ADAPTIVE SEARCH BEHAVIOR: A RESPONSE TO QUERY FAILURE

by

CATHERINE L. SMITH

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Communication, Information, and Library Studies

written under the direction of

Dr. Paul B. Kantor

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2010

UMI Number: 3397517

UMI®

Dissertation Publishing

ProQuest®

ABSTRACT OF THE DISSERTATION

ADAPTIVE SEARCH BEHAVIOR: A RESPONSE TO QUERY FAILURE

By

CATHERINE L. SMITH

Dissertation Director:
Dr. Paul B. Kantor

When an interactive search system returns a results list that fails to meet its user's information need, the user experiences a *query failure*. With the present generation of search systems, except for the most common and simple information needs, users often encounter query failure. This dissertation examined the behaviors searchers use when overcoming query failure. Specifically, this dissertation compared searches conducted on systems operating at three different levels of performance in a single mixed-model factorial experiment, with system performance as the independent variable. The General Linear Model and planned contrasts were used in an exploratory analysis of the effects of system performance on system responses, search behavior, and searcher productivity. Thirty-six volunteers from the Rutgers University community participated in the study. The study found that when system performance is degraded, searchers increase the pace of query submissions. Inter-query time intervals are shorter when results lists are shorter and when a spelling error message is displayed with a results list. These findings suggest that a system capable of monitoring a user's query submission rate and the characteristics of its own responses may be able to detect and assist a user experiencing a difficult search.

Dedication

This dissertation is dedicated to the memory of my mother, Sara Luverne Floyd Smith, who always let me go barefoot in the summer even though my feet got dirty.

Acknowledgements

This work could not have been completed without the guidance and support of my dissertation committee, Gretchen Chapman, Paul Kantor, Michael Lesk, and Nina Wacholder. I am grateful for their criticism and comments. I give special thanks to my advisor, Paul Kantor, for his guidance and instruction on the ways of research and the challenges of analysis. His support and mentorship have been essential to my development as a scientist. Special thanks also to Nina Wacholder for many far-ranging, engaging, and creative discussions. Thanks also to Claire McInerney for her mentorship throughout my studies, for her great sense of humor, and for knowing just when I needed a laugh. To my classmate, Michael Cole, many thanks for the long, interesting discussions about searchers, systems, and a great many other things. Thanks also to Paulette Kerr for her prayers, buoyant spirit, and for teaching me the Jamaican way to "*big up myself*".

Last but not least, I thank my family for their enduring support. My husband Kay Whitefield picked up his life, moved to New Jersey, and supported me materially and spiritually over the past 6 years; thank you for your love, which has carried the day on many occasions. My sons Joseph, William, and Thomas rolled happily along; thanks for your inspiring curiosity and for all your laughter. Finally, thanks to my father, Joe, for his counsel on the joys and challenges of experimentation, and for his sage observations and advice.

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION

1.1 MOTIVATION FOR THE STUDY

An ideal information retrieval system would return to its user a list of all documents that cover all aspects of its user's information need, and only those documents. The list would be ordered according to how well each document met the user's need, with the "best" documents at the top, so that each document was as least as good as any document below it on the list. Of course, a user's information need may be satisfied without the system reaching this ideal performance. For example, if one document is sufficient to satisfy the searcher's need, and that document appears at the top of the list, documents lower on the list are superfluous and their order is irrelevant. In this case, the system's performance is satisfactory for the user, but not necessarily *ideal*.

When a system returns a list that fails to meet its user's information need, the user experiences a *query failure*. Query failures occur when there is a breakdown in the processes that determine how the user's need is related to the documents in the system's collection. The failure may arise in either, or both, of two types of processes[1]: (1) the user's cognitive processes for expressing the information need in a query, and (2) the system's internal algorithmic processes.

1) In the first instance, a user may fail to express his or her information need adequately. For example, a user submitting the query "investing in bulls" might receive documents containing information about *investing during a downturn in financial markets* when the information need is related to *purchasing cattle*. In

---

[1] We assume a rational searcher who is capable of determining whether a document meets his or her information need.

order to overcome this type of failure the user must restate the query in a way that disambiguates its meaning.

2) In the second instance, a user may state his or her need unambiguously, and yet, the system may fail. The failure may be due to poorly designed algorithms such as those used for matching documents to queries. For example, if a user seeking information about *investing in cattle* enters the query "investing in cows," the system might return documents about *investing in a financial services company called "Cattle, PLC."* For the user, recovery from this second type of failure *also* requires restatement of the query[2].

From the user's point of view, whatever the cause of query failure, the consequence is fundamentally the same; the user must *solve the problem of how to improve the query*.

Barring a gross and obvious error in the query, such as typing *buils* for *bulls*, a user may be unable to discover why the failure has occurred. Indeed, the cause of query failure may be indeterminate precisely because a statement of information need (the query) is *optimal only in its relation to the search system*, and the user may have a poor understanding of that relation.

When a query fails, a user may (a) suspend the search (e.g., pause temporarily, turn to another system, quit altogether), or (b) continue the search by submitting a revised query. If the search continues, the user changes the query according to his or her beliefs about the deficiencies in the prior query. With each query, the response from the system provides the user with information about how the system works. By using the system

---

[2] We recognize that a system may fail because the desired information does not exist. When this happens, an ideal system would return an empty list to its user, along with a message indicating that the information does not exist within the system. We consider failure to return this ideal response to be a failure of the system's internal algorithmic processes. Of course, recovery from this type of failure requires that the user select a different system before resubmitting the query. Our study does not address the problem of selecting an alternative system, or the searcher's solution to this problem.

repeatedly the searcher learns how to induce the system to produce sufficiently useful search results[3]. The user learns the relationships among information needs, statements of need (queries), and the system's internal processes. Ideally, over the course of many searches the user perceives and learns the *regular features* of these relationships. Those regularities allow the user to develop efficient cognitive procedures for routine aspects of interactive search (Anderson, 1998).

Of course, information needs exist within the larger context of a task and its associated goal(s). When a task generates an information need, information search becomes a sub-goal of that task. In turn, this sub-goal generates the problem of selecting the best information source for the need. For example, a searcher may select a bookstore, an online social information source (e.g., instant messaging, Twitter), or a face-to-face encounter with someone close by (e.g., asking a passerby on the street). When an interactive text-based search system is selected as the information source, the problem of *optimizing the query* emerges as a sub-goal of the information search task. If the system does not provide the desired information in the results returned from the initial query, the searcher then experiences the problem of query failure (Anderson, 1998; Card, Moran, & Newell, 1983; Newell & Simon, 1972).

For the current generation of search systems, there are two basic types of designs for avoiding or repairing query failure. A *learning-system* design involves detection of contextual information (e.g., about the user or the task), which is used to automatically optimize the query with the objective of returning satisfactory results. A *learning-user* design involves presenting meta-data to the user, with the objective of conveying information about relationships between information needs, queries, and the system's

---

[3] Presumably, if this cannot be learned, the searcher will eventually stop using the system.

internal processes. Interactive search systems may use neither, either, or both of these design approaches.

In a *learning-system* approach, *the system* learns about the relationship between the context in which a query is produced and the needed information. Context is relevant because the same set of words may represent different information needs, for example, a "cone" may be a pine cone or ice cream cone (Lesk, 1986). In its ideal form, a system of this type knows enough about any query's context that it can always return a satisfactory set of documents for any initial query. Examples of designs that pursue this goal include personalization, query learning, and query augmentation (Bruza & Dennis, 1997; White, Ruthven, & Jose, 2002; Xu & Croft, 1996). This approach is likely to be highly efficient for recurring, simple needs, because a system of this type learns by observing many examples of query/context pairs. For complex, non-routine needs, query/context pairs are rare or unique, and a learning-system will have little information with which to learn.

In contrast, the *learning-user* design approach assumes that *users* can better represent their needs in queries when *they* learn the relationships between needs, queries, and the system's response processes. This approach has two basic forms. In a *query-focused design*, the system *asks the user* to state the request in a form that matches the terminology and syntax of the system's optimal response process. In doing so, the system provides evidence of its internal processes, and demonstrates how those processes relate to queries. Examples of this approach include controlled vocabularies (Liu & Wacholder, 2008), ontologies (Muller, Kenny, & Sternberg, 2004), interactive query term suggestion (White & Ruthven, 2006), structured queries (Goncalves, et al., 2004), and "advanced" search menus (Google, 2008). In contrast, a *display-focused design* presents explicit

information about (1) the corpus over which the system operates, (2) the system's response process, and/or (3) the relationship of both to the query. Examples of this approach include the graphical display of search results (Spoerri, 2006), faceted displays (Hearst, Baeza-Yates, & Ribeiro-Neto, 1999; Hearst, et al., 2002), contextual displays (Dumais, Cutrell, & Chen, 2001), cluster displays (Kural, Robertson, & Jones, 2001), browsing displays (Zhang & Marchionini, 2005), and other forms of information visualization (Leuski & Allan, 2004). In its ideal form, the learning-user approach teaches users how to induce the system to return satisfactory results for *any* information need. In effect, the ideal system of this type teaches its user how to solve and prevent the problem of query failure. Developing this type of system requires understanding how users currently solve the problem of query failure.

