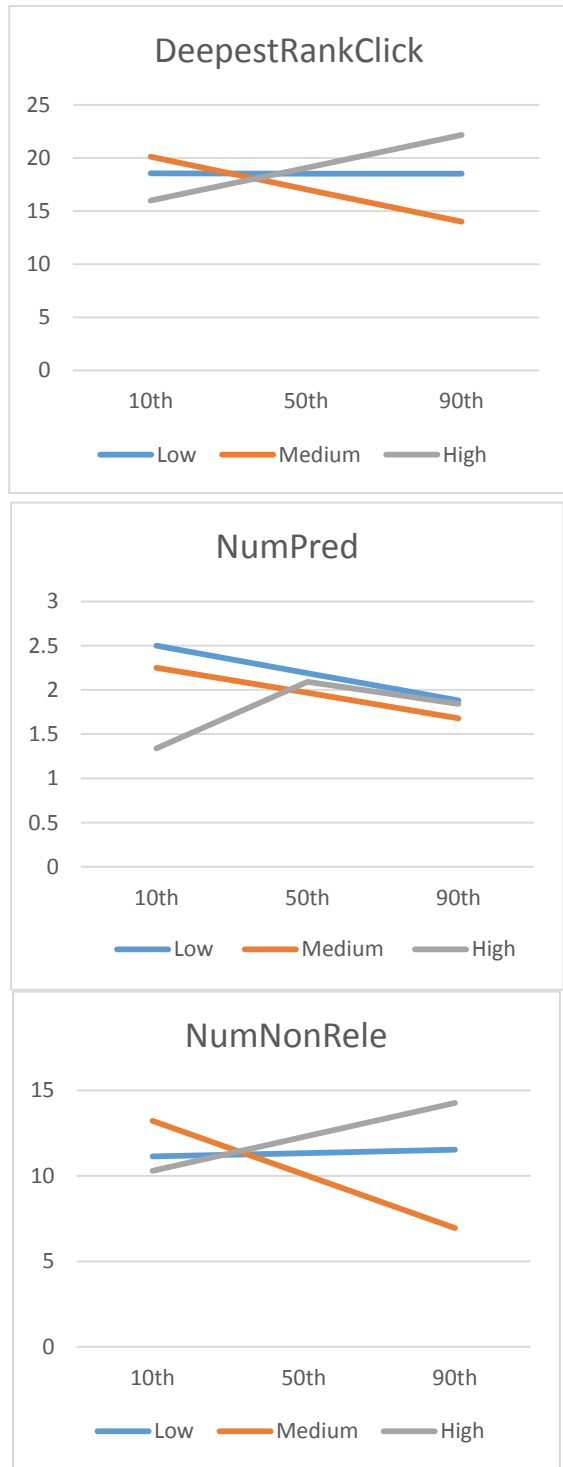


Figure 18, 19 & 20. Interaction effect between the first ISL treatment and NFC on DeepestRankClick (Top), NumPred (Middle), and NumNonRele (Bottom).

GEE results for ISP tasks and NFC show that ISP had an impact on NumQuery, albeit not significant ($X^2=5.84, p=.054$). When ISP was persistent at the beginning of a task, participants submitted the most queries, while the fewest queries were issued when the treatment was bursting. There was also an interaction effect between ISP and NFC on NumPagination ($X^2=7.07, p<.05$) and DeepestRankHover ($X^2=6.38, p<.05$). The relationships are shown in Figure 21 and Figure 22. Figure 22 shows that the depth of mouse hover aggregated at the task level was greatest when the ISP was bursting, followed by disrupted and persistent. However, as NFC increased, the three lines converged, indicating that the difference was relatively small for higher NFC participants. Similar trends can be observed in Figure 21 except that when NFC was high, participants paginated more when ISP was persistent rather than bursting.

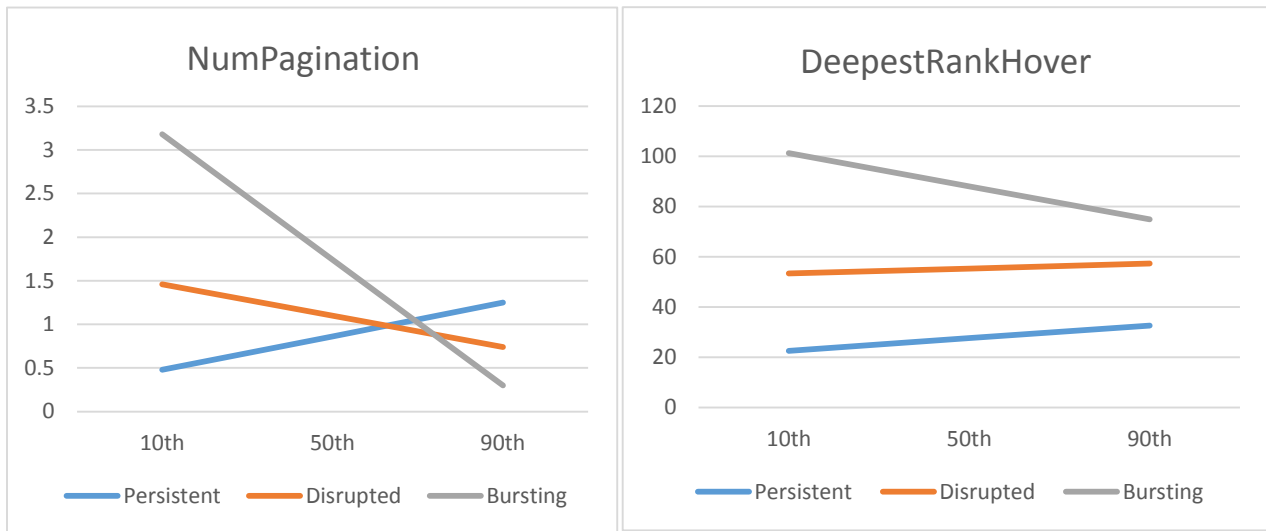


Figure 21 & Figure 22. Interaction effects between first ISP treatment and NFC on NumPagination (Left) and DeepestRankHover (Right)

5.6.4 The last treatment prior to task stopping. In section 5.6.2 where the effects of task length and order of treatment were described, it was shown that there was a trend that ISL tasks more often ended immediately after the high ISL, and ISP tasks were less likely to end immediately after the bursting ISP. In this section, the question of whether any specific treatment preceded task

stoppings more often than others is investigated. Among the 73 ISL tasks where the tasks were one, two or three queries long, there were 10 tasks where low ISL preceded the task stoppings, 16 tasks where medium ISL preceded the task stoppings and 47 tasks where high ISL preceded the task stoppings (Table 28). The relationship between ISL and task stopping was significant ($X^2=32.41, p<.0001$); more tasks ended immediately after a high ISL than low or medium ISL ($p<.0001$). However, no ISP treatment preceded task stopping more often than others ($X^2=2.168, p=.338$).

Table 28

Number of Tasks by the Last ISL Treatment in a Task and Task Length

		Task Length			Total Number of Tasks
		1	2	3	
The Last ISL Treatment in a Task	L	3	4	3	13
	M	8	5	3	16
	H	19	14	14	47

5.6.5 Relationships among pre-task expectations, search behavior measures and post-task evaluations. Pre-Task and Post-Task Questionnaire items were analyzed to better understand whether participants' pre-task expectations and post-task evaluations were associated with search stopping behaviors measures. Pearson's correlation test indicated that one's level of interest in a topic was positively correlated with the time spent on the task ($r=.151, p=0.01$). Furthermore, participants who had more experience searching the topics prior to starting the study, issued fewer queries ($r = -0.188, p=.002$). The more one knew about the topic before search was also correlated with fewer query submissions ($r = -.12, p= 0.043$). Pre-task expectation items were related to post-task evaluation items as well. The more difficult a participant felt about a task before searching, the more difficulty they reported after searching ($r=.464, p<.001$), and the more difficulty determining when to stop searching ($r=.406, p<.001$). Also, the more difficult a task was before

searching, the less success participants evaluated their own search ability when completing the task ($r = -.286, p < .001$) as well as the system's ability to find relevant information ($r = -.159, p < .001$). Search behaviors prior to task stopping were also associated with post-task evaluations. The more SERPs paginated prior to task stopping, the easier it was to decide whether information was enough to stop ($r = -.136, p = .022$) and the more successful participants felt they were at solving the task ($r = .144, p = .015$). The deeper the mouse hover, the easier participants felt the tasks were after searching ($r = -.123, p = .039$), the easier it was to decide whether information was enough ($r = -.164, p = .006$), and the more successful participants felt about themselves at solving the problem ($r = .163, p = .005$).

5.6.6 Predicting post-task evaluations. The previous section demonstrated the overall relationship among pre-task expectation, search behaviors and post-task experience. In this section, the Pre-Task Questionnaire items, search behavior measures (except NumNonRele as it was highly correlated with DeepestClickRank), NFC, and the first and the last treatment in a task were included in a model using GEE for the ISL tasks (Table 29) and another model for the ISP tasks (Table 30), respectively, to examine the importance of each variable in explaining post-task evaluations. Note that the first treatment and the last treatment were entered as categorical variables in the models. Results are presented by Post-Task Questionnaire items as below:

Post-Task Difficulty: Participants' ratings of task difficulty before searching and the first treatment encountered during a task significantly predicted task difficulty after searching in both ISL and ISP tasks. The higher the Pre-Task Difficulty, the higher the Post-Task difficulty. When the first treatment was medium or high ISL, Post-Task Difficulty was significantly lower than when the first treatment was low ISL. When the first treatment was bursting, participants felt the tasks were significantly more difficult than when the first treatment was persistent. Moreover, NFC

was shown to have a significant effect on Post-Task Difficulty in ISP tasks; the higher the NFC, the more difficult participants felt about the task after search.

Difficulty_Enough: For both ISL and ISP tasks, task difficulty before searching significantly predicted how difficult it was to determine when they had enough information to stop. For ISP tasks, when the first treatment was persistent, participants felt it was less difficult to decide when to stop than when it was bursting or disrupted.

Success_Self: For both ISL and ISP tasks, participants' task difficulty ratings before searching were negatively correlated with their evaluations of their own search ability during the search. The more difficult participants rated a task before searching, the less successful they felt subsequently about their abilities to find relevant information. The first treatment was also a significant predictor of participants' responses to this question in ISL tasks. When the first treatment was medium or high, participants evaluated their abilities more positively. ISP did not impact participants' responses to this question. Instead, for ISP tasks, the accumulative depth of mouse hover was predictive of participants' evaluations. The greater the depth of mouse hover and the more search experience participants had with a topic, the more positively participants evaluated their abilities.

Success_System: For ISL tasks, participants' ratings of task difficulty before searching and the number of queries they issued prior to task stopping significantly predicted their evaluations of the search system. Participants' task difficulty ratings before searching were negatively related to the scores they gave to the system; while the number of queries submitted was positively related to the system rating. For ISP tasks, participants' interest in the topic and previous search experience were both positively correlated with participants' system evaluations.

Table 29

Significant Predictors of Post-Task Evaluations in ISL Tasks (“+” Indicates Positive Relationship Between a Continuous Predictor and a Criterion; “-” Indicates Negative Relationship Between a Continuous Predictor and a Criterion)

Predictors	Criteria			
	Post-Task Difficulty	Difficulty_Enough	Success_Self	Success_System
Pre-Task Difficulty	Z=9.40 *****(+)	Z=7.55*****(+)	Z= -6.33*****(-)	Z= -4.48*****(-)
NumQuery	--	--	--	Z=2.16* (+)
First Treatment	M<L : Z= -2.57*; H<L: Z= -2.30*	--	M>L: Z=2.04*; H>L: Z=3.79****	--

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$

Table 30

Significant Predictors of Post-Task Evaluations in ISP Tasks (“+” Indicates Positive Relationship Between a Continuous Predictor and a Criterion; “-” Indicates Negative Relationship Between a Continuous Predictor and a Criterion)

Predictors	Criteria			
	Post-Task Difficulty	Difficulty_Enough	Success_Self	Success_System
Interest	--	--	--	Z=2.37* (+)
Search Experience	--	--	--	Z=2.02* (+)
Pre-Task Difficulty	Z=6.51*****(+)	Z=4.21*****(+)	Z= -2.16*****(-)	--
DeepestRankHover	--	--	Z=2.20* (+)	--
First Treatment	B>P; Z=2.13*	B>P: Z=2.07*	--	--
NFC	Z=3.08** (+)	--	--	--

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$

5.6.7 Search stopping behavior patterns prior to task stopping. A search stopping behavior pattern is defined as the sequence of moves performed by a searcher prior to task stopping. Maximal Repeating Pattern, a program developed by Siochi and Ehrich (1991) to help researchers identify behavioral patterns with interfaces, is employed in the dissertation to identify patterns prior to task ends. A repeating pattern (RP) is a sequence of actions that occurs more than once in the dataset, and a maximal repeating pattern (MRP) is the longest repeating pattern. For instance, if search moves are represented as A, B, and C and a full set of actions is represented as “ABACABA”, the RPs in this data set are “AB” and “ABA”, and the MRP is “ABA”. From this

example it can be seen that a short repeating pattern can be part of a longer repeating pattern. Take the present study as another example. “click document, rate document and stop searching” was a frequently observed RP, which could be part of two other long yet less frequently occurring RPs: “click document, rate document, click document, rate document and stop searching” and “issue a query, click document, rate document and stop searching”. Shorter repeating patterns occur more frequently than longer repeating patterns since they are the components of longer repeating patterns. In other words, there is a tradeoff between the length of a RP and the frequency of its occurrence.

To present RPs observed in the dataset in an organized manner, RPs are divided into three categories according to “number of transitions.” Transitions are search moves that take a participant away from a current SERP; both reformulations and paginations are regarded as transitions. The three categories of RPs are one-SERP RPs, which include search moves on the last SERP prior to task stopping (RPs without any single pagination or reformulation), two-SERP RPs, in which the search moves include either two query submissions or span across two SERPs prior to task stopping (RPs with one pagination or one reformulation), and three- or more-SERP RPs, where the search moves span across three or more SERPs (RPs with two or more paginations, two or more reformulations, or a combined total of three or more reformulation and paginations). For each category, the three most common and non-overlapping RPs are presented here; the three most common RPs of the same length were independent of one another and could not occur in a unique task at the same time. However, a one-SERP RP can be part of a two-SERP or three-SERP RP. The moves of interest in the analysis include: clicking on a document (SERPClick), judging a document’s relevance (DocJudge), reformulating a new query (Reformulation), clicking on the next page (Pagination), and clicking on “Done” (TaskEnd).

One-SERP RPs: One-SERP RPs are the most common and also the shortest RPs observed from the study. Three one-SERP RPs are shown in Table 31 to represent the most frequent moves participants took on the last SERP prior to task stopping. Prior to task ends, participants most often clicked and reviewed one document on the last SERP (#1). The second most common pattern was clicking and rating three documents on the last SERP before task stopping (#2). Clicking and rating two documents on the last SERP was the third most common RP among one-SERP RPs (#3).

Table 31
One-SERP RPs

Search Moves	One-SERP RPs		
	#1	#2	#3
1 st	SERPClick	SERPClick	SERPClick
2 nd	DocJudge	DocJudge	DocJudge
3 rd	TaskEnd	SERPClick	SERPClick
4 th		DocJudge	DocJudge
5 th		SERPClick	TaskEnd
6 th		DocJudge	
7 th		TaskEnd	
Frequency	91	89	55
Percentage of Total Number of Tasks (N=282)	32%	32%	20%

Two-SERP RPs: Two-SERP RPs were associated with longer behavioral patterns which involved one transition before entering the last SERP. As shown in Table 32, the most frequently seen two-SERP RP began with a query reformulation, followed by the examination and judgment of a single document before task stopping (#4). The second most frequent RP included one pagination, then a click and a rating, (#5) and the other equally prevalent RP started with one reformulation, followed by clicking two documents and rating (#6). As previously mentioned, a

one-SERP RP can be part of longer RPs, #1 belonged to #4 and #5. The reason why the frequency from #4 and #5 do not add up to the frequency of #1 is because one participant issued a query and examined only one document before clicking “Done.”

Table 32

Two-SERP RPs

Search Moves	Two-SERP RPs		
	#4	#5	#6
1 st	Reformulation	Pagination	Reformulation
2 nd	SERPClick	SERPClick	SERPClick
3 rd	DocJudge	DocJudge	DocJudge
4 th	TaskEnd	TaskEnd	SERPClick
5 th			DocJudge
6 th			TaskEnd
Frequency	56	34	34
Percentage of Total Number of Tasks (N=282)	20%	12%	12%

Three- or more-SERP RPs: RPs spanning across three or more SERPs occurred much less often than the other two categories. The most frequent RP included two consecutive paginations without any document clicking prior to stopping (#7) (Table 33). A much longer yet slightly infrequent RP started with a click and a rating, followed by a reformulation to a new SERP and a single document click and rating, and then one more reformulation to another SERP on which a single document click and rating were made before task stopping (#8). Three consecutive paginations without any document clicking and rating was the third most often observed RP in this category (#9). These RPs show that both consecutive paginations without document clicking and consecutive reformulations with sparse clicking predicted task stoppings; these behavioral patterns probably signaled a lack of relevant results.

Table 33
Three- or More-SERP RPs

Search moves	Three-SERP RPs		
	#7	#8	#9
1 st	Pagination	SerpClick	Pagination
2 nd	Pagination	DocJudge	Pagination
3 rd	TaskEnd	Reformulation	Pagination
4 th		SerpClick	TaskEnd
5 th		DocJudge	
6 th		Reformulation	
7 th		SerpClick	
8 th		DocJudge	
9 th		TaskEnd	
Frequency	16	13	12
Percentage of Total Number of Tasks (N=282)	6%	5%	4%

Now that the most common RPs of length three are described, still two questions remain: Did these RPs represent the behavioral patterns of the majority of participants? How often did participants exhibit a RP during the experiment? If a RP was only observed in a few participants in many tasks, then the RP could only represent the behavior of a minority of participants. To understand who contributed to these RPs, Figure 23 is plotted where each color represents a unique participant and the size of each color block indicates the number of times a participant repeated a RP. From left to right are the three most common one-SERP RPs, two-SERP RPs and three- or more-SERP RPs. Table 34 demonstrates the number of unique participants exhibiting each RP and the average number of tasks each unique participant repeated a RP during the experiment.

From both Figure 23 and Table 34 one can see that the shortest RPs happened the most often and were contributed by almost all participants, which means that almost everyone shared the same sequence of moves on the last SERP prior to quitting a task at least once during the

entire study; in other words, one-SERP RPs were highly representative of the search stopping behavior patterns prior to task stopping. To the contrary, the longest RPs occurred the least frequently and the fewest number of unique participants exhibited these RPs. The frequency of two-SERP RPs was between one-SERP and three-SERP RPs and about half the participants exhibited each two-SERP RP. Participants on average repeated the RPs from 1.44 to 2.61 tasks in this study. In other words, during the entire experiment where each participant carried out six search tasks, 1.44 to 2.61 of search tasks within each participant exhibited the same search stopping behavior pattern.

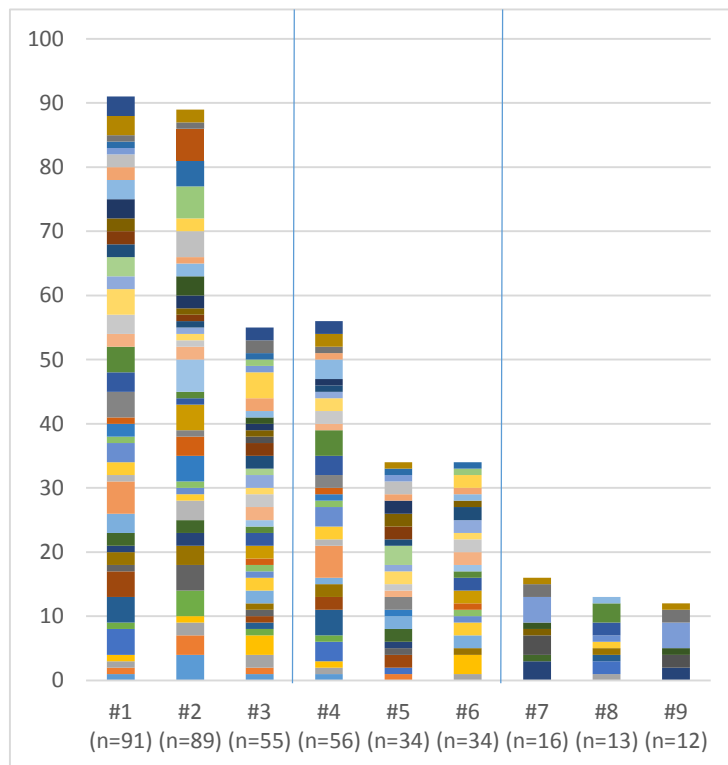


Figure 23. RP frequency by unique participant IDs (Each color represents a unique participant).

Table 34

Unique number of participants exhibiting each RP and average number of tasks exhibiting a RP by each unique participant

Length of RP	One-SERP RPs			Two-SERP RPs			Three- or More-SERP RPs		
RP#	#1 (n=91)	#2 (n=89)	#3 (n=55)	#4 (n=56)	#5 (n=34)	#6 (n=34)	#7 (n=16)	#8 (n=13)	#9 (n=12)
Unique # of Participants	40	38	37	30	23	23	8	9	6
Average # of Tasks	2.28	2.61	1.49	1.87	1.48	1.48	2	1.44	2

5.7 Query and Task Stopping Strategies

The findings from the interviews are organized as the following: First, factors influencing participants’ query-stopping decisions are reported; following that, factors affecting task stopping decisions are discussed. Thirdly, situations where participants mentioned they gathered information by pagination in their past experiences are reported. Lastly, participants’ search styles are summarized to close this section.

5.7.1 Query stopping strategies. Four factors were found to influence when participants decided to stop evaluating search results retrieved by a query: properties of the search results, properties of the queries, properties of the search tasks and properties of the person.

5.7.1.1 Properties of the search results. Participants relied heavily on the first SERP for determining their next moves. The first pages enabled evaluations as to whether a search was “off track” or “on the right track”. In other words, the impression of the first page provided an estimation of the quality of one’s query term, which was then used to infer the quality of results retrieved by it. Seeing many bad results on the first SERP often caused participants to abandon their queries right away. For instance, Participant #4 mentioned: “They were off track like the

tanning one pulled up information about the skin of animals. I knew I was off track so that rather than going to the second page I made more searches”. Below is another example:

A lot of these are joke sites about hilarious permanent tattoos. I did not like the direction it is going like these kind of things. It seemed to be going to the funny. I am like forget it, let's try something else. (P5)

One participant used the word “gestalt” to illustrate the overall impression gleaned from the first SERP. Whether to stop evaluating results or not was primarily based on the overall appropriateness of the first page of results with respect to what one was searching for:

I guess my rationale was if the results on the first page are not relevant I am probably not hitting the nail on the head. It is overall, gestalt, whether they seem to fit what I am looking for to judge the quality of the first page. (P43)

While some participants made their query stopping decisions based on their general and abstract perceptions of the first SERP, others articulated a belief that the more relevant results on the first SERP, the more likely it was to find more relevant results in subsequent SERPs.

I only had three [useful results on the first page], and the rest were about air pollution. I thought if I put in a bunch of new words, I can manipulate it to give me some different results... If the first page gives me a lot, I will give the second page a chance, but none of the pages gave me a lot. (P18)

In addition, almost every participant made an argument that the most related search results were on the first page, so if the first page, the supposedly best page, did not seem good enough, the likelihood of retrieving good results was rare, then there was no point going deeper.

For example, one participant said: “The page I was at already does not have much information, so keep going the chance of finding more information is less (P31).”

The proportion of relevant results in relation to the proportion of non-relevant results on the first SERP appeared to be another way to measure the quality of the first page and whether to issue another query.

Here there are 3 relevant, and 3 totally irrelevant ones here, these are relevant and these are the irrelevant block [point to the screen], then my phrase is turning up as much bad stuff as relevant stuff, so I am willing to look at more stuff. (P18)

The observation that the amount of useful information on the first SERP, be it measured by number, by proportion, or justified by general impression, affected query stopping supports the hypothesis that the information scent level on the first SERP affects when searchers decided to issue a new query.

Moreover, the relative locations of relevant search results to non-relevant results on a page or the distribution of relevant results or non-relevant on the first page was factored into the decision process by some as well. The unusual ranking of the manipulated results pushed some participants to search more.

If there are good ones mixed with irrelevant ones, you can like go on to the second page and find something good, but if it is like good ones and all like junky ones, you might as well move on to the next thing (P27).

The comment also demonstrated that when several non-relevant results followed relevant results, participants expected seeing more non-relevant results if they kept going; therefore, rather than keep searching they preferred to reformulate on the first SERPs. Such phenomenon

was further elaborated by P17: "As I went further down, it starts to get into leather tanning products. Because I was seeing multiple links like that, it made me think it is probably less likely to find relevance beyond that."

When one participant was asked what he or she would do if the relevant results were evenly dispersed on the first SERP, the participant replied:

If there are relevant pages on the 1st, 3rd, 5th, 7th, I would definitely click on it [the second page]...because then why wouldn't I think that the 9th, the 11th and 13th over the next page be relevant, if it is in that order? (P21)

When results that were relevant were mixed with irrelevant, the participant's belief in the search engine ranking was undermined so it was no longer applicable to judge what results to view based on the ranks of results. Hence a new hypothesis, the ranking on the first SERP might continue onto the second, probably caused participants to delay query reformulation.

Query stopping could also take place when non-relevant search results appeared in a block. Collocating non-relevant search results together made individual non-relevant results even more salient. P22 described how such a result alignment repelled him/her from examining more results: "Something about the five articles, they may be different, but they were all social [stock media marketing], social [medial stock], social [media stock], the fact they are constructed together just made me think my search terms were not on the head."

These above comments regarding how the distribution of relevant and non-relevant results influenced subsequent query stoppings demonstrate that information scent pattern was critical to when participants decided to move on to the next query.

Even though participants mostly acknowledged that relevant results should be ranked as early as possible, when they believed what they were looking for by nature might have relevant results ranked lower or on subsequent SERPs, they were more motivated to review more results before they abandoned their current queries. Participants seemed to have developed an understanding of what type of results were more likely to be ranked higher and what others required more searching based on their interpretation of the search task and what types of results might interest others. P27 explained why he/she delayed query stopping:

Because the first page wasn't helping me, I was just going to see if there is anything on the second page. I feel someone could be interested in seeing funny pics so a lot of people clicked on it. (P27)

A similar experience was reported by participants #29: "I know that extreme water sports would not be the most searched thing on the internet, so I think I need to look for my own information."

Sighting seemingly commercial search results sometimes also persuaded participants to search deeper. They believed the results were ranked higher because of payment and they should skip them (P5). However, others viewed the existence of commercial results a sign of bad query terms. Participant #3 said with confidence, "When you start to see eBay you know you are in trouble".

While some participants argued that sifting through search results was laborious, others preferred to stick to reviewing results because they felt "An extra click is a very small investment for finding something very good (P3)." Several other participants even exhausted all ten SERPs to make sure there was nothing important missed. One participant maintained that scrolling down

the SERP was so easy that he did not even think he was searching for information, rather, he was letting the computer have the full control, and called such search behavior “lazy search” (P44).

5.7.1.2 Properties of the queries. Many participants attributed their query stopping decisions to the nature of their own query terms. While most reformulations took place early on the first SERPs, they occurred later for reasons such as lack of better terms or lack of confidence in the current query terms. Participant #13 continued because, “I wasn't sure if I could come up with other terms. I could not immediately think of something better,” while Participant #3 explained her insecurity as the following:

Maybe I am insecure about the terms I use and I want to make sure I exhaust these terms before I change them. If I change before I look at several pages, maybe there is something there. If I change too quickly I am missing out. (P3)

In addition, P3's comment shows that delaying query reformulation not only was a strategy for looking for useful information but also for looking for alternative query terms.

On the other hand, a strong confidence in one's own query terms also prevented participants from reformulating immediately. Participants believed the search results were worth exploring when they were confident their own query terms would retrieve useful information. This finding is consistent with the finding in Kantor (1987) that the more a searcher believes he or she can find a satisfactory document, the more non-relevant results the searcher is willing to tolerate.

One participant who reformulated early attributed his behavior to both the “zero penalty” of reformulations modern search engines afford and to his/her innate desire to pursue better results:

My going back and retyping, or not going to the next couple of pages maybe isn't necessarily an indication of impatience, but a knowledge that this is not your last chance....I can come back to this search configuration if I want to....it is a desire, a sense that with a new term maybe I can get better results. (P47)

This quote demonstrates a belief that if one tries hard enough, eventually one can find the best query term that retrieves the best set of search results in relation to one's information need. Knowing that the results exist and the results can "wait" to be discovered, moving away from the current query is therefore favored over result filtering and gathering. The ability to re-type a previous query and re-examine earlier results makes exploring new search queries more appealing. Another participant argued that reformulation allowed for new directions of thinking and serendipitous discoveries.