In the present generation of search system designs, except for the most common and simple needs, users often encounter query failure. Searchers have learned to overcome failure in their daily use of systems such as Yahoo!, Google, and Microsoft's Live Search, relying on skills and habits learned over the course of many searches. While many aspects of search behavior are increasingly well described in a growing body of literature, there has been relatively little focus on how searchers overcome query failure. This dissertation addresses that gap.

1.2 RESEARCH OBJECTIVES

The objective of this dissertation is an exploratory analysis of how users change their interaction with the system when system performance is degraded. More specifically, we focus on describing (1) changes that are observable by both the system and the searcher (the system's *responses* and the searcher's *behavior*), and (2) changes

that the system cannot observe directly, but that the searcher experiences (the searcher's

productivity). The study addresses the following questions:

When system performance is degraded:

- How do observable system responses change?

- How does search behavior change?

- How does searcher productivity change?

- How are system performance, system responses, and search behavior

  interrelated?

From a practical point of view, the answers to these questions will help in the

design of more effective search systems. We envision a system that monitors its users'

behaviors and its own responses, with the goal of detecting query failure. If a system can

detect query failure, it can change its responses to be more helpful. Helpful responses

would teach users how to express their information needs more effectively. In order to

design this system, we need to understand how the system's responses affect searchers'

behaviors. This dissertation contributes to these larger research goals.

1.3 RESEARCH APROACH

We have approached our research objectives using a factorial experiment. System

performance is our independent variable. The experiment compares searches conducted

on systems operating at three different levels of performance: a *standard* level and two

degraded levels. We observe search behavior, system responses, and searcher

productivity as dependent variables under these three performance conditions. Our goal

was to give our subjects the experience of a difficult search conducted on a poorly

performing system. We did this with experimental treatments that produced high rates of

query failure, and experimental topics that were informational and complex. We know from prior research that larger effects can overwhelm the relatively small effect of system performance. Our experiment is designed to control for these incidental effects, including the effect of the subject, the search topic, and the position of a search (1st, 2nd, 3rd, etc.) during an experimental session.

Because our objectives focus on how *behavior* changes when searchers solve the problem of query failure, in this experiment we are not concerned with what searchers can verbalize about how they overcome query failure. We want to observe search behavior that is unaffected by a demand that the subject reflect on the process. For this reason, we have not asked our subjects to "think aloud" during the experiment. We record, unobtrusively, information about behavior and the system's responses. Our analysis looks for meaningful differences in, and relationships between, system responses, search behavior and searcher productivity, as affected by system performance.

1.4 OVERVIEW OF THE STUDY

The experiment involved 36 subjects, assigned to 3 groups of 12 subjects each. One group was a control group, and the other two were treatment groups. Subjects were assigned a task that involved searching on 12 pre-defined topics, which we assigned in a balanced order. The topics were administered in 3 blocks of 4 topics each. For the first block of 4 searches, every subject used the same system, which performed at a standard level. During the second block of 4 searches, subjects in the two treatment groups used systems that we degraded intentionally, while subjects in the control group continued to use the standard system. In the third block, all subjects used the standard system. We did not tell subjects about the blocks or about the change in the system.

During the experiment, we asked subjects to pretend they were working for journalists as they completed the experimental task. Their assignment was to find as many "good" information sources as possible and to avoid "bad" sources, while using their own definition of a "good source." Subjects identified good information sources by clicking a checkbox on the search system interface. Subjects competed for a chance to win a $40 bonus. Only the experimental mock-up of the Google system could be used for searching. There were no limits on time or on the number of queries that could be used.

Subjects worked on 12 assigned topics, which were presented as statements. We designed each statement to contain lexical ambiguity, that is, every topic contained words with more than one meaning (for example, *tire* meaning *wheel* and *tire* meaning *fatigue*). This made it easy for subjects to retrieve results for a topic related to an alternative meaning of a word, and unrelated to the assigned topic. Our goal was to induce difficulty of a similar type and level among our subjects and across the assigned topics.

We produced the two degraded systems by displaying different parts of Google's theoretically infinite results lists. Results were always displayed as if they were from the top of Google's list. The experiment ran on a proxy server that manipulated each query before it was sent to Google and processed results returned from Google. Prior to display in the experimental interface, all results lists returned from Google were scraped[4], parsed, manipulated, and stored, with advertising and sponsored items removed. The interface looked like Google, except that checkboxes and control buttons were added, and the active topic statement appeared in an upper frame. Each results page contained no more than 20 items and there was no option to continue to a next page of results.

---

[4] "Scraping" is a process that extracts data from a page of *html*.

Throughout each experimental session we captured detailed information about search behavior and system responses. After all experimental sessions were completed, a quality rating was given to each information source identified as "good" by subjects. We used these ratings to confirm that our manipulation of results lists produced degraded performance, and to analyze the productivity of searchers. We conducted our analysis using the General Linear Model and planned contrasts.

Our key finding is that when performance is poor, the pace of query submissions increases. This finding is important because a system that can monitor a user's query submission rate and detect a change in that rate may be able to detect a difficult search and offer assistance to the user. We note, however, that the change in query submission rate we found in our study has been detected in the mean of the query rate over blocks of four searches. For a system designed to detect this change in behavior, the difficulty lies in detecting a meaningful change in query rate during a single search session.

1.5 ORGANIZATION OF THE DISSERTATION

The dissertation is organized as follows:

- *Chapter 2* reviews the research in interactive information retrieval that guides our study and the interpretation of our results.

- *Chapter 3* presents the specific research questions addressed by the study.

- *Chapter 4* details our experimental research method.

- *Chapter 5* presents detailed information about our experimental subjects, methods used in preparing our data, the derivation of variables, and the rationale for our analysis.

- *Chapter 6* covers our analysis of the effect of system performance on searcher productivity, search behavior, and system responses.

- *Chapter 7* covers the exploratory analysis of relationships between system responses and search behavior.

- *Chapter 8* discusses our findings and conclusions, the limitations of the study, and future work.

2. LITERATURE REVIEW

In this chapter, we review research drawn from the Information Science literature describing interactive search behavior. We focus on factorial studies, analyses of search engine and browser logs, and predictive models derived from log data. We start the chapter by introducing the broad context that motivates much of the research: the development of the *ideal* search system. The section that follows presents the structure of the literature review.

2.1 INTRODUCTION

The ideal search system would predict, with great accuracy, the value of every document in its collection, for every possible query. Of course, the value of a document is in the mind of the person who needs information, so the ideal system would, in effect, predict those values for each of its users, on any occasion, for any information need. The system would then present the documents in the order of their predicted value, with the most valuable document first[1]. Traditionally, we define search system performance as a measure against this idealized goal.

Typically, in a research setting, we measure performance by asking judges to assess the relevance of retrieved documents for a topic, using relevance judgments as a proxy for predicted value. We calculate performance by comparing the system's predictions of relevance with the assessors' judgments. Naturally, system designers are keenly interested in methods for accurately predicting users' relevance judgments. In pursuit of this goal, researchers study the relationship between *user behavior* and relevance judgments, with the following reasoning: Behavior is evidence of a user's state

---

[1] Of course, we also assume that the system can compute a minimum threshold for *value* for every possible query; documents with a predicted value below this threshold would not be presented to the user.

of mind (the user's information need and related relevance judgments). A system can capture this evidence by observing behavioral indicators. Ideally, the system monitors these indicators to better understand its users' information needs and thereby improve its relevance predictions. This ultimate objective of predicting document relevance motivates much of the recent research on behavior, including the modeling of user behavior. Generally, these efforts focus on *learning-system* approaches to design.

While relevance prediction remains the underlying motivation for many studies, the challenges of behavioral modeling have produced many interim research goals. Principal among these is a model capable of predicting a searcher's next interactive *behavior*. Two essential research questions have arisen from these goals: (1) How does the user's mental state affect interactive behavior? and (2) How do the system's responses affect behavior? We review studies that address these questions.

Interactive search behavior has been studied since the 1970s. Early studies investigated interaction logs from electronic library catalogs (see Hunter, 1991 for an excellent review of early work); the analysis of library system logs continues to this day. Typically, these studies involve demographic analysis, descriptions of usage patterns for system functions and query operators, visual inspection of query patterns, and descriptions of the observed patterns. Because our study focuses on web search engines, this literature review excludes studies of library systems and their users.

Studies of interactive search behavior are, of course, constrained by the source of the data analyzed. Generally, data has been recorded for three types of behavior: (1) *query formulations*, (2) *interaction with results list displays*, and (3) *interaction with open documents*. Early research technologies generally recorded only one, or perhaps two,

types of behavior. As data-gathering techniques have advanced (e.g., search engine logging, instrumented browsers), it has become possible to record the naturally occurring behaviors of a large and diverse user population, with simultaneous recording of details about many types of behaviors. In addition, laboratory systems have advanced, making it possible to synchronize ever more detailed log data with detailed visual scanning data. These integrated datasets provide very fine-grained descriptions, making it possible to analyze and model interleaved sequences of all three types of behavior. The literature reviewed here uses, primarily, search engine data and covers both early and recent work on the three types of interactive search behavior.

2.2 ORGANIZATION OF THE LITERATURE REVIEW

Next, we describe the organization of the literature review. Table 2.1 places each section of the review in the context of the above discussion. We return to this structure in summarizing the literature at the conclusion of the chapter.

Table 2.1 Organization of the literature review

Searcher's mental state	Three types of **search behavior** (Section 2.3)			System responses
Effect of task-type (Section 2.4.a)	Interaction with open documents (Section 2.3.a)	Interaction as query formulation (Section 2.3.b)	Interaction with results lists (Section 2.3.c)	Effects of: system performance (Section 2.4.b)
Factors affecting search behavior (Section 2.4)				

In the next section of the chapter (2.3), we present findings on the three types of interactive behavior. The earliest attempts to develop quantitative models of behavior focused on interaction with open documents, including *reading, scrolling, printing, bookmarking,* and *saving*, which researchers term "implicit indicators of relevance"

(Kelly & Teevan, 2003); these studies are reviewed briefly in Section 2.3.a. A smaller number of studies have attempted to use query formulation to infer a searcher's information need, or to predict search behavior; we present these studies in Section 2.3.b. Section 2.3 concludes with recent work on interaction with results lists, focusing on studies of visual scanning and "clicking" behavior, in Section 2.3.c.

The next section of the chapter (2.4) reviews experiments in which researchers manipulate the searcher's mental state and study resulting changes in behavior. This has been done by varying the type of task assigned to searchers (in Section 2.4.a), or by varying the performance of the system and the system's observable responses (in Section 2.4.b). In contrast with work that attempts to infer the user's state from behavioral evidence, these experiments provide direct evidence of relationships between behavior and the experimental factors in the searcher's experience.

The next section (2.5) reviews two query-log studies, both of which use a combination of behavioral evidence, and evidence from the system's own responses, in models that predict a searcher's next action. These two studies are particularly important here because they examine the predictive power of temporal features of search behavior. The main finding of this dissertation focuses on changes in the temporal dynamics of search behavior as a response to query failure.

The chapter ends by summarizing the studies and findings reviewed (in Section 2.6). Together, the research we review forms the outlines of an emerging sketch of interactive search behavior and its relationship to characteristics of the search system. We conclude by situating our study within this context.

2.3 THREE TYPES OF SEARCH BEHAVIOR

2.3.a Interaction with open documents

The earliest work on interactive behavior with *full-text* information retrieval systems focused on the user's interaction with retrieved electronic documents (Morita & Shinoda, 1994). These studies examine behavior after a user finds and opens a document or webpage in order to investigate its content. The user's subsequent interaction with the document is hypothesized to be an implicit indicator of the value or relevance of the document. The following behaviors have been studied most often: *dwell time*, the elapsed time between opening a document and closing it, (Kelly & Belkin, 2004; Kim, Oard, & Romanik, n.d.; Konstan, et al., 1997; White, et al., 2002), *scrolling* (Kelly & Belkin, 2001), and *saving*, *printing*, and *book-marking* documents (Oard & Kim, 1998). Oard and Kim (2001) identified four types of document-focused behaviors: examine, retain, reference, and annotate. Kelly and Teevan (2003) provide a comprehensive review of studies published through 2002.

Early work on document interaction produced mixed results. Researchers found that the effects of a user's task, and an individual's behavioral predilections, made document interaction an unreliable indicator of value or relevance. Generally, these early studies examined each interaction type (e.g., scrolling, dwell time) in isolation, and the amount of data available for analysis was quite limited. Recently, highly detailed browser logs have made it possible to examine document interaction in the context of other search behaviors, using a large number of examples; we present one such study in section 2.5.

2.3.b Interaction as query formulation

A small number of quantitative studies have examined query formulation as a type of search behavior[2]. In work attempting to infer characteristics of a user's information need from behavioral evidence, *the length* of a query has been found to suggest the *level of specificity* needed by a user (Lau & Horvitz, 1999; Phan, Bailey, & Wilkinson, 2007); in these studies query *specificity* was defined subjectively by the researchers. Short queries are more likely to be associated with broad or general information needs; however, precise or specific information needs may be expressed by a query of any length. These findings suggest that in isolation, query length is a weak indicator of the specificity of a user's information need.

Downey, Dumais, and Horvitz (2007a) used a large search engine query-log (10 million queries from over 250,000 users) to examine differences in search behavior in relation to *common* and *rare* queries. A rare query is any query submitted to the search engine no more than once in a seven-day period; all other queries are defined as *common*. Table 2.2 presents findings from the study. Relative to common queries, rare queries are less likely to result in a *click-through*[3] and are more likely to result in query reformulation (altering the query and resubmitting it). When a user reformulates a

Table 2.2 Probability of next action for rare and common queries
(from Downey, Dumais, & Horvitz 2007a)

		p(next action)		*p(rare query | reformulation)*
		click	reformulation	
initial query type	rare	.50	.45	.84
	common	.58	.33	.50

---

[2] Other studies have examined the content and linguistic properties of queries (e.g. Bruza & Dennis, 1997; Rieh & Xie, 2006). These aspects of query formulation involve internal processes of memory and word association. Our literature review does not cover these studies.

[3] A *click-through* occurs when a searcher clicks on an active hyptertext link (url) in the search results. A click-through opens a webpage or document so that the user can examine and interact with it.

common query, there is a 50% chance that the new query will be a rare query. When a user reformulates a rare query, it is very likely (84%) that the new query will also be rare. This finding suggests that the characteristics of a searcher's prior query may predict, with some probability, the searcher's subsequent query behavior. Of course, searchers generally have no way of knowing whether their queries are rare or common. These differences are of interest for modeling behavior.

Two query-log studies have found that the *temporal dynamics of query submission* are predictive, with some probability, of future behavior. Lau and Horvitz (1999) created a model that infers a user's next behavior, using inter-query time intervals and characteristics of the prior query. Downey, Dumais, and Horvitz (2007b) found that the time interval between queries is predictive of click-through, submission of the next query, and termination of a search. These integrated models use several types of search behavior; we review the studies in detail in Section 2.5, below.

2.3.c Interaction with results lists: visual scanning and click-through

Early work on interaction with results lists tested the idea that click-though behavior might be used to infer the relevance of a clicked document, with the goal of improving system performance (Joachims, 2002; Kemp & Ramamaohanarao, 2002; White, Jose, & Ruthven, 2001). While click-through behavior in isolation is valuable evidence of a searcher's *expectation of relevance*, it was found to be an unreliable indicator of document relevance (Fox, et al., 2005); searchers often click on sources that are not relevant and they often fail to click on relevant sources. The promise of modeling document relevance from click-through evidence led to efforts to improve understanding of how people interact with results lists. Eye-tracking, which has a long history in

research on reading and visual attention, has been used to gain insight into these interactions.

The earliest eye-tracking studies simply described the duration of visual attention to various parts of a results page. Recent work has investigated the order in which users examine *captions*[4] on ranked results lists, and the amount of visual attention (measured as fixation duration) given to each rank position. Other studies relate measures of visual attention to click-through behavior. The main objective of these studies is to describe the visual behavior that occurs prior to a click-through. While much of this work continues to focus on the goal of predicting relevance from click-through behavior, these studies also reveal that users have developed strong habits in their interactions with ranked results lists.

2.3.c.i Effect of visual display characteristics

Most eye-tracking studies have not controlled for the visual characteristics of results lists. This is an important concern because the two studies that have investigated this issue show that visual characteristics affect user behavior. In a query-log study, Clarke, Agichtein, Dumais, and White (2007) found that caption features affect the probability that a user will click a caption. Users are less likely to click a caption that: (1) contains fewer of the terms used in the query, (2) has a shorter snippet, (3) contains text with lower readability, or (4) has a longer or more complex *url*. In an experimental study, Cutrell and Guan (2007) found that scanning and clicking behavior is affected differently by the length of a caption snippet, depending on the type of search being conducted; the

---

[4] A *caption* is the visual format of an item on a search engine results list. A caption has three components: (1) *title* – a single line of blue underlined text that is a live hyperlink to the underlying information source, (2) *snippet* – two or more lines of text that exemplify the information content of the underlying source; the snippet is not a live link, and (3) *url*, displayed in green and also not a live link. Currently, all three dominate search engines (Google, Yahoo!, and Live Search) format their captions in this way.

study is discussed in detail below. Because most eye-tracking studies have not controlled for the effect of caption features, it is too early to draw definitive conclusions about the *details* of the visual scanning process for results lists, however, the effect of caption *ranking* is well established.