I think it is probably better to submit a new search [when not finding enough on the first page], partially because you are deliberate in your terms that you are searching, you are also putting into a different thought process. It is almost like it allows you to be flexible in your thinking. Because if you are looking for something and you know what the answer is, you are only to get certain amounts of results. And then if you take inspiration by words, you see whether it is in an article or a phrase that might lead to a different path, and a deeper understanding about something that maybe related in different ways. (P9)

5.7.1.3 Properties of the search tasks. Participants judged how much effort they should put into searching by task properties. The amount of effort one was willing to put in subsequently influenced how many results they were willing to explore before they abandoned a query. For example, participants' own task typologies allowed them to gauge how much

information was needed. Two typologies reported by participants were factual vs. opinion and work vs. personal.

This one is all about opinions, I assumed I had to go more pages to get both sides. ... It is such a broad question. It is less about a good query but about finding one or two good things. ... I feel [for this type of search] I click on the second page normally. For specific ones, I don't need to. But for opinion-based, I would, to see both sides. But I probably like to go to the second page [more] than to change the search. If it is vague, I will go to the second page because there may be relevant results which are just not as connected to the way I search. (P28)

At work I cover my bases more deeply. I need to present information to other people and I don't want to be caught off guard. I don't want to be asked questions and, oh, I wished I had gone deeper. But I feel like at home or things for personal use, I sort of just skim it. (P2)

P2's and P28's comments support the view that the amount of information needed to reach the feeling of enough depends on task type (Bates, 1984). P2's comment also demonstrates that when searching on behalf of other people, there was ambiguity in what enough meant for these people, so he or she adopted higher standards. Yet when searching for oneself, because there were fewer consequential effects, less effort was deemed necessary.

Participants reported that in their daily search engine experiences, they almost always found needed information on the first SERP; therefore, when asked to recall any previous experiences of pagination, they had a hard time coming up with any examples. While the inability to recall pagination experiences could be caused by a tendency to reformulate on the

first SERP, it could also be that most of the participants' memories were about simple look-up tasks.

Some participants decided to reformulate each time a subtask was fulfilled. Each reformulation marked the division between one search goal and another. As described by Participant #15, "part of the changing of search terms was also to hit the different aspects in the prompt."

While what P15 said demonstrates that participants read the task descriptions carefully and tried to complete the requirements accordingly, it nevertheless suggests that rather than treating a simulated scenario as a larger task and trying to come up with a search plan on their own, some participants were actually solving multiple independent smaller tasks in a linear order where each subtask appeared in the description. Further, task relevance or task importance determined how much search took place before participants terminated subtasks. For example, Participant #14 said that for something trivial searching at home, 1 or 2 documents was enough; but for things affecting society, he or she tried to make sure he or she had critical information by searching for more. When tasks were not important to participants, they did not see the need to search deeply to fulfill them.

5.7.1.4 Properties of the person. Lack of motivation was a strong deterrent to query reformulations in some cases, while a heightened interest in, or a motivation for information often led to prolonged searching. Prior knowledge sometimes resulted in pre-supposed answers before searching, leading to earlier reformulations.

I usually stay on the first or two pages, unless there is something I really need to know about or something I am really interested and it's just not coming up no matter what I put in the search engine; then I will go beyond the first page. (P11)

While the characteristics of queries, search results, and task that cause searchers to reformulate sooner or later vary under different circumstances, some individual differences were stable enough to contribute to common responses across scenarios. For example, participants' search experiences with commercial search engines had a direct impact on when they reformulated. Some participants who had experienced the benefits of searching deeper in real life developed the habit of always going further before reformulating their queries. For example, Participant #17 said in his experience he could find helpful things on the second page, so he/she generally went to the second page.

Another participant compared searching beyond the first SERP to browsing a library shelf:

Sometimes you can find jewels, gems on pages that are not on the first [page]. And that takes you to things you might not have thought about. It's like going to the library and you know what books you are getting, you have the call number, and then you browse on the aisle side, and it is when you think, that might be an interesting connection to this, it's almost like it is a spider web, where you have the beginning and everything starts to feel more tangential around it. (P9)

However, the majority of participants expected to find everything on the first SERP and when searching with Google, the one search engine participants referred to when they talked

about real life search experience, tended to reformulate queries before they entered the second SERP out of habit. One participant even said that Google had “spoiled” him.

When you look for something it should be on the first page. The most relevant things are going to be on the first page. If you go to the second, third, fourth [pages] you are getting out of topic. The second page may be irrelevant or repetitive information. (P30)

A librarian participant perceived the tendency to reformulate queries early as a characteristic of many searchers today, especially the younger generation. This participant grew up using the library and was taught to go through books and journals exhaustively for information; as a result, she had not developed the expectation of finding everything fast and early on the first SERP, which contributed to her not reformulating queries early.

5.7.1.5 Summary. While many participants chose to reformulate on the first SERP because they believed that search engines presented the most useful results on the first SERP or because of the ease of reformulations with search engines, many other stopping decisions appeared to be contingent on the tasks. The subjective task typologies and task relevance allowed participants to make high-level decisions about the amount of information needed, and the structure of the task influenced the number of query stoppings observed for a search task. The confidence in one’s own query terms could impact query stopping as well, yet either high or low confidence led to delayed query stopping for different reasons.

Characteristics of the searcher, including motivation, knowledge of a topic, and previous search experience with search engines, were also found to influence when participants decided to reformulate. High motivation kept participants longer in a search result set; greater knowledge

caused participants to reformulate early; and a long and close relationship with modern search engines led to the habit of examining only the first SERP.

After issuing query terms, more contextual factors influenced query stopping decisions. The perceived quality of search results, which could be a result of simply the first impression or based on a more careful analysis of the results, had a critical impact on when stopping took place. For participants who articulated how specific elements of search results influenced their stopping decisions, observing a lower number of relevant search results was often the reason why they issued another query after examining the first SERP. Even when there appeared to be some good relevant results, when non-relevant results appeared in a block or seemed to occupy the lower end of the first SERP, participants felt continuing to the second SERP was unlikely to be promising. Moreover, the sight of commercial search results played a critical role in participants' decision-making. Participants also calibrated their search result evaluation efforts based on their assumptions of how search engines responded to third parties. For instance, the knowledge that search engines sold highly-ranked search result space and search results could be ranked highly because many other users clicked on them, sometimes caused participants to explore search results ranked lower or beyond the first SERP.

5.7.2 Task stopping strategies. Participants tended to explain their task stopping decisions by relating to the content they had reviewed, the goal they wanted to achieve, how they felt, and the study constraints.

5.7.2.1 Content. A common strategy to determine when to stop was by exhausting the resources, including the exhaustion of relevant search results and the exhaustion of known query terms. Participants would not have known objectively whether they had exhausted all relevant search results or not, so the perception of having exhausted results was based on their

interactions with search results already seen. One participant explained that she stopped because “I could not find anything else and I could not find other terms“ (P12). Another stopped because “the links seemed not relevant anymore” (P31). Still another felt he/she had exhausted everything when he/she went to the second and third pages and still could not find what he was looking for“ (P35).

The stopping rules identified in previous research were also articulated by many participants. For example, the above phenomenon can be explained with the *Difference Threshold Rule* (Nickles et al., 1995). Searchers stopped when they did not learn anything new.

Participants did not always push themselves to the extreme; rather, they took the satisficing approach. The concepts of “sufficiency” and “feeling of enough” which have been frequently observed in the literature were repeatedly mentioned during the course of the interview. Participant #43 stated the feeling of enough was, “much of a qualitative sense. I feel I kind of know what is going on, rather than I know exactly. It's like a gut thing.” Another participant made sure he had enough by asking him/herself: “Have I to my satisfaction answered the question? Have I felt I reached a critical mass of the knowledge of this thing?” (P47)

Others took a minimalist approach, believing enough means “necessary information to answer the question” (P45) or “more than one thing, had a variety, and be able to address the question” (P42). Some other participants were able to articulate what enough means in more concrete terms under different circumstances:

I think it is an important issue; that is why I got four instead of three articles...It is like you have a three or five panel judge. With two [judges] you cannot decide, you got to decide with three, five or seven. (P44)

P44 first explained that three articles normally met the threshold of enough, because three represented the least number of voices to reach a diplomatic consensus. Yet in this particular task, which was important to him, four articles rather than three was considered enough.

Not only was number of articles used as a criterion for determining task stopping, for some participants the articles had to belong to a certain genre:

I guess the source of the documents was from journals, medical standpoint and psychology journals, experts, I thought there was evidence, a good answer from not only someone's opinion. Someone has really done a study... I tried to [monitor how many useful documents I had found]; I was trying to get 4 to 6 documents. (P37)

But for P7, under the circumstance when the documents found were all opinions rather than solid evidence, she considered four documents good enough.

I guess in my searching if I have to read 4 articles, like longer articles about this, then that would be an exhaustive search for me. Because these are people's opinions, I can imagine the 5th, 6th, 7th articles about the same articles I would sort of get... After the 4th [article] I would probably have formed an opinion...Again, I guess after the 4th one I started to see enough overlap to make me think the fifth one wasn't gonna get me more. (P7)

On the surface it may seem that all these examples supported Kraft and Lee's (1979) satiation rule and Cooper's (1973) satisfaction stopping rule since participants stopped after obtaining a certain number of useful documents. Yet the mentioning of how different genres carried different weights demonstrates that quantity alone did not tell the whole story.

Earlier it was mentioned that some participants treated each task as a smaller number of independent tasks. Participants reformulated once they had fulfilled every subtask listed in the

task and they stopped the task once all subtasks were completed. Such a stopping decision can be explained by the *Mental List Rule*, stop when a list of requirements is met (Nickles et al., 1995).

In the list of questions I am looking for, I knew I had found several different types of pollutants and I have links saved about how they were harmful, and I felt I had enough relevant ones I can answer these questions. (P18)

In this study, the requirements were the questions in the task descriptions. The mental list rule was observed primarily in tasks that appeared to be relatively easy for participants to divide into smaller units, which also supported Browne et al.'s (2007) findings.

In contrast to trying to satisfy each individual subtask, other participants took the approach of obtaining a general understanding of the topic. The *Representation Stability Rule* (Nickles et al., 1995) can be used to explain this approach. Once a mental model was established about the task topic, participants stopped searching for information. An example of stopping following this rule was made clear by Participant #27: "Once I had a good idea, either what I already thought was confirmed, or something new came up, [I stopped]."

At times participants reported that when a pre-conceived belief was confirmed or supported by retrieved search results, they stopped. Such phenomenon can be explained by the *Single Criterion Rule* (Nickles et al., 1995). Participants stopped looking once enough information about a single predetermined criterion is found. For instance, Participant #24 said "I think because I already had an answer in my mind. I was looking for information to reaffirm mine. I felt like I found several articles that said that." In this example, the single predetermined criterion was the existing answers or beliefs participants held in mind before searching.

Participants who did not have preconceived answers sometimes stopped when they sensed a coherent theme emerging from repetitive information. Participant #43 used the tattoo removal task as an example: “It seems the information was consistent. They all agreed laser was the best. After a couple of research [studies] saying that, I was like, I am not going to find out too much more.” This was another example of applying the *Representation Stability Rule* to determine when to stop.

In tasks where participants were asked to collect information about the pros and cons of using social media and the impact of violence in video games, participants were concerned about gathering balanced opinions in order to achieve an unbiased decision. Participant #11 stopped when he felt he had a good variety of information to look at and it wasn't just one-sided. For Participant #22, a fair assessment of an argument had to be comprised of one positive opinion, one in the middle and one negative opinion.

These two comments, along with the three-panel judge comment mentioned earlier, demonstrate that participants treated debatable topics with greater caution than average open-ended topics where there may exist standardized answers (e.g., most toxic marine pollutant).

It was also found that sometimes participants stopped as soon as a satisfactory document was found. While in many IR models this strategy is hypothesized to occur more often in fact-finding tasks, it seems to apply to open-ended tasks, too. Some participants tended to gather information until they found the best article that met all requirements in the task description. One participant stated that he/she stopped because, “this is one of the things there is no answer to it. There is no conclusive evidence, they all say the same things. This one has both sides I feel is better than other sites“(P6). This happened after participants struggled through repetitively encountering documents that only fulfilled some aspects of the task requirements but not all.

5.7.2.2 Goal. Some participants stopped searching once they reached the goals they set to achieve. One of the goals participants wanted to attain with their search efforts was being able to engage in a conversation about the subject matter of the task.

By the time I stopped I have managed to find enough sources that gave me 5 or 6 options. And in the hypothetical task if my sister is turning 25 and I have some idea of what she is talking about, then at least I have a basis to have a conversation with her. (P18)

One participant justified the amount of information she needed to obtain based on the length of the conversation he or she wished to sustain as the following: “If you and I are meeting for dinner and you brought up this I can easily talk about it for at least 20 minutes” (P5). As an English literature instructor, P5 also imagined she was assigning students to write a report about the topic and predetermined a fixed number of resources to be gathered before initiating the search.

Imagine I am writing a paper. Let's say I make students write a 3 or 4 pages paper investigating in some reasonable manner, 5 to 6 sources, you know, max, so that I will be able to discuss it intelligently at a dinner party. The criteria is to have some reasonable knowledge on both sides and be able to write a 3 to 4 pages paper about it, just to present the facts. (P5)

The strategy P5 took, stopping once a certain amount of information was obtained, fit in the *Magnitude Threshold rule* proposed by Nickles et al. (1995). Yet it was not commonly observed that participants had a specific number in mind before they embarked on searching.

5.7.2.3 How they felt. Frustration from not finding useful information, perceived passage of time, and lack of interest in the topic were some subjective factors that resulted in premature

task stoppings. Participant #6 appeared embarrassed when he said: “I was kind of unsatisfied. Like I said, it was really hard to find conclusive evidence. I figured, well, maybe the fact there is a lot about lasers [and] much less about others tells me, maybe laser is it.” Another participant articulated his indifference toward the topic:

I don't really care that much about it. And at the same time there are so much hearsay. Well, I read some of that, that says yes, it is bad. And I think after this article I found a couple others that actually have more factual kind of evidence than just opinions. I think that's all I need to know. (P8)

5.7.2.4 Study Constraints. In addition to lack of interest, which was probably a consequence of administering artificial tasks, participants described several other study constraints that caused them to spend more or less time they would have normally spent on searching.

Participant #15 explained the difference in time between a real need and a simulated need: “If I was truly picking a water sport for my sister's birthday, I would've spent so much more time working. So a lot of it is the time constraint of the study.” Even though no time constraint was placed on participants, P15 was tracking the passage of time to complete all six tasks in a reasonable time. In addition, what could have been a multi-episode task was forced into a single-episode task, which might have also led to the decrease in time and effort participants invested.

Participant #47 put more effort into searching because he knew he was examined during the study:

None of these are emergency, this is an artificial situation, when I am at home I would bookmark. I can go back. In the spirit of the experiment, I wanted to make sure I tried

enough angles, at least have some representative corpus. By the time I finished [that] it looked like someone with a brain doing research. (P47)

5.7.2.5 Summary. A task stopping decision was often made directly as a result of interactions with SERPs. Participants described incidents of “forced” stopping, stopping when no more information could be found, as well as incidents of “voluntary” stopping that resulted from securing enough, or necessary information. It was also the case that some participants decided how many results to obtain at the start of a task and stopped immediately after the number was satisfied, although this was not a common case. Task structure was used by participants to justify when optimal task stopping should occur, just as it was used to justify query stopping. Moreover, the nature of the search task appeared to influence task stopping decisions as well. For opinion-based tasks, participants focused on gathering a “balance of arguments” and used this to determine stopping. Imagining the potential use of the information was also used by some participants to decide when to stop gathering information. Participants related to scenarios where one needed to give others suggestions or one needed to produce a report where concrete criteria needed to be satisfied before quitting. Participants also allowed emotions and motivations to influence their decisions about when to stop. Lastly, participants’ awareness of being in a study motivated some to search more and others to search less.

5.7.3 Pagination-prone searches. At the end of the interview, participants were encouraged to recall in their real life when they actually paginated to gather information. The most commonly described searches where participants paginated were when they conducted people search, product search, image search and literature search. In people search, participants commented they knew little about a person and many people could share the same names, so they did not mind filtering through search results. In product search, because some participants

did not want to miss great deals, they were willing to go deeper just in case; some others went through several pages because the product name was specific enough that they were convinced as long as they were patient, they could find information about it. Participants said they paginated in image search because processing images was perceived as less effortful than text. With regard to literature search, some participants commented they often used Google Scholar or library databases, which they believed were more trustworthy; thus they were comfortable going through multiple pages, assuming results deeper were also credible. Some other participants were exhaustive with result evaluation because they did not want to miss any related study.

5.7.4 Search styles. Based on participants' query stopping tendencies, participants exhibited three distinct types of behaviors:

First-SERP searcher: First-SERP searchers were participants who had the tendency to reformulate queries after reading a few documents on the first SERP. The majority of participants belonged to this type. First-SERP searchers limited their searches to, at most, the first ten search results. If the search results appeared non-satisfactory or if more information was still needed, participants submitted new queries. First-SERP searchers can be further divided into two groups based on their level of faith in search engine algorithms. One group of the first-SERP searchers disregarded all results appearing beyond the first SERP. For example, one participant said she never checked results on the next SERPs because she believed everything beyond the first page was “sketchy” (P45). The other group of first-SERP searchers were aware that search engines did not always do the best job, but still chose to submit new queries on the first SERP rather than paginate because they perceived it easier to search when all relevant results were ranked high.

Interpreter: Interpreters were those participants who adjusted when to reformulate query terms based on the characteristics of the first SERPs. If the first SERPs looked promising and information was still needed, participants paginated; if they learned new concepts as they examined search results, they reformulated right away by incorporating the new concept into new searches.

Explorer: Explorers examined the first X SERPs regardless of the quality of the first SERP. X could range from two to ten SERPs in our study. Participants in this category always reviewed search results beyond the first SERP; they had the tendency to reformulate queries after viewing certain numbers of SERPs out of habit.

5.8 Summary of Results

RQ1: What is the relationship between the information scent level of the first SERP and search stopping behaviors?

The findings of the study demonstrate that there was a positive relationship between information scent level and query stopping. The amount of interaction between the participants and the search results prior to query stopping could be predicted by information scent level; when information scent level varied from low to high, the values of almost all measures increased as well; a greater extent of interaction was observed before participants ended search result examination for a query when information scent level was higher on the first SERP. Observations from the interviews also reveal evidence that information scent level affected query stopping. For example, lack of relevant results on the first SERP motivated some participants to issue a new query.

The order of information scent level in a task had a significant effect on task stopping as well. Significantly more search tasks ended immediately after participants were exposed to high information scent level than medium and low information scent level. Moreover, when the first treatment at the start of a task was medium or high scent level, participants evaluated their abilities of retrieving useful information more positively and rated the task easier than when the first treatment was low scent level.

RQ2: What is the relationship between the information scent pattern of the first SERP and search stopping behaviors?

Information scent pattern was found to have a relationship with query stopping through a few measures: Abandonment, NumRele and NumNonRele. Participants were more likely to leave a SERP without clicking on any result when the information scent pattern was bursting than when it was persistent and disrupted. Participants examined more relevant results prior to query stopping when they were exposed to the persistent pattern or the disrupted pattern, and examined more snippets of non-relevant results when exposed to the persistent or bursting patterns than the disrupted pattern. For participants who examined only results on the first SERP, they tended to explore results deeper when they were exposed to the persistent and bursting patterns. Self-report data from the interview supports these observations. Participants indicated that, when non-relevant results appeared in a block or occupied the lower end of the first SERP, they felt continuing to the second SERP was unlikely to be promising and thus reformulated the query.

The first information scent pattern at the start of a search task appeared to affect the number of queries submitted prior to task stopping ($p=.054$). When the persistent pattern was shown at the beginning of a task, participants issued more queries than when it was the disrupted

or bursting pattern. Information scent pattern also affected how difficult the task was perceived to be and how difficult it was for participants to decide whether they had enough information to stop searching. When the persistent pattern was shown at the start of a task, participants reported the task was easier and it was easier to tell they had enough information than when participants were shown the bursting information scent pattern at the start of a task.

RQ3: What is the relationship between NFC and search stopping behaviors?

The results indicate that there was a negative relationship between NFC and query stopping. It was found that, in tasks where information scent level was manipulated, the lower the NFC, the more SERPs participants paginated and the deeper the mouse hovering before query stopping took place. Interactions were found between NFC and information scent level on the likelihood to reformulate, paginate or stopping and the time spent examining results prior to query stopping.

It was also found that NFC had a positive relationship with the number of queries issued prior to task stopping in study tasks where information scent level was manipulated. No matter what information scent level was displayed at the beginning of a task, participants with higher NFC issued more queries than participants with lower NFC. Significant interactions between NFC and information scent level were found on the depth of click, the number of non-relevant documents examined and the number of snippets of non-relevant results examined prior to task stopping, and between NFC and information scent pattern on the number of SERPs paginated and the depth of mouse hovering.

RQ4: How can we model task stopping using interaction signals?

Several Pre-Task Questionnaire items were correlated with the amount of effort devoted to searching prior to task stopping. These include previous search experience on a topic, knowledge of a topic and interest in a topic. The greater a participant's search experience, the fewer queries they issued; the greater their knowledge of the topic, the fewer queries issued; and the greater their interest, the more time they spent on searching.

The number of queries issued in a task affected the amount the interactions that happened prior to the end of a task. When ISL was manipulated, participants who issued more queries engaged in more interaction. When ISP was manipulated, participants who issued two or three queries in a search task interacted with search results to a greater extent than participants who issued only one query in a search task. Even though no inferential statistical test was conducted to assess whether the combination of task length and order of treatment affected task frequencies, the descriptive statistics suggest the bursting ISP often resulted in reformulation. Moreover, as reported in RQ1, more task stoppings occurred immediately after participants were exposed to high ISL. These findings suggest that a SERP with relevant results ranked consecutively at the lower ranks motivated more interactions prior to task stoppings, while a SERP with many relevant documents often resulted in task stoppings.

The analysis of search stopping behavior patterns prior to task stopping resulted in nine commonly repeated sequences of moves of three lengths, which can possibly be used to predict task stopping. The most common behavioral pattern prior to task stopping was clicking and viewing only one search result on the last SERP before ending a task, and the top three most common patterns account for the search stopping behavior patterns in 84% of all search sessions.

The task stopping strategies shared by participants during the interviews provide additional criteria that can be used to predict task stopping. For example, it was found that task

stopping occurred after all aspects of a topic were queried. For opinion-based topics, participants stopped when they obtained balanced information on competing views.

Chapter VI. Discussion

This chapter discusses the findings of the study. The discussion is first organized by the two search stopping behavior types studied in this dissertation research: query stopping and task stopping. Under each search stopping behavior type, interpretations and implications of the major findings are presented.

6.1 Query Stopping

To answer the question in the title of this dissertation, How far will you go?, this dissertation first examined the relationships between ISL and query stopping, between ISP and query stopping, and between NFC and query stopping through the use of search behavior measures. This work also investigated other factors that affected query stopping decisions during the experiment as well as in participants' daily lives. The findings reveal situational and individual variables that determined when participants stopped evaluating searching results retrieved by a query and issued another query.

6.1.1 First impression determined clicking. Earlier in the literature review it was mentioned that search results retrieved by a single query submission can be regarded as an information patch (Pirolli, 2007). Query abandonment, or not clicking on any results on the SERP, can therefore be interpreted as early query stopping in an information patch; it is considered early because participants left the information patch before they examined any content pages. The results of the present study show that information scent level and pattern could be used to explain why such early stopping took place. When the first SERP showed only

one relevant result, the abandonment rate was the highest among all three information scent levels; participants probably assumed they were unlikely to retrieve relevant information from searching an information patch where they could only observe one relevant document at the first glance, so they decided not to engage with the information patch. Similarly, when participants were presented with the bursting ISP, the abandonment rate was the highest among all three information scent patterns. It is possible that participants did not realize there were four relevant documents at the lower ranks and therefore the information patch appeared not worthy of their effort. Moreover, the abandonment rates for the persistent ISP and the disrupted ISP were both low. A persistent ISP could have maintained participants' attention because the interleaving of relevant and non-relevant results prompted participants to continue examining the SERP. It might also be the case that as long as the first two results were relevant, the information scent provided by the SERP was enough to engage participants at the information patch level.

6.1.2 The higher the scent, the greater the number of interactions. The relationship between information scent level and query stopping can be summarized at two levels: at the search result set level and at the first SERP level. At the search result set level, when information scent level was high, participants clicked on search results ranked lower, spent more time interacting with search results, and examined more results. These findings suggest that more effort was devoted to examining results in these cases and the depth of search was greater prior to query stopping when the first SERP exhibited a greater potential as a useful information patch.

At the first SERP level, information scent level could explain how deep into the first SERP participants explored search results. For participants who only interacted with search results on the first SERP, not only did that information scent level have significant effects on all the search behavior measures mentioned in the previous paragraph, the depth of mouse hover, an

approximation of search attention, also demonstrated that greater scent was associated with greater interaction. The positive relationship between information scent level and amount of interaction is aligned with previous findings where it was found that the amount of information scent on the homepage of a Website predicted the number of pages visited on a site (Card et al., 2001). This also indicates that information scent level is useful for predicting people's online interaction behaviors beyond Website navigation and that information scent level can be used to explain how much people will interact with SERPs when conducting open-ended search tasks.

6.1.3 Distribution of scent mattered, but not as much as ISL. The relationships between ISP and task stopping can also be described at the same two levels: search result set level and first SERP level. While the persistent ISP led to the greatest amount of interaction for many measures, the GEE method only identified significant results for NumNonRele, NumRele and Abandonment at the search result set level. What was more informative was the effect of ISP on task stopping on the first SERP. For participants who only interacted with results on the first SERP, information scent pattern could explain the depth of search before they decided to stop their queries. The results showed that when relevant results were ranked on the first SERP using the optimal ranking, the disrupted pattern, participants stopped interacting with search results at higher rank positions. On the one hand, this finding possibly suggests that when search results were contiguously positioned at higher ranks, by the time information scent disappeared after the fourth rank, participants might have assumed they had seen all the useful information. On the other hand, this finding demonstrates that when exposed to the persistent and bursting ISPs, participants did not perform worse than when they were exposed to the disrupted ISP, as there was no difference in the number of relevant results saved.

The increased depth of results exploration for persistent and bursting patterns implied the possibility of applying alternative rankings to motivate deeper search result exploration. This implication is especially relevant in the context of the design of online experiments. By using the persistent pattern, researchers can possibly increase the number of manipulations experienced by searchers in large-scale online experiments while at the same time maintaining a low abandonment rate. Take the problem of query disambiguation, for example. Given an ambiguous query (e.g., mac), interleaving Web results of different senses (e.g., Mac the computer and MAC the makeup) might engage users to review more results and allow search engines to gather more user interaction information before a decision is made about intent.

Even though the inferential statistical results did not find information scent pattern useful for explaining the number of SERPs paginated, evidence of its potential impact can be gleaned from the interviews. For example, one participant believed more junk would follow if he or she had continued after seeing non-relevant results at lower ranks. Some other participants inferred based on the distribution on the first SERP the distribution would continue onto the next SERP. Still one other participant reformulated upon seeing non-relevant results positioned in a block. These observations to some degree suggest that information scent pattern could affect one's likelihood to continue to the second SERP.

6.1.4 Search forward vs. Search deeper. The findings of the study show that NFC did not have a significant effect on either the amount of information searched or the time spent searching, which is not aligned with the findings from previous studies (e.g., Curseu, 2011). However, the results do show that the effect of NFC manifested by causing variations in people's search strategies. Participants with higher NFC paginated less and stopped hovering at higher ranks before query stopping than participants with lower NFC; participants with higher NFC also

submitted significantly more queries before task stopping than participants with lower NFC (a detailed discussion of this comes later under Task Stopping). These findings suggest that participants with a higher NFC had a tendency to devote their cognitive efforts to frequent query reformulations rather than prolonged result evaluation. The resulting difference in search depths implies that a different weighting scheme or persistence parameter in evaluation measures such as nDCG and RBP can be adjusted based on an individual's NFC. By doing so, the same search result can be assigned different weights depending on an individual's likelihood to access the result; search results at the same rank positions will therefore be discounted to a greater extent for a searcher with higher NFC than lower NFC.

The effect of NFC on search depth and query submission has other implications. For searchers with lower NFC who have a tendency to explore search results deeper, displaying more results per page or automatically loading subsequent pages can save them the need to click. Moreover, when it comes to large scale experiments, using lower NFC searchers as test subjects can perhaps increase the number of manipulations experienced. For higher NFC searchers who have a tendency to issue more queries to solve an information problem, more information should be provided to facilitate query reformulations. Making query suggestions more salient to higher NFC searchers can possibly help them become more resourceful in generating useful query terms.

With regard to the situations where participants mentioned they paginated to gather information in their daily life, including people, product, image and literature searches, other SERP design options can be applied. For example, once a people search is detected, presenting advanced search filters such as job title, work place, schools attended, and graduating year can

eliminate the need for exhaustive review of search results. Other criteria can be extracted for product search as well.