2.3.c.ii Effect of caption ranking

It is well established that when searchers scan a ranked list, they use the rank position of a caption as a cue to the expected relevance of the underlying information source (Cutrell & Guan, 2007; Granka, Joachims, & Gay, 2004; Guan & Cutrell, 2007; Joachims, et al., 2005; Klockner, Wirschum, & Jameson, 2004; Lorigo, et al., 2006). Searchers focus their visual attention and click on captions according to these expectations. The top two captions on a results page are fixated more frequently, and are fixated for a longer period than are any other rank positions. The top caption is particularly privileged by the user; it is clicked with the highest frequency and it is more likely that it will be clicked even if the 2nd caption is more relevant than the 1st.

In the first detailed experimental study on the subject, Joachims, et al. (2005) examined the relationship between search behavior, relevance, and the rank position of a caption. When the 1st caption on a results page was more relevant than the 2nd, and the subject clicked on either of the top two captions, 95% of clicks were made on the 1st caption. Of course, this is a reasonable response. In contrast, when the 2nd caption was more relevant than the 1st, and the subject clicked on either of the top two captions, 72% of clicks were made on the less relevant 1st caption. The authors termed the tendency to click on the first caption, even when a more relevant caption appeared below it, a "click-through trust bias" for the top position.

The order in which users scan captions below the top two positions, and factors that affect scanning order, are not well understood. While most studies suggest that scanning usually proceeds from top to bottom in rank order, it is also clear that scanning patterns are more complex. Joachims, et al. (2005) found that for half of all cases, searchers scanned the caption directly below a clicked caption prior to the click. In a detailed analysis, Lorigo, et al. (2006) found that many searchers do not use a linear (top-down) scanning strategy exclusively. In one third of cases, when a searcher clicked a caption not all of the captions above the clicked caption had been scanned. Only one fifth of scan-paths analyzed were strictly linear, where captions were scanned in the exact descending rank order (with no skips or scans of a previously scanned caption). Klocker, Wirschum, & Jameson (2004) also found that many searchers (35% in one experiment, and 48% in another) employ what they termed a *breadth-first* scanning strategy, in which visual attention returns to a previously scanned caption higher on the list.

The above findings show that searchers have developed visual scanning and clicking patterns that reflect the dominant statistical property of ranked lists: over the long run, the probability that a caption will be useful decreases monotonically with its position on the list. Searchers match their attention to this expectation, and focus their visual attention and interactions at the top of the list. On average, attention decreases monotonically for items lower on the list.

These findings are pertinent to our study, because they describe what searchers are doing in the time interval between query submissions. Because our experiment has not collected data on visual attention and click-through, these insights inform the interpretation of our findings.

While rank position is a dominant factor in scan patterns and clicking behavior, other factors are involved as well. We discuss these next.

2.4 FACTORS AFFECTING SEARCH BEHAVIOR

This section reviews additional experimental work on two factors in the searcher's experience: *task-type* and *system performance*. We present the effect of task-type first. In contrast with studies that attempt to infer the user's state from behavioral evidence, experimental studies of task-type manipulate the user's task and then measure the effect of those manipulations on behavior. The section concludes with studies that examine the effect of system performance on behavior.

2.4.a The effect of task-type

Several studies have examined how behavior is affected by *task-types* as defined by Broder (2002). *Navigational search* is a form of known-item search, where the user's goal is to find a single discrete website. Broder found that approximately 22% of search engine queries are navigational. *Transactional search* involves the search for a product or service; approximately 33% of queries fall into this category. *Informational search* is any other type of search and comprises the remaining 45% of queries. Here we review experimental studies that compare navigational and informational tasks.

Lorigo, et al. (2006) examined the effect of navigational and informational task-types on interaction with results pages and retrieved documents. The study found significant differences in the fraction of *task time* (time spent completing a search task) searchers allocated to results pages and open documents. During informational tasks, in comparison to navigational tasks, searchers spend a larger *fraction* of their task time on, and give more visual attention to, open documents, and a smaller fraction on results lists. During navigational tasks, the opposite is true; searchers spend a larger fraction of task

time on results lists, and a smaller fraction on open documents. The authors suggest that this difference occurs because captions tend to contain the information sought during navigational search, therefore, clicking open a document is less likely to be necessary. This is generally not true for informational search, where a caption is less likely to contain the desired information. Importantly, no significant differences were found in the behaviors searchers used during interaction with results lists. There were no differences in the number of lists viewed, the time spent on each list, the number of fixations on lists, and average fixation duration.

Cutrell and Guan (2007) studied how scanning behavior was affected by (1) navigational and informational task-types and (2) the length of snippet text (short, medium, long), and the interaction of these two factors. Consistent with Lorigo, et al. (2006), they found no significant main effects for behaviors used during interaction with results lists. They did find a significant *interaction* effect for task-type and snippet length. Their most interesting finding occurred when snippet text was longest (6 to 7 lines). During informational search, when snippet text is longest searchers scan fewer captions, but spend more time scanning each caption. Within each caption, visual attention shifts to focus more on the snippet text and less on the title or *url*. With fewer captions scanned, but scanned with more attention, searchers complete informational searches *more quickly* with long snippets. In contrast, for navigational search, when snippets are longest searchers scan *more* captions, giving less visual attention to the *url* and more to the snippet text. As a result, searchers complete navigational search *more slowly* when snippets are long.

The authors suggest that, for navigational tasks, the *url* contains the most relevant information in the caption. During navigational tasks, searchers scan the caption looking for the *url*, but the scanning process is impeded by the length and content of the long snippet text. In contrast, for informational search, the caption text contains highly relevant information. The authors suggest that searchers increase their attention to this text during informational search, and that this change in scanning behavior increases the efficiency of the search.

Guan and Cutrell (2007) examined the effect of two factors on click behavior: (1) task-type (informational vs. navigational), and (2) the rank position of the first highly relevant caption (termed the *target*). The target was the "best" information source for the topic, as judged by the researchers. The experiment examined each subject's interaction with only *the first page* of results received for the first query on each topic. The target was displayed at either the 1st, 2nd, 4th, 5th, $7^{th,}$ or 8th position on the first page of the results list. For navigational search, subjects were very likely to fixate on the target, and click on it, when it was displayed in the 1st or 2nd position (see Table 2.3). For informational search, subjects were very likely to *fixate* on the target when it was displayed in the 1st, $2^{nd,}$ or 4th position, but were very unlikely to *click* on the target when

Table 2.3 Probability of visual fixation and click-through (by task-type and position of most relevant caption)
(from Guan & Cutrell, 2007)

		target position			
		1st caption	2nd caption	4th caption	p(*action* | target at position)
task-type	navigational	1.00	.89	.72	*fixation*
		.78	.83	.39	*click*
	informational	.94	.94	.89	*fixation*
		.89	.33	.17	*click*

it was displayed below the 1st position. This suggests that searchers are influenced heavily by the rank position of the caption when deciding whether to click on it, and that this influence is more pronounced during informational search. It is important to note that this study examined behavior on only the first results page returned during a search.

The above findings are consistent with the notion that search behavior is influenced by, but not controlled by, the position of captions on the results list. The characteristics of results pages (system responses) and the type of information a user needs (searcher's mental state) affect behavior. There are complex dependencies between these factors.

2.4.b The effect of system performance

A small set of studies have examined how system performance affects search behavior in an interactive setting. Two goals motivate this work. One set of studies focuses on analyzing the efficacy of using "batch" techniques, in which no interaction occurs, for the evaluation of interactive systems. Other work focuses on developing models using click-through behavior to predict document relevance. All of the studies provide evidence that searchers change their behavior in response to system performance.

Joachims, et al. (2005) examined the effect of caption ranking and relevance on visual fixation and click-through behavior; results from the same study are also reported in Lorigo, et al. (2008). Their experiment used two levels of system performance: a standard Google system (*normal*) and a degraded system. They produced the degraded system by reversing the order in which the retrieved captions were displayed on a 10-caption results list. The information source estimated by the system to be the best match to the query was placed at the bottom of the list (as the 10th caption). The source estimated to be the 10th–best was placed at the top of the list (as the 1st caption). The

other captions were similarly reordered. The study compared searches conducted in the degraded condition with those conducted in the normal condition. On average, in the degraded condition, subjects scanned more captions (3.8 captions per list vs. 2.5 in the normal list), took more time to scan each list (11 seconds per list vs. 6), were less likely to click any caption on the list (.64 clicks per list vs. .80), and were more likely to click on captions at lower positions (average rank of click 4.03 vs. 2.66). Subjects using the degraded system did not, however, overcome the effect of their rank-based expectations. They were more likely to click on one of the first 5 captions in the top of the list than one of the last 5 captions, and they were less likely to complete their task as successfully (62% vs. 85%). These results suggest that searcher's adapt their behavior in response to degraded results, but that those adaptations are not always sufficient to reach the level of success possible with a normal system. However, in principle, with enough experience, a user might learn to search efficiently with a "reversed" list.

Other studies suggest that searchers *can* effectively adapt their behavior to compensate for poor system performance. Turpin and Hersh (2001) used a question answering task, and assigned subjects to either a low performing system or an enhanced system. System performance had no effect on the accuracy of subjects' answers, but those using the low performing system searched less efficiently, submitting 3 times as many queries to achieve comparable success. Results are similar for an *instance-recall task*[5] (although the difference in the number of queries submitted was not statistically significant). Figure 2.1 summarizes the data from both tasks. The trends suggest that users issue more queries during a search in which system performance is low.

---

[5] In an *instance-recall task*, subjects are asked to find examples (instances) of a category or concept. For example, subjects may search for the birthdates of all U.S. presidents.
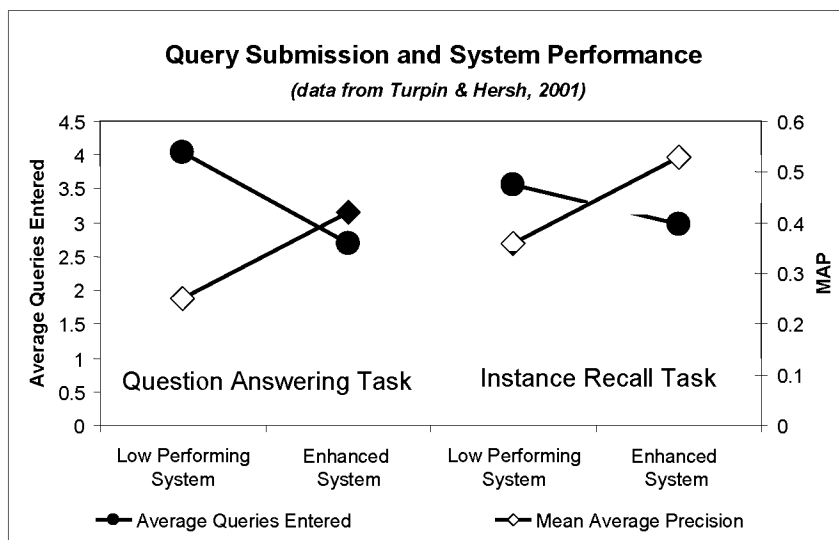
Figure 2.1. Query submission and system performance (from Turpin & Hersh, 2001).

Allan, Carterette, and Lewis (2005) found that searcher productivity was different only at the extremes of poor performance ($bpref^{6}$ < 60%) and superior performance ($bpref$ > 90%); no significant difference was found across the center of the range. In this experiment, searchers highlighted relevant passages from retrieved texts. Searcher productivity was measured as the average number of relevant text passages identified per minute. Error rates (the number of non-relevant passages identified and the number of relevant passages not identified) were not affected significantly by system performance at any level.

Turpin and Scholer (2006) examined how quickly searchers could find a single relevant document (a *target search*), and the number of documents a searcher could find within a five minute time limit. When using degraded systems searchers completed target searches just as quickly as did those using better systems. System performance had little effect on the number of documents found within the time limit.

---

[6] *Bpref* stands for "binary preference." It is a measure of precision in a passage (sections of text) retrieval task. It measures the fraction of total passage material that is non-relevant and that appears ahead of relevant material.

The above findings suggest that users adapt their behavior to compensate for poor system performance. Adaptive behavior is a rational response if, for example, a user has learned through repeated experience that system performance varies considerably depending on the topic of a search (Lagergren & Over, 1998).

2.5 INTEGRATED MODELS OF BEHAVIOR – QUERY-LOG STUDIES

Searchers interact with a search system by submitting queries, scanning results lists, clicking on captions, and reading, bookmarking, saving, and printing open documents. Any or all of these search behaviors may occur during any search. Collectively these behaviors provide external evidence of the internal cognitive processes used by the searcher during the interaction. In theory, this evidence can be used to infer a searcher's mental state and to predict relevance judgments and search action.

Modern web search engines are capable of collecting detailed data on naturally occurring search behavior (*server-side* data capture). These include anonymized identification of a user's browser session, the content of each query submitted, and for each query, the results page returned by the system, including details on each caption, the *urls* clicked on the results page, and time-stamps for each of these actions. For research purposes, server-side data capture may be supplemented with data from actions that occur within an instrumented browser (*client-side* data capture). For example, client-side data might include mouse movements, scrolling, clicks on links in visited websites, *urls* typed, and queries submitted to other types of search systems (e.g., libraries, shopping sites). Collectively, these data are the system's evidence of a user's mental state.

Two integrated models of search behavior have been developed using combined server- and client-side log data and machine learning techniques; we present these models

below. Both are of particular interest because they use the temporal dynamics of query submission in the prediction of search behavior.

The earlier of the two studies (Lau & Horvitz, 1999) used measures of behavior, including temporal features of behavior, to predict a searcher's next action. The study was the first paper to investigate the informativeness of inter-query time intervals (time interval between two queries). The model was developed using supervised[7] machine-learning in a Bayesian-network, using log data for 4,690 queries. The data for each query included query terms, a time stamp, and an anonymized session identifier. To supplement this data, each query was hand-coded by the researcher, to indicate: a) the user's information goal, as inferred by the researcher, and b) query-type. Query-types included *new* (first query on a topic), *reformulate* (reformulated query on the same topic), *page* (move to another results page for the same query), and *off-topic* (the interleaving of two or more unrelated topics). Figure 2.2 depicts the relationship between inter-query time intervals and two types of queries, as defined by the authors: *page queries* and *reformulate queries*.

Because query *pairs* were analyzed (query@time_1 : query@time_2), the graph shows the probabilities for queries that occur within each discrete time interval, or bin. The chart shows the probability that the query@time_2 is of a type, given that it occurs within a time interval. For example, if the 2nd query in the pair was submitted after 10 seconds had elapsed, but before reaching 20 seconds, the query would fall within the 10-to-20 second bin; we read its probabilities from the 10-20 second time interval on the chart. For a 2nd query in this time interval, the probability that it is a page query is 70%,

---

[7] Supervised machine learning occurs when a modeling algorithm receives labeled positive and negative examples of the relationships or classifications it is learning.
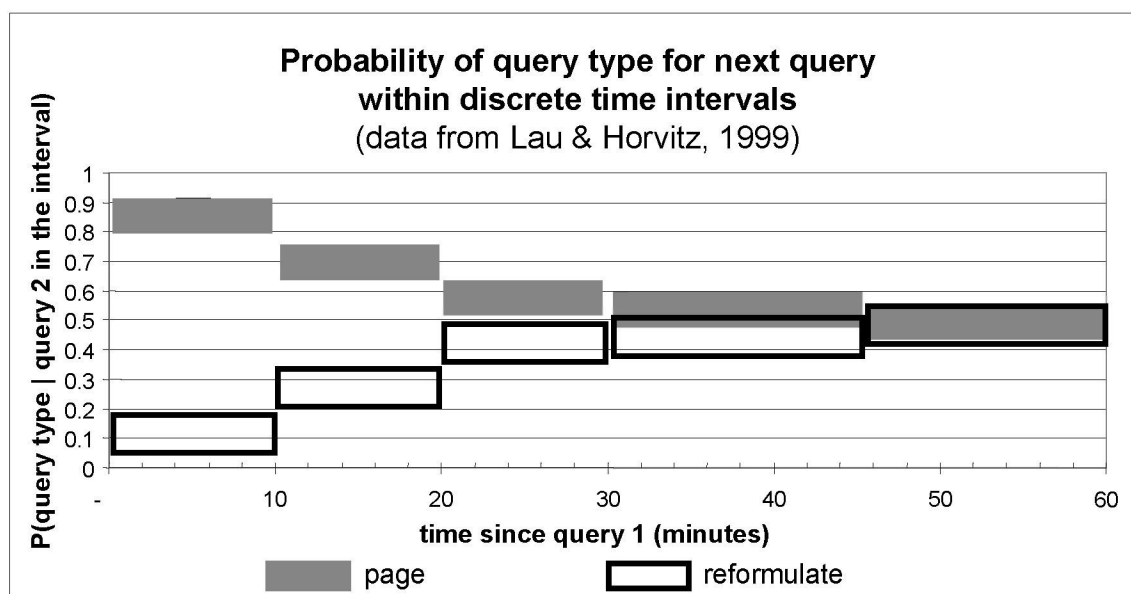
Figure 2.2. Probability of query type (from Lau & Horvitz, 1999).

while the probability that it is a reformulation is 27%. The probability that the next query

will be a reformulation is highest (approximately 0.50) after 45 seconds have elapsed

(after 2 minutes, the probability of reformulation decreases). The probability of a page-

query is highest (approximately 85%) in the first 10 seconds after the first query, and

drops off rapidly over the first 30 seconds. The model uses inter-query time intervals and

query-type coding to predict the *type* of query the user will submit next.