6.1.5 Moderating effects of NFC on query stopping. The most interesting findings from the present study were probably the interaction effects between ISL and NFC. From the predicted probabilities of reformulation, pagination and stopping in Figures 10, 11 and 12 one can tell that given a query submission, the most likely outcome was reformulation regardless of ISL treatment; the predicted probabilities for pagination and stopping were both very low for every ISL treatment. It is also shown in Figures 10 and 11 that the slopes of medium ISL were steeper than the slopes of low and high ISL when predicting the probabilities of reformulation and pagination, which means that NFC had a more profound effect when the SERP contained three relevant result pages. The consistently high predicted reformulation probability of low ISL regardless of NFC probably suggests that seeing only one relevant document convinced participants the search results were not worth their time continuing to explore. Yet when there were three relevant documents, participants of different NFC scores had varying interpretations of whether the first SERP looked promising enough to devote more time and effort; the higher the NFC, the more likely a participant chose to reformulate.

When seeing three relevant results on the first SERP after a query submission, participants of extremely high NFC scores exhibited a higher probability of reformulation than when seeing only one document, and participants with extremely low NFC exhibited a lower probability of reformulation than when seeing five relevant results. However, for participants with mid-range NFC scores, the predicted probability of reformulation increased when information scent level shifted from high to low. These results are consistent with Axsom, Yates and Chaiken (1987) where it was found that behavioral variance was the most profound when the

contextual factor was at the moderate level while behavior was more uniform when the situation was extreme; for example, when a situation has high relevance to an individual, he or she will attend to the situation with great cognitive effort regardless of NFC. Similarly, in this study, seeing three relevant results on the first SERP perhaps triggered participants with very high and very low NFC to interpret the quality of their searches very differently. When a participant with low NFC experienced three relevant documents, he or she did not feel there was enough information but was hopeful of finding more by continuing searching; seeing three relevant results might suggest to the participant that he or she was already on the right track, following the same path could save one effort from creating more specific queries. Participants with higher NFC, on the other hand, were more likely to have decided three documents were enough for a query submission and it would be a more cost-effective strategy to move on to another search in order to retrieve results related to other aspects of the topic. In other words, since on average participants saved 4.82 documents for a search task, when only three relevant results were displayed on the first SERP, which was not enough for completing a search task, participants with lower NFC were more likely to paginate to find more information while participants with higher NFC were more likely to reformulate to gather more information.

Another interaction effect between ISL and NFC was found on time spent in a search result set. Participants with lower NFC showed greater variability on the time spent evaluating results depending on the number of relevant results presented on the first SERP. This finding may be explained by the Elaboration Likelihood Model (Petty & Cacioppo, 1981, 1986), with which NFC has a close connection. According to Cacioppo et al. (1983), participants with higher NFC would be more likely to base their decisions about whether to devote more time to searching based on careful scrutiny of results. As a result, they could have come up with new

queries while reading results or snippets ranked high, which might be the reason they spent relatively uniform time on SERPs regardless of ISL. On the other hand, as the model suggests, participants with lower NFC were less motivated to engage in cognitive activities and thus, more likely to use heuristics when deciding about whether to continue searching. They perhaps used the number of relevant results on the first SERP as a heuristic to decide how much time to invest in their searches. When they encountered more relevant documents, they might have been more convinced of the quality of the search, and subsequently spent more time evaluating results.

6.1.6 The non-paginating behavior. This dissertation showed that even in open-ended tasks where participants need more than one result (or snippet) to solve an information problem, participants were more likely to reformulate to gather information than to paginate; only 20% of the query submissions led to result examination after the first SERP. While researchers often decide an arbitrary rank which users reach upon which to base hypothetical user models and task models for system-centered evaluation measures (e.g., precision at k , the number of relevant documents among the first k results), setting a rank beyond 10 seems unrealistic in most cases, even for information-gathering tasks. Even in high ISL where there were five relevant results at the optimal ranking, the reformulation rate was still as high as 50%. If most results beyond the 10th rank are never examined by people when using search engines, evaluating algorithms based on an inclusion of results beyond the 10th rank probably does not reflect users' perception of system performance at least for the types of tasks typically evaluated in experimental IR.

It may also be the case that the pagination rate was low because the study tasks were not among the search scenarios where participants usually examined multiple SERPs in their daily life. Perhaps it is only for special search scenarios, such as people, product, image and literature searches, as identified by the participants in the interviews, where searchers will ever consider

pagination. If this pagination is indeed an “abnormal” search behavior in search engine usage, information problems leading to pagination should be studied to provide useful approaches to facility problem solving.

The knowledge of search engine algorithms and the ease of re-querying offered by modern search engines probably also explain why this and previous studies have found that reformulation was more common than pagination. Participants in this research were aware that search engines ranked search results by quality and many believed that issuing new queries was an easier search strategy than sifting through search results. Therefore, for participants who had developed the habit of searching only among the first ten results, it was as if only ten, rather than millions of results were retrieved for each query submission. The rarity of pagination and the tendency of reformulation observed in this study call for a reconsideration of the current search result presentation practice. What does it mean to offer results that almost no one will examine? How can search results that are not displayed on the first SERP be exposed or integrated to allow for more diverse solutions and serendipity? How can we get people to go deeper in the search results list? And, finally, is there a way we can leverage the habit of non-pagination to modern searchers’ benefits?

6.2 Task Stopping

This dissertation shows that participants’ task stopping behavior varied to a great extent. The number of queries issued in a search task ranged from one to 19, with about 27% of search tasks containing only a single query submission. To further understand task stopping, this dissertation analyzed task stopping by task length and treatment order in a search task. Following that, this study analyzed, categorized and compared the behavioral patterns prior to task

stopping. Lastly, this work investigated factors affecting task stopping during the experiment and in real life search experiences.

6.2.1 The enduring effects of first impression. Information scent at the start of a search task could affect task stopping. When the persistent ISP was presented at the beginning of a task, participants appeared to issue more queries prior to task stopping. Perhaps seeing the persistent ISP at the beginning of a task suggested a sense of sparseness of useful results of a given topic; therefore, frequent query reformulating was adopted as a means of trial-and-error to gather enough information. This result indicates that while persistent scent on the first SERP did not lead participants to paginate more frequently before query stopping, when persistent scent was presented at the start of a task, it triggered more query submissions to gather information.

Not only did the first treatment in a task affect subsequent search behaviors, it also affected participants' post-task evaluations. When the first ISL was low, participants felt a task to be more difficult after the search than when they were shown a medium or high ISL at the start of a task. Displaying low ISL in the beginning of a task also led participants to feel they were less successful at solving the tasks than when they were presented with the medium or high ISL. Moreover, when the first treatment was persistent ISP, participants rated the tasks easier after the search than when the first treatment was bursting ISP. More interestingly, presenting a SERP with persistent ISP at the start of a task made participants feel it was easier to decide when they had enough information to stop than when the first treatment was bursting ISP. These findings imply that early interaction during a search session could have set a tone for subsequent search experience; being exposed to only a few relevant results or being exposed to multiple relevant results at lower ranks might have made participants believe it would be challenging to find useful information.

Earlier in the literature review section it was shown that most past research indicated that higher NFC led to more information processing, yet the findings were not derived from interactions with SERPs. The extent of information processing in the previous studies was often measured by objective counts of messages processed (Verplanken et al., 1992; Verplanken, 1993; Bailey, 1997; Nair and Ramnarayan, 2000). This dissertation research used multiple search behavior measures to assess the amount of interaction prior to task stopping, and the results showed that participants with higher NFC neither examined more search results nor spent more time searching, but they did submit more queries to solve an information problem. In other words, a participant with a higher motivation to engage in cognitive activities invested more effort in generating search queries than another participant with a lower motivation to engage in cognitive activities, but such a strategy did not result in examining more documents or spending more time searching. However, compared with many previous studies where the documents of interest used in the experiments were more structured (e.g., product attributes), the Web pages used in the present study were probably more variable in the layout, length and other characteristics. These variances could also be the reasons why no difference was found in the number of results examined and task time prior to task stopping.

NFC also had an effect on post-task experience. The results showed that higher NFC participants believed the search tasks were more difficult after searching than lower NFC participants, which could be a result of greater processing. A tendency to more carefully process information while interacting with SERPs could have triggered a closer examination of search results, with more in-depth questions left unanswered, subsequently resulting in higher task difficulty ratings after searching.

The effect of NFC on task stopping was not always consistent across participants. The first treatment in a search task had varied effects on participants of different NFC on several measures. When the first treatment at the start of a search task was high ISL, as NFC increased, the accumulative depth of mouse click increased as well. Being exposed to high ISL probably suggested to participants with higher NFC more relevant results could still be found at lower ranks, which motivated them to dig deeper and tolerate more non-relevant results. To the contrary, when the first treatment was medium ISL, participants with higher NFC did not click as deep and examined fewer snippets of non-relevant results than participants with lower NFC. Seeing only three relevant documents ranked consecutively from the top of the first SERP probably caused participants with higher NFC to focus only on results ranked higher. When the first treatment was low ISL, no matter what NFC a participant had, it did not affect the depth of mouse click and the number of snippets of non-relevant results examined. Participants probably interpreted the low recall of relevant results at the start of a task as a mistake resulting from their own queries, thus they did not take this early interaction seriously. For ISP tasks the effect of the first treatment did not have as profound effect on participants with higher NFC than participants with lower NFC on the number of SERPs paginated and the accumulative depth of mouse hover. This shows that the ranking of search results could be viewed as a peripheral cue that was relied on by participants of lower NFC to infer the ranking of future results to a greater extent. The bursting ISP probably suggested to participants with lower NFC that relevant results would be located at lower ranks for future query submissions, therefore they were more likely to hover lower and paginate more frequently.

6.2.2 Predicting task stopping. The characterization of commonly repeated patterns prior to task stopping adds to the research community's limited understanding of search stopping

behavior patterns, which was contributed only by Toms and Freund (2009) in the literature before the present study. The patterns are divided into three groups by length: one-SERP repeated patterns, or sequences of moves on the last SERP before task stopping, two-SERP repeated patterns, or sequences of moves on the last two SERPs before task stopping, and three-SERP repeated patterns, or sequences of moves spanning across three or more SERPs before task stopping. One-SERP repeated patterns were more predictive of task stopping among different lengths of repeated patterns because they represented the sequences of moves on the last SERP prior to task stopping in 84% of the tasks. While search engine companies often use a 30-minute window of non-activity in the search logs to identify session boundaries, the repeated patterns observed in this study can serve as additional features for session segmentation.

While in Toms and Freund (2009) the three most common patterns before stopping were observed from the final stage of participants' search tasks, they did not define what "final stage" was in their study, so there are some differences in the methods used in this study and Toms and Freund's study. This dissertation examined search stopping behavior patterns on the last SERP prior to task stopping, the last two SERPs prior to stopping, and the last three or more SERPs prior to stopping, respectively. This fundamental difference prevented comparisons in detail; only higher level similarities and differences can be discussed here. Toms and Freund found that the most common pattern was issuing a query, examining result snippets, and viewing a page, which is similar to RP #1 where participants examined only one search result before ending a task. However, the present study did not include examining search result snippet as a search move. Another similarity between Toms and Freund (2009) and the present work is that they also identified viewing search results beyond the first SERP as a popular behavioral pattern; RP #5 shows that some participants clicked on one more result after they paginated and before they quit

a search task. This dissertation did not observe the third most common pattern reported in Toms and Freund's work: following a link on a landing page before stopping; this pattern was not observed because following links in a landing page was not allowed during the experiment.

In addition, the results also showed that certain treatments were more likely to precede task stoppings than others. Regardless of task length, there were more tasks where the last treatment in a search task was high ISL (Table 21). Inferential statistics also demonstrate that high ISL preceded task stopping more frequently than low and medium ISLs, which indicates the presence of an implicit threshold of number of relevant results to obtain before participants reached the feelings of enough; the more they accumulated relevant search results, the more likely they were to stop.

Similarly, the descriptive statistics in Table 22 showed there were fewer tasks where bursting ISP was displayed immediately before task stopping when task length was two or three. Figure 7 also showed that the bursting ISP led to a slightly higher reformulation rate than the persistent and disrupted ISP. Finding relevant results concentrated at lower ranks seemed to encourage participants to prolong task length. Perhaps the positions of relevant results suggested to participants their query terms were not satisfactory; therefore, they re-issued another query hoping to improve the ranking of results. Even though there were slightly more one-query tasks when the only treatment experienced was bursting ISP than the other two ISPs, the number of SERPs paginated in these bursting ISP search tasks was also higher than tasks where persistent or disrupted ISP were experienced, which also demonstrated that the bursting ISP motivated prolonged result examination for participants who only issued one query in a task.

The number of relevant pages saved or exposed to may potentially be used to predict task stopping as well. Participants saved four documents in 71 tasks out of the 288 tasks completed in

this study, which shows that about 25% of the time four documents was enough for participants to feel enough. In addition, 46 of the 76 one-query tasks were ISP tasks, meaning that participants stopped searching for a task after being exposed to four relevant documents on the first SERP, which again supports the view that four documents might be the magic number to satisfy a search task.

6.2.3 Task stopping rules: old vs. new. Evidence was found in the interviews to support existing stopping rules proposed by previous work. These different stopping rules offered various qualitative and quantitative characterizations of “the feelings of enough”. Some participants articulated a specific number of documents they expected to obtain, while most relied on more abstract rules for determining when the amount of information retrieved was enough. For example, some participants stopped when they felt they had acquired a critical mass of knowledge regarding a topic, some stopped when they could not find anything new, some stopped when their pre-conceived beliefs were confirmed, and still others stopped only after they had searched for every aspect of a topic. The identification of these stopping rules shows that stopping rules used in information seeking contexts also apply to online search tasks.

In the meantime, observations derived from the interviews also confirm that different task types require different stopping rules (Browne et al., 2007) and different amounts of information to reach the feeling of enough (Bates, 1984). When search tasks appeared to be easier to break down into subtasks, participants were more likely to apply the mental list rule, stopping after all aspects of a task were searched. Other tasks such as opinion-based tasks, work-related tasks, and tasks participants deemed important, often entailed examining more results before they ended the tasks.

A new task stopping rule was discovered for opinion-based tasks. Opinion-based tasks, task#5 and task#6, were observed to demand additional efforts in capturing and balancing the “valence” of arguments, or the “pros” and “cons” in the content. During the search process for tasks like this, searchers not only have to judge whether an article belongs to “pros” or “cons”, but also need to evaluate the arguments offered by each side to determine the validity of the article. Moreover, as they search they need to keep track of the number of arguments on each side in order to make unbiased decisions. Displaying retrieved search results by perspectives can better support the need to compare, and offering searchers opportunities to annotate and record retrieved results can alleviate the burden of keeping track of different perspectives. For example, a pin board-like application that allows searchers to sort useful results by self-defined piles can assist searchers in deciding what perspective on a topic is still needed and how much more information for each perspective is needed as they progress through the search process.

VII. Conclusions and Future Work

This dissertation explored the applicability of using information scent and need for cognition to explain query stopping and task stopping. Specifically, four research questions were addressed in this work:

RQ1: What is the relationship between the information scent level of the first SERP and search stopping behaviors?

RQ2: What is the relationship between the information scent pattern of the first SERP and search stopping behaviors?

RQ3: What is the relationship between NFC and search stopping behaviors?

RQ4: How can we model task stopping using interaction signals?

An empirical investigation with 48 participants provided evidence that both information scent and need for cognition are useful constructs for explaining when query stopping and task stopping take place. Participants in this study represented a wide range of professions and age groups, which is advantageous for generalizing the findings to Web search services that serve diverse user populations. The study tasks covered popular topics searched online, which also increases the generalizability of the results to many Web search tasks. More specific findings are elaborated below.

First, the findings demonstrate that participants relied on the information scent of the first SERP to decide when to stop evaluating results and submit a new query. The higher the

information scent level on the first SERP, or the more relevant results on the first SERP, the more participants interacted with the search results and the greater the depth of their search (RQ1). The level of information scent was not the only clue participants used to decide when to reformulate; the pattern of the information scent on the first SERP also affected how deeply participants explored the first SERP for potentially useful information. The disrupted information scent pattern, where all relevant results were placed consecutively from the first to the fourth positions followed by six non-relevant results, probably suggested a lower probability that useful information could be found at the lower end of the first SERP; therefore, participants did not explore search results as far down on the first SERP as when the same number of relevant results were scattered across the first SERP or concentrated at the lower end of the first SERP (RQ2). This finding shows that the best ranking used by most modern search engines supports search efficiency because participants travelled the least distance to obtain the same number of relevant results as the other two patterns; however, such ranking does not motivate deeper exploration in the search result list. Since algorithmic relevance does not always align with subjective relevance, the best ranking can prevent searchers from finding useful results that are erroneously ranked lower.

One of the most significant findings in this work is that NFC played a critical role in explaining why some participants stopped searching earlier than others (RQ3). Participants with lower NFC went deeper in a search result set than participants with higher NFC. Moreover, participants with higher NFC were found to issue more queries during a search task than participants with lower NFC. Specifically, participants with higher NFC exerted more effort in generating query terms than participants with lower NFC. Both findings demonstrate that participants with different levels of NFC used different search strategies to gather information.

Participants with lower NFC tended to search further down the list, while participants with higher NFC tended to reformulate their queries. While this dissertation did not investigate whether moving further down the search results list or reformulation resulted in better search success, there was a positive correlation between the number of SERPs paginated during a task and how highly participants rated their own success in addressing the search task. Future work will compare the quality of results retrieved by those who went further in the search results list with those who chose to reformulate.

There were interaction effects between information scent level and need for cognition which might explain why some participants were more likely to search deeper than others given certain SERP characteristic but not others (RQ1 & RQ3). Even though participants overall preferred to submit multiple queries rather than to examine more SERPs to obtain additional information, the probabilities of query reformulation and pagination under different circumstances could be differentiated by considering of both ISL and NFC. When there was only one relevant result on the first SERP, participants reformulated to a higher probability than when there were five relevant results on the first SERP regardless of NFC. This is to say that displaying one relevant results gave rise to higher probability of reformulation no matter which participant was searching. However, if three relevant search results were available on the first SERP, participants with higher NFC had a higher probability to reformulate than participants with lower NFC. This finding indicates that the effect of NFC is more evident when the first SERP appeared to have moderate quality. Similarly, the probabilities of pagination were similar when the first SERP displayed one relevant result and when the first SERP displayed five relevant results across NFC scores. However, when there were three relevant results on the first SERP, participants with higher NFC were less likely to paginate than participants with lower

NFC. Lastly, the results also show that seeing five relevant results led to a greater probability of task stopping, followed by three relevant results, and last by one relevant result; this suggests that, as participants accumulated more relevant results, they were more likely to end a task.

Individual differences, SERP characteristics and behavioral patterns were found to be predictive of task stopping (RQ4). While this research classified ISL and ISP tasks each into 15 categories by task length and order of treatments, many cells contained fewer than 5 observations, which made inferential statistical analysis impossible. As a result, task length was not considered in the GEE models and only the effects of the first treatment on task stopping and NFC were examined. The results showed that at the start of a task, if search results were dispersed across the first SERP, participants tended to issue more queries than when the search results were placed at the optimal ranking or when they were concentrated at the lower end of the SERP. Participants were also more likely to stop looking for information after they were exposed to a SERP with more relevant results. In addition, participants' knowledge of the topic, search experience of the topic and interest in the topic were also related to the amount of interaction prior to the end of a search task. Lastly, participants appeared to be exhibiting common behavioral patterns before they terminated search tasks; these patterns can be used to predict task stoppings.

Besides information scent level, information scent pattern and need for cognition, this dissertation research identified other factors that could explain when participants reformulated and when they quit a search task. Properties of the search results, queries, search tasks and person all affected when query reformulations took place. Some identified properties of the search results provide additional support for the relationship between information scent level and query stopping, and the relationship between information scent pattern and query stopping

established from the quantitative analyses. Factors contributing to task stopping included the content participants had reviewed, the goal they wanted to achieve, how they felt, and the study constraints. Many factors reflected stopping rules reported in the literature, demonstrating the generalizability of such stopping rules from offline information seeking scenarios to online search tasks.

Even though only 20% of queries in the six study tasks resulted in viewing results beyond the first ten links, participants articulated search scenarios where they were willing to examine more SERPs. These scenarios included when they were searching for people, products, images and literature. Even though the study tasks used in the present study were open-ended tasks, not covering any of the above search scenarios could probably explain why pagination rate was low. Future research should focus on these scenarios where searchers rely on exhaustive result filtering to obtain the desired information. Interface features that are beneficial for completing these search tasks can probably be derived from studying the specific information needs simulated in these scenarios.

Still unresolved is the question of whether some cognitive stopping rules resulted in better search quality than others. For future work, one way to answer this question is by asking human assessors to evaluate how well each document supports task completion. An alternative is to provide assessors the entire set of saved documents of a search task and ask them to rate how well each set satisfies the task requirements. Still another method to assess search result quality is by analyzing the changes in query reformulations. Hassan, Shi, Craswell and Ramsey (2013) found that queries in unsuccessful tasks were often more similar to one another than in successful tasks. Through comparing the similarity of queries in search tasks in which different stopping rules were applied it is possible to derive a relative success score for each search task.

While ISL and ISP treatments were manipulated experimental conditions, the fact that search engines do not necessarily rank search results according to searchers' subjective relevance means that in real search settings searchers may experience these treatments, hence the experimental conditions had ecological validity. In this research, information scent level and information scent pattern were operationalized by the number of relevant results and the distribution of relevant and non-relevant results on the first SERP, yet alternative operationalizations may be considered to study whether the findings from this study hold. It is possible to assign graded relevance scores to search results and manipulate information scent pattern by arranging the order of results of varying scores to reflect a sense of continuation, discontinuation, delay or other patterns of scent. However, making sure participants experience the information scent pattern as expected will be a challenge. Alternatively, information scent of the first SERP can be operationalized by using a fixed DCG@10 while varying both the number of relevant results and the positions of relevant results at the same time.

The use of assigned tasks allowed this research to understand the causal relationships between task characteristics and the task stopping rules applied by participants by keeping as many things as constant as possible. However, it changed the time and effort expended by participants and impacted their stopping behaviors. While this can be viewed as a limitation of the study, it is the case that people commonly conduct imposed search tasks, so the findings do have some external validity and suggest that people impose different stopping rules depending on the origins of the search task. A focused comparison between assigned tasks and self-selected tasks may find differences in the stopping rules used and search result quality. It is also worth noting that the findings of this dissertation are specific to Web search. Search stopping behaviors in the context of databased search may exhibit different characteristics.

While the potential outcomes of early query stopping were discussed at the start of this dissertation, this study does not equate stopping at higher ranks to suboptimal result quality. If a searcher learns something novel from results at higher ranks and is able to apply it to reformulate successful queries, more and better results can possibly be retrieved. While this study did not compare task success between early query stopping and later query stopping, this work was able to demonstrate that NFC, information scent, and other factors affected when query stoppings occurred; furthermore, this study was able to characterize query stoppings by the use of search behavior measures. Especially, RankLastClick, RankLastHover and NumPagination provided evidence that individuals explored search results to varying depths in a search result set, which can be used to inform personalization in several aspects. First, features of search result pages can be flexibly assigned to searchers based on their depth of search. For searchers who tend to examine multiple SERPs, a seamless user experience can possibly be achieved by displaying more search results per SERP or by automatically loading the next SERP to waive the need to click. For searchers who tend to check only results ranked higher and reformulate frequently, query reformulations can be made more efficient by allowing them to highlight to select content they find useful in the snippets or landing pages to provide more context to their initial queries, or by enabling easy access to query suggestions. Secondly, since not all searchers stop searching at equal depths, instead of applying a one-size-fits-all user model for evaluation, search algorithms can be evaluated based on individual search behaviors. Lastly, researchers and practitioners can leverage alternative search result rankings to collect more relevance evaluations in the natural search setting. The deeper a searcher explores, the more implicit and explicit relevance signals a system can capture.

In conclusion, this research has contributed to a better understanding of search stopping behavior by defining two types of search stopping behaviors: query stopping and task stopping, by explaining when search stopping behaviors take place empirically using information scent and need for cognition, by summarizing common behavioral patterns prior to task stopping, and by uncovering factors considered in the decision making processes during search tasks. The findings about query stopping have practical implications for how search engine results pages can be personalized, how search effectiveness measures can be tuned, and how online experiments can obtain more user interaction data. The findings about task stopping have potential practical implications for searching for opinions and identifying session boundaries.

Appendix A

The First Three Queries

Subject ID	Task ID	1st Query	2nd Query	3rd Query
3	1	health risks associated with spray tanning	types of artificial tanning	
12	1	artificial tanning methods	risks of artificial tanning	risks of artificial tanning methods
15	1	artificial tanning	fake tans	health risks of artificial tanning
24	1	different types of artificial tanning		
20	1	what are different methods of artificial tanning		
23	1	bronzer risks	spray tan risks	tanning
1	1	risks of artificial tanning	risks of artificial tanning	types of artificial tanning
4	1	artificial tanning methods	different types artificial tanning	
7	1	artificial tanning methods	tanning spray tanning lotion health risks	
10	1	artificial tanning method	artificial tanning method	artificial tanning risks
13	1	artificial tanning	spray tanning	
16	1	artificial tanning methods	vitamin d	
6	2	ocean pollutants	ocean pollution causes	
9	2	different types of ocean pollutants		
18	2	types of ocean pollution		
21	2	different types of ocean pollutants		
2	2	ocean pollutants		
5	2	types marine pollution	types of ocean pollution	
8	2	what are some different types of ocean pollutants		
11	2	ocean pollutants	ocean pollutants effects	
14	2	list of ocean pollutants	ocean pollutant	ocean pollutants
17	2	examples of ocean pollutants		
19	2	enviromental risk aossiated with ocean pollutants	environmental risk associated with ocean pollutants	what are different types of ocean pollutants
22	2	ocean pollutants	ocean pollutants and their environmental risks	
25	3	extreme sports		
28	3	extreme sports for amateurs	risk of hang gliding	
31	3	list of extreme sports	risks associated with bungee jumpng	risks associated with indoor climbing
34	3	extreme sports for amateurs	extreme sports for beginners	extreme sports risks
27	3	extreme sports for amateurs	f	rock climbing risks
30	3	extreme sports for amateurs	extreme sports for beginners	extreme sports list
33	3	bungee jumping	bungy jumping	
36	3	amateur extreme sports	extreme sports for amateurs	risks bmx
26	3	cliff jumping risk	hangliding risks	kitesurfing risk
29	3	amateur extreme sport activities	google	
32	3	ametuer extreme sports	extreme sports for amteuers	risks of mountain biking
35	3	amateur extreme sports	amateur extreme sports risks	easy extreme sports for amateurs
3	4	methods of tattoo removal		
12	4	tattoo removal methods		
15	4	at home tattoo removal	laser tattoo removal	list of tattoo removal methods
24	4	current methods for tattoo removal		
20	4	tattoo removal methods		
23	4	tattoo removal		
1	4	tattoo removal methods		
4	4	methods to remove tattoos		
7	4	tattoo removal	tattoo removal ip	tattoo removal ipl
10	4	tattoo removal methods		
13	4	crsurgery tattoo	cryosurgery tattoo	dermabrasin
16	4	best tattoo removal method	tattoo removal	
6	5	computer communication social skills	googl	is technology ruining social skills
9	5	computers communication impact	impact of computers on face to face communication	impact of computers on social skills
18	5	effect of facebook on communication	facebook replacing social interaction	
21	5	positive and negative impact of computers	positive impact of computers	positiveimpact of computers

2	5	does facebook effect face to face communication		
5	5	impact of computers on communication and social skills	impact of computers on face to face social skills	
8	5	how does in internet affect people s social skills		
11	5	impact of social media on communication	social media effect on communication	social media impact on face to face communication
14	5	technology communication social skills		
17	5	is social media negative	is social media negative on social skills	
19	5	does use of computers for communication have a positive or negative impact on people s face to face social skills	does use of computers for communication have a postive or negative impact on people s face to face social skills	
22	5	computer communication and face to face social skills		
25	6	violent video games		
28	6	effects of violent video games	effects of violent video games on children	violent video games
31	6	violent video games		
34	6	m rated video games effects		
27	6	effects of violent video games on teenages		
30	6	violent video games and teenagers		
33	6	detrimental effects of mature video games on teenagers	google sc	
36	6	reported effects of violent video games on teens		
26	6	effects of violent video games	effects of violent video games on teenagers	
29	6	reported effects of violent video games on teenagers		
32	6	reported effects of violent videogames		
35	6	effects of violent video games on teenagers		

Appendix B

Entry Questionnaire

Q1 Participant ID:

Q2 Please provide your AGE:

Q3. Please indicate your SEX:

- Male
- Female

Q4 Please provide your CURRENT OCCUPATION:

Q5 Please indicate the highest degree you've earned:

- High school or GED
- Associate's Degree. Major: _____
- Bachelor's Degree. Major: _____
- Master's Degree. Major/Area: _____
- Doctorate. Major/Area: _____

Q6 The statements below describe some different activities that are associated with searching for information online. Please indicate the level of confidence you have in your abilities to execute each activity.

	1 Totally Unconfident	2	3	4	5	6	7	8	9	10 Totally Confident
Identify the major requirements of the search from the initial statement of the topic.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Correctly develop search queries to reflect my requirements.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Use special syntax in advanced searching (e.g., AND, OR, NOT).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Evaluate the resulting list to monitor the success of my approach.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Develop a search query which will retrieve a large number of appropriate articles.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Find an adequate number of articles.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Find articles similar in quality to those obtained by a professional searcher.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Q7 The statements below describe some different activities that are associated with searching for information online. Please indicate the level of confidence you have in your abilities to execute each activity.