The more recent Search Activity Model (SAM) (Downey, et al., 2007b) also uses

time intervals, among other measures, to model behavior. SAM predicts a user's next

search action from among three possible actions: *query*, *click-through*, or *end session*.

The model was produced using machine-learning over data extracted from highly

detailed logs from more than 250,000 users, including both server-side and client-side

data. The data includes records for three types of events: (1) queries, (2) click-throughs,

and (3) clicks on the browser's back button to return to the search engine after opening a

document or webpage. Each event was associated with a session, and each session with a user. Fifty-one features were extracted or derived from the log data; these features parameterized the events, users, and sessions. Six types of features were used, including: user, search session, query, click-through, non-action, and temporal (for the full list of features, see Appendix A). Several versions of the SAM model were tested against a baseline model, *previous action* (PA), which predicted the searcher's next action using only the immediately prior event (query, click-through, or back-button). The model was developed by adding features to this baseline. The best fit to the data was a version called *SAMLight*.

While the specific details of the SAMLight model are not published, the authors do discuss several key findings from learning and testing the model. First, in training the model, lagged data were used in the computation of action probabilities conditioned on a searcher's prior actions. Testing found that the predictive performance of the model was highest when only the immediately preceding action was used. The inclusion of more than one preceding action actually caused a reduction in performance. Table 2.4 lists the eight most predictive features in the SAMLight model.

The feature with the most predictive power is the elapsed time between two search actions[8], r(SearchAct). Indeed, r(SearchAct) was found to improve the PA model significantly when it was the *only* feature added. The study finds that inter-query time intervals are predictive of both click-through and re-query[9]. Figure 2.3 shows the conditional probability of a next action as a function of inter-query interval, as modeled

---

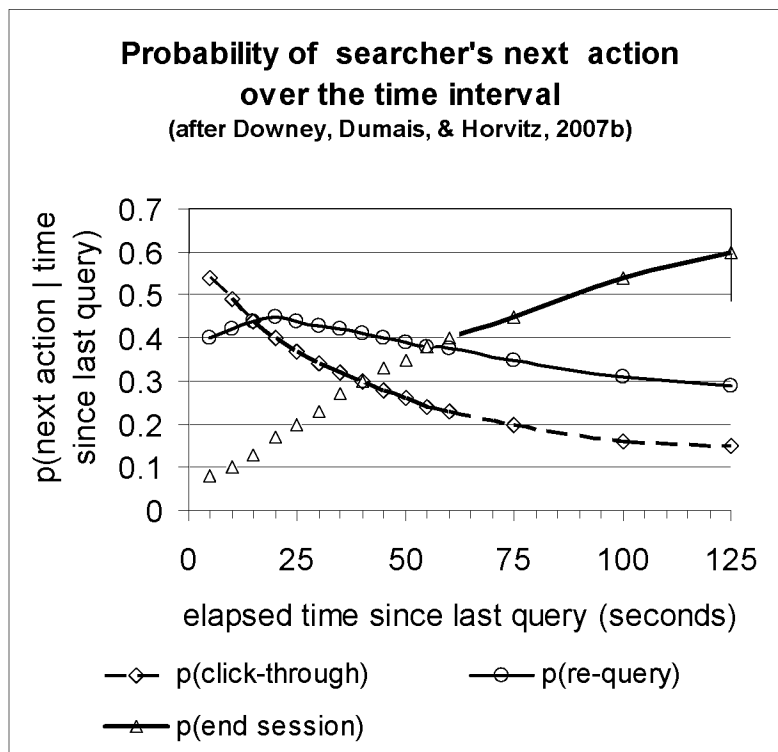[8] Including latency from network transmission and page-load on the user's browser
[9] A *re-query* is a query submission of any type after an initial query, including *reformulation* queries (a query with words that have been changed by the searcher) and *page* queries (a request for another section or page of a results list)

Table 2.4 Top 8 predictive features in the SAMLight click-through model
(from Downey, Dumais, & Horvitz, 2007b)

Feature set type	No. of features	Predictive features (1= most predictive) and relationship to the probability of click-through	
Temporal/transition	4	**1.** r(SearchAct) elapsed time between two search actions[10]	p(*click*) decreases for longer interval between actions
Query	24	**2.** q(FirstResult) rank of first result [on list] requested	p(*click*) decreases for queries that request results lists starting at lower rank positions
		3. q(HasSuggestion) query has spelling suggestion	p(*click*) decreases for query with spelling suggestion
		5. q(HasDefinitive) query has definitive result (e.g., navigation)	p(*click*) increases for queries with definitive result (e.g., amazon.com)
		6. $q(c_r$Prob) probability of a click for the query	p(*click*) increases for queries likely to result in a click
Search session	5	**4.** S(qFrac) ratio queries / search actions	p(*click*) decreases as more search actions are queries
		7. S(Num$q$) number queries entered in session	p(*click*) decreases as more queries are entered
		8. S(Max$q$Words) number of words in longest query submitted	p(*click*) increases as query length increases over the session
User	11	none in top 8	
Results click	4	none in top 8	
Non-action features	3	none in top 8	
TOTAL	51		

---

[10] The r(SearchAct) feature excludes any interval longer than 30 minutes or for which the next action was an end-of-session.

**Probability of searcher's next action
over the time interval**
(after Downey, Dumais, & Horvitz, 2007b)

Figure 2.3. Probability of searcher's next action (from Downey, Dumais, & Horvitz, 2007b).

in SAMLight. Immediately following a query submission, a click-through is the most

likely next action. The probability of a click-trough is at its maximum immediately after

the page is returned (approximately 54%) and as time elapses, its probability drops off

steadily. During the first 15 seconds after a query submission the probability of a re-

query increases. A re-query becomes more likely than a click-through after 15 seconds,

and it remains more likely than a click-through thereafter. Approximately 20 seconds

after query submission, the probability of a re-query peaks at about 45%. After about a

minute, if the user has not clicked or re-queried, it is most likely that the session will end,

and this probability increases with time. The authors do not report predictions based on

the time intervals between click-throughs.

The other most predictive features of the SAMLight model are also characteristics of queries. The probability that a searcher will click on the results page tends to decrease when:

1. the interval since the last action is long (as discussed above),
2. the query requests the 2nd, 3rd, or lower page of results (a starting rank position lower on the list)
3. the query returns a spelling suggestion,
4. successive queries are submitted without intervening non-query actions, (e.g. a click-through or a click on a back-button)
5. the query does not have a "definitive result" (e.g., a common navigational query with a high probability that a specific *url* will be clicked),
6. the query has resulted in few click-throughs when used by other searchers,
7. each successive query is submitted, and
8. longer queries have been submitted previously during the session.

We discuss these features in our summary of the literature.

2.6 SUMMARY AND DISCUSSION

Collectively, the findings presented above form an initial, though incomplete, description of search interaction. We summarize these findings using the structure presented in Table 2.5.

Much of the research covered in our review focuses on the goal of developing systems capable of observing search behavior to model and infer a searcher's mental state (primarily relevance judgments). The length of queries, click-through on results lists, and various forms of interaction with open documents, have been studied with this objective. Section 2.6.a summarizes findings from these studies. A small number of experimental studies have examined the relationship between the type of information needed (a user's mental state) and behavior, including effects on visual scanning, time on task, and

Table 2.5 Organization of summary: factors in interactive search behavior

Searcher's mental state	Query formulation *length of query*	Interaction with results lists *click-through*	Interaction with results lists *visual scanning*	Interaction with results lists *task time*	Interaction with open documents *dwell, scroll, print, save*	Interaction with open documents *task time*	System responses
Inference from behavior to state (Section 2.6.a)							
relevance of information source		X			X		
specificity of information need (general vs. specific)	X						
Effect of searcher state on behavior (Section 2.6.b)							
type of information need (information vs. navigation)		X	X	X		X	
Effect of system responses on behavior (Section 2.6.c)							
		X	X	X			*snippet length*
		X					• *presence of query terms* • *caption readability* • *url complexity* • *url length*
		X	X				*rank position of caption*
		X	X	X			*ordering of rank positions*

X = this relationship has been studied (e.g., relationship between click-through and relevance of information source)

Table 2.6 Organization of summary: studies of factors in interactive search behavior