	1 Totally Unconfident	2	3	4	5	6	7	8	9	10 Totally Confident
Devise a query which will result in a very small percentage of irrelevant items on my list.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Efficiently structure my time to complete the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Develop a focused search query that will retrieve a small number of appropriate articles.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Distinguish between relevant and irrelevant articles.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Complete the search competently and effectively.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Complete the individual steps of the search with little difficulty.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Structure my time effectively so that I will finish the search in the allocated time.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix C

Instruction Sheet

The current study is focused on understanding how people search online to solve specific information problems. During the study, please imagine that you are at home looking for information. You will be doing six different information finding tasks. For each task, you'll be given a task description and you should use a custom search engine to address it. **There is no time limit on each of the tasks, and no minimum time limit overall either.** So spend what feels to be an appropriate amount of time on each task, until you have collected a set of pages that in your opinion satisfy the information requirement of a task, and then move on to the next task.

On the next screen, you will be presented with a practice search task, and a link to a questionnaire about the task. After you finish answering the questions, you will be directed to the custom search interface. From there, you can start the search by entering words in the search box. **Once you click on a result, the page will open in a new tab. Review the page content, and answer the following question on the screen: Do you want to save the page? Please save pages you think are relevant to the task. Please answer the question based on the content of page (and do not click on any links on the page).** When you feel you have gathered sufficient information for the task, you can click on "Done" to end the task. After completing each task, you will be presented with a link to another questionnaire about the task.

After the practice task, you will complete six tasks following the same procedure. While you search, I will use a piece of software to record the screen of the computer. After you finish all your searches, I will review three of these recordings with you so that I can get a better understanding of your search process.

Please ask me for clarifications now if you have any questions about the instructions.

Are you ready to start?

Appendix D
Pre-Task Questionnaire

Q1 Participant ID:

Here is the description to the task you are about to search for:

“Having heard some of the recent reports on risks of natural tanning, it seems like a better idea to sport an artificial tan this summer. What are some of the different types of artificial tanning methods? How risky are they? Which one would you recommend?”

The following are some questions regarding the search task for you to answer:

Q2 How interested are you to learn more about the topic of this task?

- | | | | | | |
|--------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Not at all
interested | | | | | Very
interested |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Q3 How many times have you searched for information about this task?

- Never
- 1-2 times
- 3-4 times
- 5 or more times

Q4 How much do you know about the topic of this task?

- | | | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Nothing | | | | | Very
much |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Q5 How easy or difficult do you think the search task is?

- | | | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Easy | | | | | Difficult |
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Appendix E
Post-Task Questionnaire

Q1 Participant ID

You just searched for the following task:

"Having heard some of the recent reports on risks of natural tanning, it seems like a better idea to sport an artificial tan this summer. What are some of the different types of artificial tanning methods? How risky are they? Which one would you recommend?"

Q2 How easy or difficult was the search task?

Easy Difficult

Q3 How easy or difficult was it to determine when you had enough information to finish?

Easy Difficult

Q5 How unsuccessful or successful do you think you were at solving this search task?

Unsuccessful Successful

Q6 How unsuccessful or successful was the search engine at finding relevant documents?

Unsuccessful Successful

Appendix F

Exit Questionnaire

Q1 Participant ID:

Q2 For each of the statements below, please indicate whether or not the statement is characteristic of you or of what you believe.

	Extremely uncharacteristic of me	Somewhat uncharacteristic of me	Uncertain	Somewhat characteristic of me	Extremely characteristic of me
I would prefer complex to simple problems.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like to have the responsibility of handling a situation that requires a lot of thinking.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Thinking is not my idea of fun.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I try to anticipate and avoid situations where there is likely chance I will have to think in depth about something.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I find satisfaction in deliberating hard and for long hours.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I only think as hard as I have to.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I prefer to think about small, daily projects to long-term ones.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I like tasks that require little thought once I've learned them.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The idea of relying on thought to make my way to the top appeals to me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I really enjoy a task that involves coming up with new solutions to problems.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Learning new ways to think doesn't excite me very much.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I prefer my life to be filled with puzzles that I must solve.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The notion of thinking abstractly is appealing to me.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel relief rather than satisfaction after completing a task that required a lot of mental effort.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It's enough for me that something gets the job done; I don't care how or why it works.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I usually end up deliberating about issues even when they do not affect me personally.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix G
Recruitment Email

To : UNC staff mailing list

Cc : dianek@email.unc.edu

Subject : Adult research subjects needed for Web search study

----- Message Text -----

I need adult volunteer research subjects to help me study Web search behavior. To qualify, you must be 18 years or older, be a proficient English speaker and have at least two years of online search experience.

This study takes approximately **1-1.5 hours to complete** and you will receive **\$20.00 cash** for participating!

This study will take place in a lab in Manning Hall on UNC campus. Please email me at wanchinw@live.unc.edu to schedule your participation.

** You will not be offered or receive any special consideration if you take part in this research; it is purely voluntary. This study has been approved by the UNC Behavioral IRB (IRB Study xx-xxxx).

Many Thanks,

Wan-Ching Wu, Ph.D. Candidate

School of Information and Library Science

University of North Carolina at Chapel Hill

Appendix H
 Example Log File

Userld	TaskNum	SerpNum	TaskId	SerplD	PageNum	Query	ClickRan	ClickDoc	ClickDocUrl	IsJudgeSave
30	2	1	6	40	0	violent video games labeled "M"	1	64001	http://articles.cnn.com/2	
30	2	1	6	40	0	violent video games labeled "M"	1	64001	http://articles. yes	
30	2	1	6	40	0	violent video games labeled "M"	2	64002	http://www.sciencedaily.	
30	2	1	6	40	0	violent video games labeled "M"	2	64002	http://www.s. yes	
30	2	1	6	40	0	violent video games labeled "M"	5	64005	http://www.pamf.org/pre	
30	2	1	6	40	0	violent video games labeled "M"	5	64005	http://www.p. yes	
30	2	1	6	40	0	violent video games labeled "M"	8	64008	http://www.medicalnews	
30	2	1	6	40	0	violent video games labeled "M"	8	64008	http://www.m. yes	
30	2	1	6	40	2	violent video games labeled "M"	15	64015	http://www.teenhelp.con	
30	2	1	6	40	2	violent video games labeled "M"	15	64015	http://www.te. yes	
30	2	2	6	44	0	violent video games rockstar	2	64302	http://articles.latimes.co	
30	2	2	6	44	0	violent video games rockstar	2	64302	http://articles. yes	
30	2	2	6	44	0	violent video games rockstar	4	64308	http://www.sciencedaily.	
30	2	2	6	44	0	violent video games rockstar	4	64308	http://www.s. yes	
30	2	4	6		0	video games "rockstar Mature Audie	1		http://chachaskitchen.co	
30	2	4	6		0	video games "rockstar Mature Audie	1		http://chacha. yes	

Appendix I

Example Query-Level Search Behavior Data File

UserId	TaskNum	SerpNur	TaskId	SerpId	SerpTime	NumPagination	NumExamined	DeepestRankClick
2	7	1	1	4	21.177	0	0	0
3	6	1	1	4	65.338	1	2	12
4	5	1	1	4	20.304	0	1	3
5	4	1	1	4	19.083	0	0	0
6	3	1	1	4	118.595	0	1	1
7	2	1	1	4	134.915	1	2	18
8	7	1	1	4	106.824	0	1	1
9	2	1	1	4	26.354	0	1	1
10	7	1	1	4	25.183	0	0	0
11	6	1	1	4	50.477	0	2	3
12	5	1	1	4	105.121	4	2	12
13	4	1	1	4	89.037	1	2	12
14	3	1	1	4	300.336	1	2	12
15	6	1	1	4	28.925	0	0	0
16	5	1	1	4	65.427	0	1	1

Appendix J

Example Task-Level Search Behavior Data File

UserId	TaskNum	TaskId	TaskTime	NumQuery	NumPagination	NumExamined	DeepestRankClick
2	7	1	91.851	2	0	1	3
3	6	1	226.766	4	8	5	32
4	5	1	149.895	4	0	9	14
5	4	1	136.239	3	0	6	7
6	3	1	436.688	4	0	5	11
7	2	1	143.458	1	1	2	18
8	7	1	258.785	2	0	3	4
9	2	1	339.958	5	3	10	39
10	7	1	82.052	2	0	3	4
11	6	1	435.313	6	0	16	28
12	5	1	161.014	2	4	5	16
13	4	1	107.922	1	1	2	12
14	3	1	313.797	1	1	2	12
15	6	1	344.85	4	0	6	8
16	5	1	249.683	2	0	5	6

Appendix K

Estimates of parameters in GEE Models

Parameters Measures	Intercept	Low ISL	Med ISL	NFC	Low*NFC	Med*NFC
Time	97.37	-35.26	5.47	-24.43	15.74	-2.53
NumPagination	0.53	-0.03	-0.03	-0.70	0.10	0.05
Abandonment	2.13	-1.80	-0.21	0.18	-0.21	0.13
NumExamined	1.87	-1.09	0.15	-0.13	0.03	0.03
NumPred	0.45	-0.23	0.03	0.00	-0.06	0.08
NumRele	1.39	-0.82	0.08	-0.10	0.06	0.03
NumNonRele	2.58	-0.61	-0.45	-0.93	0.08	0.52
DeepestRankClick	4.44	-1.69	-0.32	-1.04	0.12	0.56
DeepestRankHover	10.45	-1.01	-0.33	-8.00	1.12	0.73

Parameter	QueryAction (0=reformulation; 1=pagination)	Estimate
Intercept	0	1.3778
Intercept	1	-0.1771
Low ISL	0	6.1153
Low ISL	1	6.0720
Med ISL	0	-1.3491
Med ISL	1	2.7082
NFC	0	-0.2108
NFC	1	-0.0122
NFC*Low ISL	0	-0.5805
NFC* Low ISL	1	-0.7048
NFC* Med ISL	0	0.7213
NFC* Med ISL	1	-0.5301

Parameters Measures	Intercept	Persistent ISP	Disrupted ISP	NFC	Persistent *NFC	Disrupted *NFC
Time	110.27	10.62	2.26	-8.01	-10.70	7.21
NumPagination	0.48	-0.03	-0.01	-0.50	-0.33	-0.03
Abandonment	1.87	0.31	0.13	0.30	-0.08	-0.36
NumExamined	2.20	0.14	0.10	-0.20	-0.18	-0.05
NumPred	0.39	0.01	-0.02	0.06	0.09	0.01
NumRele	1.78	0.15	0.13	-0.24	-0.27	-0.03
NumNonRele	3.42	0.55	-0.91	-0.54	-0.30	-0.23
DeepestRankClick	5.61	0.69	-0.80	-0.74	-0.48	-0.28
DeepestRankHover	10.71	-0.10	-0.71	-5.65	-3.63	-0.06

Parameter	QueryAction (0=reformulation; 1=pagination)	Estimate
Intercept	0	-0.17
Intercept	1	-2.81
Persistent	0	0.94
Persistent	1	4.80
Disrupted	0	0.80
Disrupted	1	1.22
NFC	0	0.46
NFC	1	0.83
NFC*Persistent	0	-0.45
NFC*Persistent	1	-1.36
NFC* Disrupted	0	-0.42
NFC* Disrupted	1	-0.45

Appendix L

Task-Level Search Behavior Measures for tasks with four or more query submissions

Measures	ISL Tasks (n=68)	ISP Tasks (n=46)
Time	554.19 (627.19)	568.90 (595.54)
NumPagination	2.37 (4.19)	2.13 (3.42)
NumExamined	9.49 (4.05)	9.80 (4.57)
DeepestRankClick	24.19 (14.84)	24.65 (12.61)
DeepestRankHover	48.65 (44.05)	45.59 (36.26)
NumRele	6.09 (2.93)	6.57 (3.40)
NumPred	3.21 (2.93)	2.96 (2.77)
NumNonRele	14.71 (12.34)	14.85 (9.61)

REFERENCES

- Agosto, D. E. (2002). Bounded rationality and satisficing in young people's Web-based decision making. *Journal of the American Society for Information Science and Technology*, 53(1), 16–27.
- Alam, M. A., & Downey, D. (2014). Analyzing the content emphasis of web search engines. In *Proceedings of the 37th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '14)*, Gold Coast, Queensland, Australia, 1083–1086.
- Amichai-Hamburger, Y., Kaynar, O., & Fine, A. (2007). The effects of need for cognition on Internet use. *Computers in Human Behavior*, 23(1), 880–891.
- Anderson, L. W. & Krathwohl, D. A. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Aula, A., Majaranta, P., & Rähkä, K. (2005). Eye-tracking reveals the personal styles for search result evaluation. In M. Costabile & F. Paternò (Eds.), *Human-Computer Interaction - INTERACT 2005* (pp. 1058–1061). Berlin: Heidelberg: Springer.
- Axson, D., Yates, S., & Chaiken, S. (1987). Audience response as a heuristic cue in persuasion. *Journal of Personality and Social Psychology*, 53, 30–40.
- Azen, R., & Walker, C. M. (2010). *Categorical data analysis for the behavioral and social sciences*. Taylor & Francis.
- Azzopardi, L., Kelly, D., & Brennan, K. (2013). How query cost affects search behavior, In *Proceedings of the 36th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '13)*, Dublin, Ireland, 23–32.
- Bailey, J. (1997). Need for cognition and response mode in the active construction of an information domain. *Journal of Economic Psychology*, 18(1), 69–85
- Bates, M. (1984). The fallacy of the perfect thirty-item online search. *RQ*, 24(1), 43–50.
- Berryman, J. M. (2006). What defines “enough” information? How policy workers make judgments and decisions during information seeking: preliminary results from an exploratory study. *Information Research*, 11(4).
- Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. *Journal of American Society of Information Science and Technology*, 31(4), 271–277.
- Borgman, C. (1989). All users of information retrieval systems are not created equal: an exploration into individual differences. *Information Processing and Management*, 25(3), 237–251.

- Borgman, C., Hirsh, S., Walter, V., & Gallagher, A. (1995). Children's searching behavior on browsing and keyword online catalogs: the Science Library Catalog project. *Journal of American Society of Information Science*, 46(9), 663–684.
- Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 8-3.
- Brookes, B. C. (1980). The foundations of information science. *Journal of Information Science*, 2(3-4), 125–133.
- Browne, G., & Pitts, M. (2004). Stopping rule use during information search in design problems. *Organizational Behavior and Human Decision Processes*, 95(2), 208–224.
- 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.
- Cacioppo, J., & Petty, R. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116–131.
- Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Review*, 119(2), 197–253.
- Cacioppo, J., Petty, R., & Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48, 306–307.
- Cacioppo, J., Petty, R., & Morris, K. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, 45, 805–818.
- Card, S., Pirolli, P., Van Der Wege, M. M., Morrison, J. B., Reeder, R. W., Shraedley, P. K., & Boshart, J. (2001). Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability. In *Proceedings of the 19th Annual International ACM Conference on Human Factors in Computing Systems (CHI'01)*, Seattle, WA, 498-505.
- Carenini, G. (2001). An analysis of the influence of need for cognition on dynamic queries usage. *Proceeding of the 19th Annual International ACM Conference on Human Factors in Computing Systems (CHI'01) (Extended Abstracts)*, Seattle, WA, 383–394.
- Chaiken, S. (1987). *The heuristic model of persuasion*. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social Influence: The Ontario Symposium* (pp. 3–39). Hillsdale, NJ: Erlbaum.
- Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. E. (2001). Using information scent to model searcher needs and actions on the web. *Proceeding of the 19th Annual International ACM*

Conference on Human Factors in Computing Systems (CHI'00) (CHI Letters), 3(1), 490–497.

- Chi, E. H., Pirolli, P., & Pitkow, J. (2000). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a Web site. *Proceeding of the 18th Annual International ACM Conference on Human Factors in Computing Systems (CHI'00)*, Hague, The Netherlands, 161–168.
- Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Robles, E., Dalal, B., et al. (2003). The bloodhound project: Automating discovery of web usability issues using the InfoScent™ simulator. In *Proceedings of the 21th Annual International ACM Conference on Human Factors in Computing Systems (CHI'03)*, Fort Lauderdale, Florida, 505–512.
- Chiravirakul, P., & Payne, S. J. (2014). Choice Overload in Search Engine Use? *Proceeding of the 32th Annual International ACM Conference on Human Factors in Computing Systems (CHI'14)*, Toronto, Canada, 1285–1294.
- Cohen, A. (1957). Need for cognition and order of communication as determinants of opinion change. In C. I. Hovland (Ed.), *The order of presentation in persuasion*. New Haven, Connecticut: Yale University Press.
- Cohen, A., Stotland, E., & Wolfe, D. (1955). An experimental investigation of need for cognition. *The Journal of Abnormal and Social Psychology*, 51(2), 291–294.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Connolly, T., & Thorn, B. K. (1987). Pre-decisional information acquisition: Effects of task variables on suboptimal search strategies. *Organizational Behavior and Human Decision Processes*, 39(3), 397-416.
- Considine, M. (1994). *Public policy: A critical approach*. Melbourne: Macmillan.
- Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1), 30–41.
- Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness part II. Implementation of the philosophy. *Journal of the American Society for Information Science*, 24(6), 413–424.
- Crystal, A., & Kalyanaraman, S. (2005). Personality variables and information processing on the web: Methodological issues and empirical evidence. *Paper presented at the International Communication Association Annual Meeting*. New York, NY.
- Curşeu, P. (2011). Need for cognition and active information search in small student groups. *Learning and Individual Differences*, 21(4), 415–418.

- Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in web search. *Proceeding of the 25th Annual International ACM Conference on Human Factors in Computing Systems (CHI'07)*, San Jose, CA, 407–416.
- Dalton, M., & Charnigo, L. (2004). Historians and their information sources. *College and Research Libraries*, 65(5), 400–425.
- Das, S., Echambadi, R., McCardle, M., & Luckett, M. (2003). The effect of interpersonal trust, need for cognition, and social loneliness on shopping, information seeking and surfing on the web. *Marketing Letters*, 14(3), 185–202.
- Dostert, M., & Kelly, D. (2009). Searchers' stopping behaviors and estimates of recall. In *Proceedings of the 32nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'09)*, Boston, MA, 820–821.
- Duff, W., & Johnson, C. (2002). Accidentally found on purpose: Information-seeking behavior of historians in archives. *The Library Quarterly*, 72(4), 472–496.
- Fitzsimons, J. (2008). Editorial: Death to Dichotomizing. *Journal of Consumer Research* 35(1), 5-8.
- Ford, N., & D. Miller. (1996). Gender differences in Internet perception and use. *Aslib Proceedings*, 48, 183-192.
- Ford, N., Miller, D., & Moss, N. (2001). The role of individual differences in Internet searching: An empirical study. *Journal of the American Society for Information and Technology*, 52(12), 1049–1066.
- Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond Clicks : Query Reformulation as a Predictor of Search Satisfaction. In *Proceedings of the 22nd Annual International ACM Conference on Conference on Information & Knowledge Management (CIKM'13)*, San Francisco, CA, 2019-2028.
- Haugtvedt, C., Petty, R., & Cacioppo, J. (1992). Need for cognition and advertising: Understanding the role of personality variables in consumer behavior. *Journal of Consumer Psychology*, 21, 205–218.
- Huang, J. White, R. W., & Buscher G. (2012). User See, User Point: Gaze and Cursor Alignment in Web Search. *Proceeding of the 30th Annual International ACM Conference on Human Factors in Computing Systems (CHI'12)*, Austin, Texas, 1341-1350.
- Huang, J., White, R. W., & Dumais, S. (2011). No clicks, no problem. *Proceeding of the 29th Annual International ACM Conference on Human Factors in Computing Systems (CHI'11)*, Vancouver, BC, 1225–1234.

- Jansen, B.J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *IP&M*, 42(1), 248–263.
- Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4), 422–446.
- Józsa, E., Köles, M., Komlódi, A., Hercegfí, K., & Chu, P. (2012). Evaluation of search quality differences and the impact of personality styles in native and foreign language searching tasks. *Proceedings of the 4th Information Interaction in Context Symposium on - IIIX '12*, 310.
- Kahneman, D., Slovic, P., & Tversky, A. (Ed.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, New York: Cambridge University Press.
- Kammerer, Y., Wollny, E., Gerjets, P., & Scheiter, K. (2009). How authority-related epistemological beliefs and salience of source information influence the evaluation of Web search results—An eye tracking study. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam, Netherlands. 2158-2163.
- Kantor, P. B. (1987). A model for the stopping behavior of searchers of online systems. *Journal of the American Society for Information Science*, 38(1), 211–214.
- Katz, M., & Byrne, M. (2003). Effects of scent and breadth on use of site-specific search on e-commerce Web sites. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10(3), 198–220.
- Kaynar, O., & Amichai-Hamburger, Y. (2008). The effects of Need for Cognition on Internet use revisited. *Computers in Human Behavior*, 24(2), 361–371.
- Kehoe, C., Pitkow, J., & Sutton, K. (1999). *Results of GVU's tenth World Wide Web searcher survey*. Retrieved from College of Computing, Georgia Institute of Technology, Graphics Visualization and Usability Center website: www.gvu.gatech.edu/user_surveys/survey-1998-10/tenth_report.html.
- Klöckner, K., Wirschum, N., & Jameson, A. (2004). Depth-and breadth-first processing of search result lists. In *Proceedings of the 22th Annual International ACM Conference on Human Factors in Computing Systems (CHI'04)*, Vienna, Austria, 1539-1539.
- Kraft, D., & Lee, T. (1979). Stopping rules and their effect on expected search length. *Information Processing and Management*, 15(1), 47–58.
- Lin, C. L, Lee, S., & Horng, D. (2011). The effects of online reviews on purchasing intention: The moderating role of need for cognition. *Social Behavior and Personality: An International Journal*, 39(1), 71–81.

- Lin, C., & Wu, P. (2006). The effect of variety on consumer preferences: The role of need for cognition and recommended alternatives. *Social Behavior and Personality: An International Journal*, 34(7), 865–876.
- Lawrance, J., Bellamy, R., Burnett, M., & Rector K. (2008). Using information scent to model the dynamic foraging behavior of programmers in maintenance tasks. *Proceedings of the 26th Annual International ACM Conference on Human factors in Computing Systems (CHI'08)*, Florence, Italy, 1323-1332.
- Loumakis, F., Stumpf, S., & Grayson, D. (2011). This image smells good: Effects of image information scent in search engine results pages. In *Proceedings of the 20th Annual International ACM Conference on Information and Knowledge Management (CIKM, 11)*, Glasgow, UK, 475–484.
- Mansourian, Y., & Ford, N. (2007). Search persistence and failure on the web: a “bounded rationality” and “satisficing” analysis. *Journal of Documentation*, 63(5), 680–701.
- March, J. (1994). *A primer on decision making: How decisions happen*. New York, NY: The Free Press.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2), 246–268.
- Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1).
- Morahan-Martin, J. M. (1998). Women and girls last: Females and the Internet. *Internet Research and Information for Social Scientist Conference*. University of Bristol, UK.
- Nair, K., & Ramnarayan, S. (2000). Individual differences in need for cognition and complex problem solving. *Journal of Research in Personality*, 34(3), 305–328.
- Neuberg, S., & Newsom, J. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, 65 (1), 113-131.
- Nickles, K. R., Curley, S. P., & Benson, P. G. (1995). Judgment-based and reasoning-based stopping rules in decision making under uncertainty. *Working Paper, Wake Forest University*.
- Niu, X., & Kelly, D. (2014). The use of query suggestions during information search. *Information Processing & Management*, 50(1), 218–234.
- Nov, O., Arazy, O., López, C., & Brusilovsky, P. (2013). Exploring personality-targeted UI design in online social participation systems. In *Proceedings of the 31th Annual International ACM Conference on Human Factors in Computing Systems (CHI'13)*, 361-370.

- Oulasvirta, A., Hukkinen, J. P., & Schwartz, B. (2009). When More Is Less : The Paradox of Choice in Search Engine Use, *Proceedings of the 32th Annual International ACM Conference on Research and Development on Information Retrieval (SIGIR'09)*, Boston, Massachusetts, 516–523.
- Palmquist, R. A., & Kim, K. S. (2000). Cognitive style and on-line database search experience as predictors of Web search performance. *Journal of the American Society for Information Science*, 51(6), 558-566.
- Park, L., & Zhang, Y. (2007). On the distribution of searcher persistence for rank-biased precision. *Proceedings of the 12th Australasian Document Computing Symposium*, Melbourne, Australia, 17–24.
- Petty, R., & Cacioppo, J. (1981). *Attitudes and persuasion: Classic and contemporary approaches*. Dubuque: IA: William C. Brown.
- Petty, R., & Cacioppo, J. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 123–205). New York, NY: Academic Press.
- Pirolli, P. (1997). Computational models of information scent-following in a very large browsable text collection. *Proceeding of the 15th Annual International ACM Conference on Human Factors in Computing Systems (CHI'97)*, Atlanta, GA, 3–10.
- Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford: New York: Oxford University Press.
- Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, 106(4), 643–675.
- Pitts, M., & Browne, G. (2004). Stopping behavior of systems analysts during information requirements elicitation. *Journal of Management Information Systems*, 21(1), 203–226.
- Pocius, K. E. (1991). Personality factors in human-computer interaction: A review of the literature. *Computers in Human Behavior*. 7. 103-135.
- Prabha, C., Connaway, L. S., Olszewski, L., & Jenkins, L. R. (2007). What is enough? Satisficing information needs. *Journal of Documentation*, 63(1), 74–89.
- Rieh, S. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information and Technology*, 53(2), 145–161.
- Rodden, K., & Fu, X. (2007). Exploring how mouse movements relate to eye movements on web search results pages. *Web Information Seeking and Interaction*, 29-32.
- Saracevic, T (2010). Information Science. Edited by M. J Bates. *Encyclopedia of Library and Information Sciences*. New York: CRC Press.

- Scholer, F., Kelly, D., Wu, W. C., Lee, S. H., Webber, W. (2013). The effect of threshold priming and need for cognition on relevance assessment. *Proceedings of the 36th Annual International ACM Conference on Research and Development on Information Retrieval (SIGIR '13)*, Dublin, Ireland.
- See, Y., Petty, R., & Evans, L. (2009). The impact of perceived message complexity and need for cognition on information processing and attitudes. *Journal of Research in Personality*, 43(5), 880–889.
- Shaffer, D., & Hendrick, C. (1974). Dogmatism and tolerance for ambiguity as determinants of differential reactions to cognitive inconsistency. *Journal of Personality and Social Psychology*, 29, 601–608.
- Sicilia, M., Ruiz, S., & Munuera, J. (2005). Effects of interactivity in a web site: The moderating effect of need for cognition. *Journal of Advertising*, 34(3), 31–44.
- Simon, H. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118.
- Simon, H. (1971). Decision making and organizational design. *Organizational Theory*. New York: Penguin Books.
- Siochi, A. C., & Ehrich, R. W. (1991). Computer analysis of user interfaces based on repetition in transcripts of user sessions. *ACM Transactions on Information Systems (TOIS)*, 9(4), 309-335.
- Spink, A., & Jansen, B. J. (2006). Searching multimedia federated content web collections. *Online Information Review*, 30(5), 485–495.
- Sundar, S. S., Kalyanaraman, S., & Brown, J. (2003). Explicating web site interactivity: Impression formation effects in political campaign sites. *Communication Research*, 30(1), 30–59.
- Sundar, S., Knobloch-Westerwick, S., & Hastall, M. R. (2007). News cues: Information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology*, 58(3), 366–378.
- Toms, E., & Freund, L. (2009). Predicting stopping behaviour: A preliminary analysis. In *Proceedings of the 32nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '09)*, Boston, MA, 750–751.
- Tuten, T., & Bosnjak, M. (2001). Understanding differences in web usage: The role of need for cognition and the five factor model of personality. *Social Behavior and Personality*, 29(4), 391–398.

- Verplanken, B. (1993). Need for Cognition and external information search: Responses to time pressure during decision-making. *Journal of Research in Personality*, 27(3), 238–252.
- Verplanken, B., Hazenberg, P. T., & Palenewen, G. R. (1992). Need for Cognition and external information search effort. *Journal of Research in Personality*, 26(2), 128–136.
- Warwick, C., Rimmer, J., Blandford, A., Gow, J., & Buchanan, G. (2009). Cognitive economy and satisficing in information seeking: A longitudinal study of undergraduate information behavior. *Journal of the American Society for Information Science and Technology*, 60(12), 2402–2415.
- Wang, P., Hawk, W., & Tenopir, C. (2000). Users' interaction with World Wide Web resources: An exploratory study using a holistic approach. *Information Processing and Management*, 36(2), 229–251.
- Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67(6), 1049-1062.
- Woodruff, A., Rosenholtz, R., Morrison, J. B., Faulring, A., & Pirolli, P. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. *Journal of American Society of Information Science and Technology*, 53(2), 172–185.
- Wu, W. C., Kelly, D., Edwards, A. & Arguello, J. (2012) Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. *Poster presented in the 4th Information Interaction in Context Symposium (IIIX'2012)*.
- Zach, L. (2005). When is “enough” enough? Modeling the information-seeking and stopping behavior of senior arts administrators. *Journal of the American Society for Information Science and Technology*, 56(1), 23–35.