Searcher's mental state	Types of **search behavior**			System responses
	Interaction as query formulation *length of query*	Interaction with results lists *click-through, visual scanning, task time*	Interaction with open documents *dwell time, scrolling, printing, saving, task time*	
relevance of information source		• Fox, et al., 2005 • Joachims, 2002 • Kemp & Ramamaohanarao, 2002 • White, Jose, & Ruthven, 2001	• Morita & Shinoda, 1994 • Kelly & Belkin, 2001, 2004 • Kelly & Teevan, 2003 • Kim, Oard, & Romanik, n.d. • Konstan, et al., 1997 • Oard & Kim, 1998, 2001 • White, et al., 2002	
specificity of information need	• Lau & Horvitz, 1999 • Phan, Bailey, & Wilkinson, 2007			
type of information need		• Lorigo, et al., 2006 • Cutrell & Guan, 2007 • Guan & Cutrell, 2007		
		• Cutrell & Guan, 2007 • Guan & Cutrell, 2007 • Clarke, et al., 2007		*snippet length*
		• Clarke, et al., 2007		*presence of query terms, caption readability, url complexity, url length*
		• Cutrell & Guan, 2007 • Granka, et al., 2004 • Guan & Cutrell, 2007 • Joachims, et al., 2005 • Klockner, et al., 2004 • Lorigo, et al., 2006		*rank position of caption*
		• Joachims, et al., 2005		*System performance as ordering of rank positions*

interaction with open documents. Findings from these studies are summarized in Section 2.6.b. Another set of studies focuses on the effect of system responses on visual scanning and click-through behavior. The effects of the rank position of a caption are studied most often, however a smaller number of studies have investigated effects due to caption features. We summarize these findings in Section 2.6.c. Table 2.6, above, places each study within this framework.

After summarizing the studies listed in Table 2.6, we review the SAMLight model in Section 2.6.d. This model is important because it integrates observations from all three forms of behavior and demonstrates the informativeness (to the system) of the temporal dynamics of behavior. However, the model cannot explain the relationship between the searcher's experience of the system (e.g., system performance) and behavior. Our study provides insight into this relationship. The summary concludes with a discussion of adaptive search behavior (Section 2.6.e).

2.6.a Inference from search behavior to the user's mental state

Many studies of behavior have been undertaken with the goal of using observations of behavior to infer aspects of a user's mental state. We draw the following conclusions from these studies (see Table 2.7).

A user's relevance judgment cannot be inferred from isolated information about click-through, or subsequent interaction with an open document. While query length has some association with the specificity of an information need (with specificity identified *post hoc* by researchers), the relationship is not strong enough to produce a reliable inference. The relationship between the specificity of an information need and search behavior has not been studied experimentally.

Table 2.7 Studies of inference *from* observable behavior *to* mental states

Searcher's mental state	Search behavior		
	Query formulation	Interaction with results lists	Interaction with open documents
relevance of information source		*click-through* • Fox, et al., 2005 • Joachims, 2002 • Kemp & Ramamaohanarao, 2002 • White, Jose, & Ruthven, 2001	*dwell time, scrolling, printing, saving* • Morita & Shinoda, 1994 • Kelly & Belkin, 2001, 2004 • Kelly & Teevan, 2003 (review article) • Kim, Oard, & Romanik, n.d. • Konstan, et al., 1997 • Oard & Kim, 1998, 2001 • White, et al., 2002
specificity of information need (specific vs. general)	*length of query* • Lau & Horvitz, 1999 • Phan, Bailey, & Wilkinson, 2007		

2.6.b Effect of the user's mental state on behavior

Several studies have manipulated the type of task assigned to searchers (an aspect of a user's mental state) and observed effects on behavior. We draw the following conclusions from these studies (see Table 2.8).

Searchers allocate their attention differently for informational and navigational tasks. This is likely because, for different types of tasks, the information sought is located in different places in the system. For informational search, the information sought is located in underlying sources and not in captions; for this reason, searchers give more visual attention and task time to open documents, and less to results lists. In contrast, for navigational tasks, searchers often find the information sought in captions, with no need

Table 2.8 Studies of effects of user's mental state (task-type) on behavior

Searcher's mental state	Search behavior	
	Interaction with results lists	Interaction with documents
type of information need (informational vs. navigational)	*click-through, visual scanning, task time* • Lorigo, et al., 2006 • Cutrell & Guan, 2007 • Guan & Cutrell, 2007	*task time* • Lorigo, et al., 2006

to open a document. For this reason, during navigational tasks searchers allocate more visual attention and task time to results lists than to open documents.

2.6.c Effect of system responses on behavior

Most studies of the effect of system response on search behavior focus on the rank position of captions. Generally, this work is related to efforts to infer relevance judgments from click-through behavior. Very few studies have examined how other features of system response, such as caption text, affect behavior. Only one study has examined interaction effects from these types of factors. Several studies have investigated the effect of system performance on searcher "success," but only one has examined the effect of performance on behavior, per se. Taken together, we draw the following conclusions from this set of studies (see Table 2.9).

Search behavior is influenced by the system's responses. Of course, this is not surprising; a system works by signaling its state to its user. A rank-based search system communicates its estimate of the relevance of an information source by positioning the source on the results list. Users have learned to rely on this signal and focus their visual attention where the system places the sources most likely to be relevant, at the top of the list. Click-through occurs when a searcher has a sufficient level of belief that a document is relevant, and naturally, this behavior also focuses at the top of the list. Other

characteristics of captions also affect searchers' expectations of relevance. These include

the presence of query terms in the caption, the readability of the caption, the length of the

snippet text, and the length and complexity of the *url* displayed in the caption. There is no

published research on the relative importance of each factor in behavioral responses.

As discussed above, search behavior is affected by task-type. Importantly, one

study has demonstrated that system responses affect behavior differently for

informational and navigational tasks. For informational tasks, searchers adapt their

scanning to capitalize on the advantages of longer snippets. However, within the

constraints of the experimental conditions, during navigational search users are not able

to compensate for the disadvantages of longer snippets. There are no published studies of

interaction effects due to task-type and other caption display features.

Table 2.9 Studies of effects of system responses on behavior

Search behavior			System responses
Interaction with results lists:			
click-through & *visual scanning*	*click-through*	*visual* *scanning*	*Caption features:*
• Cutrell & Guan, 2007 • Guan & Cutrell, 2007	• Clarke, et al., 2007		*snippet length*
	• Clarke, et al., 2007		*presence of query terms ,* *caption readability, url* *complexity, url length*
• Cutrell & Guan, 2007 • Granka, et al., 2004 • Guan & Cutrell, 2007 • Joachims, et al., 2005 • Lorigo, et al., 2006		• Klockner, et al., 2004	*rank position of caption*
			System performance:
• Joachims, et al., 2005			*ordering of rank* *positions*

2.6.d Predicting interactive behavior

Table 2.10 places the most predictive features in the SAMLight model within the structure we have used for this summary. Of the behavioral factors studied in the literature we have reviewed (see Table 2.6), only three are included among the 51 features used in the development of the SAMLight model. These are: (1) the *rank position* of a clicked caption, (2) *dwell-time on open documents*, and (3) the *length of a query*. Surprisingly, none of these measures is highly predictive of behavior. The rank position of a click-through is not a strong predictor of subsequent search action. As the authors point out, rank position is a strong factor in the searcher's choice of *which* caption to click, but it is not a strong indicator of the relevance of the underlying document. While the time interval between actions[11] is the strongest predictor of the next search action, open document dwell-time is not a strong predictor. Dwell-time occurs after a click-through, and is a measure of *how* a searcher uses time before the next action.

The length of an individual query is not, in itself, a reliable indicator of the specificity of an information need, nor is it a strong predictor of search action. However, *over the course of a session* it is a strong predictor of the next action. Interestingly, the probability of a click-through increases as queries grow longer over the course of a session. This implies that *changes* in query length, and not query length per se, are meaningful indicators of the searcher's experience of the system.

Three of the most predictive features of the SAMLight model are indicators of the quality of a query, and may be predictive of query failure: (1) the presence of a spelling suggestion for the query (*HasSuggestion*), (2) the probability that the query will result in

---

[11] Recall, in the SAMLight model, actions include click-through, re-query, and re-entry to the search engine via the browser's back button.

a click-through (c_rProb), and (3) the presence of a "definitive *url*," a near certain probability that a specific *url* will be clicked as the next search action (*HasDefinitive*).

Two of these features, the probability of click-through and the definitive *url*, are known to the system but are not revealed to the user by the system. Both features are determined by the behavior of users who submitted the same query in the past. If a query used in the past has a definitive result it is more likely that the query will succeed for anyone who uses it. On the other hand, if other searchers have used a query in the past, and users rarely click on results from the query, it is likely that the query will fail *any* searcher who uses it. A system with this "knowledge" might use the information to detect and assist a user experiencing a difficult search.

A spelling suggestion message is also an indicator of query quality. Of course, the message is a signal to the user that the query just submitted is likely to have failed. Importantly, the message also provides support for overcoming the failure.

In our view, three other predictive features found in the SAMLight model characterize behaviors that are likely to occur when a searcher experiences repeated query failure during a session. If a re-query requests a subsequent page of results, as indicated by q(FirstResult*)*, it is likely that the first page of results did not contain the desired information. If the searcher continues to submit queries during a session, as indicated by S(Num$q$), it is likely that prior queries have failed. If a searcher has submitted successive queries without intervening click-throughs, as indicated by S(qFrac), it is *highly* likely that prior queries have failed. Together, these three features of behavior may be a meaningful signal to the system that the searcher is responding to

Table 2.10 Predictive features in the SAMLight model (after Downey, Dumais, & Horvitz, 2007b)

Highly predictive measures:				
Search behavior (most recent prior action)				**System responses**
query action	*results list action*	*open document action*	*integrated across actions*	results list feature
• starting rank requested	• click-through			• spelling suggestion
History of action during searcher's session:				**History of query*:**
• # prior queries this session • longest query this session			• time since last action • ratio queries / actions	• probability of click-through for the query • high probability *url* (a definitive query)
*the following features are **not highly predictive** of the next action*				
query-length	rank position of clicked caption	dwell time on open document		

* The model has access to a history of queries submitted by other users, and the history of click-throughs associated with each query

repeated query failure, that is, that the system is performing poorly. Conceivably, a system capable of detecting failure could signal its user and provide support for avoiding further failure.

Not surprisingly, the most highly predictive features of behavior are directly related to query submissions over the course of the search session. Fundamentally, this is because a query is the only mechanism a searcher has for controlling the system's responses. A re-query occurs when a searcher's prior action did not produce results good enough to fulfill the information need, that is, re-query occurs when the prior query has failed.

Importantly, information about a specific individual user is not highly predictive of the *type* of behavior a searcher is will use next. This suggests that searchers use, to some significant degree, similar behavioral responses during search. Further, this suggests that systems might be designed to detect meaningful behavioral responses, and to accommodate and augment these general tendencies of searchers, without the need of detailed personal data about individual searchers.

2.6.e Adaptive behavior

Users adapt their search behavior to match system responses. Adaptation allows searchers to exploit the advantages of a different response, or to avoid disadvantages. For example, in conducting informational searches, users change their scanning behavior when they encounter long snippet text and can complete their tasks more quickly. Adaptation may not always fully compensate for disadvantageous system responses, however. While users conducting navigational searches adjust their scanning behavior when faced with long snippets, their adaptation is ineffective, and they complete their

work more slowly. The adaptation of scanning behavior to snippet length is a simple example of the flexibility of human cognitive processing.

Adaptation to poor system performance is a more complex process. Except in the case of a spelling suggestion returned from a misspelled or mistyped query, search systems do not produce clear signals that performance is poor. Indeed, relative to what the user knows about how well the system is working, systems have little data with which to assess their own performance, and current systems fail to use the data that *is* available. As discussed in Chapter 1, current systems provide searchers with very little information about factors that contribute to query failure, and present few strategies for improving a query. Searchers have only the query mechanism, and their own strategies and solutions, as means for overcoming a failure. We know that searchers solve this problem every day, however, we know very little about *how* the problem is solved. This dissertation contributes to our understanding of how searchers overcome query failure and remain productive searchers.

While some studies have found that searchers can compensate for poor system performance, others have shown that it is not always possible to do so. These differing results are likely due to differences in the experimental tasks used, or in the specific characteristics of the experimental system failure, or both. Certainly, understanding the effect of system performance on search behavior is a complex undertaking.

The study presented in this dissertation comprises a single experiment in which we manipulate system performance and observe search behavior. We outline the goals of the study in the next chapter.

3. RESEARCH QUESTIONS

3.1 PREFACE

The experiment reported in this dissertation was originally designed to study stopping behavior during interactive search (Kantor, 1987). The hypothesis underlying the experiment was that a searcher's decision to stop searching is dependent in a specific way (Bayesian modeling) on the number of query failures encountered during a search. The data analyzed and reported here were collected during a pilot test conducted for the original study. The pilot tested the experimental design, the protocol, and the experimental computer system. While data were being collected, it became obvious that the protocol resulted in persistent searching, not stopping. Subjects continued to search despite repeated query failures, working diligently to reach the task objectives assigned in the protocol. With respect to stopping, the demand characteristics of the experiment (Rosenthal & Rosnow, 1969) overwhelmed any effect due to the degraded performance of the system. However, we found that the protocol produced rich data on searcher behavior and system responses. We decided to complete all of the planned 36 sessions, but with a change in the protocol that explicitly granted subjects permission to quit the experiment without finishing. With the change, if a subject searched for more than 80 minutes without completing the experiment, the researcher reminded the subject that he or she was free to quit the experiment without finishing (see Appendix B.4, page 135, for the change to the protocol).

As originally conceived, the experiment was intended to measure changes in stopping criteria when a system performs poorly. When we found that subjects rarely responded to the degraded performance by stopping the search altogether, a more complex research objective emerged: to explore the effects of system performance on

behavior observed *during* a search. Thus, this dissertation is an exploratory analysis of effects on search behavior attributable to system performance. Because the data were influential in shaping our understanding of the problems, we do not formulate our research objectives, stated below, as "hypotheses verification." The confidence intervals we report are thus exploratory and not confirmatory.

Exploration involves the examination of many possible relations. We follow custom here in presenting confidence intervals and p-values, to sharpen the assessment of effects. These values are not corrected for experiment-wise and per-comparison error rates, as there is no generally accepted procedure for dealing with a range of exploratory techniques simultaneously.

3.2 RESEARCH QUESTIONS

This study explores the effect of query failure on search behavior and addresses four interrelated questions. First, how do searchers adapt their behavior when a system is working poorly? Second, can searchers overcome degraded performance and remain productive? Third, how did our experimental manipulations affect system responses? Finally, how are system performance, system response, and search behavior interrelated? These questions are detailed below.

(1) How does search behavior change when system performance is degraded?

In order to answer this question, we produce query failure intentionally and repeatedly by manipulating results returned from the Google search engine. A high rate of query failure degrades system performance. We study two types of behavior: a) *identification* of valuable information sources, and b) *querying*. Specifically, we examine the following measures of behavior for each search:

- In our experiment, searchers identify valuable information sources by clicking an affirmative check mark, in what we call *flagging an information item*. The information sources a searcher believes to be good we term *flagged*. The relative frequency with which a searcher flags items is the *flagging-rate*.

- When a searcher submits a query, it is automatically recorded and time-stamped by our experimental system. The average time interval between query submissions, over the course of a topic search, is the *query-rate*.

- Searchers may make spelling errors or typing errors in a query. When this occurs, the search engine may return a message indicating that it has detected the error. We record and analyze these messages as indicators of query error (*spelling-message-per-query*).

- We measure one characteristic of the content of queries: the average number of words used in each query (*average-query-length*).

(2) How does searcher productivity change when the system is degraded?

Searcher productivity is defined as the number and quality of information sources found relative to the time used during a search. Specifically, we measure productivity as follows:

- After all data were collected, the researcher assessed every flagged information source, using a three-level quality scale. We measure the number of *good, marginal,* and *bad* sources found, as well as the distribution of quality levels among the items flagged during a search (which are good, marginal, and bad *item ratios*).

- The time used during a search is measured as the *elapsed time* between the first query submission and the searcher's indication that the search is complete.

(3) How do system responses change as a result of the experimental manipulations?

We study two types of system responses: a) the length of results lists, and b) frequencies of item displays.

- Generally, Google results lists indicate that many thousands of information sources are available in query results. It is possible, however, for Google to return a results list with fewer than 20 items (we say the list is *truncated*), or an empty list. We measure the *average list length* of results lists, as well as the frequency of *full*, *truncated*, and *empty* lists.

- The captions displayed on results lists represent information *items*. An information item may be displayed on more than one list during a search; the fraction of item displays that repeat is termed *item-display-repetitions*. One of our experimental treatments may cause the system to display different items when the same query is resubmitted. *Unique-items-per-query* measures the tendency of a system to return different items for the same query.

4) How are system performance, system response, and search behavior interrelated?

This question is addressed in exploratory analyses of the length of results lists, the time intervals between queries, and the relationship between these and query failure.

3.3 ROADMAP TO THE REMAINING CHAPTERS

The remaining chapters of the dissertation are organized as follows. Chapter 4 details the experimental method, including the design of the factorial experiment, the protocol, the computer system, and data collection. Chapter 5 covers the characteristics of the experimental subjects, preparation of the data for analysis, derivation of variables, and an overview of the analytical approach. Chapter 6 presents results for research

questions 1, 2, and 3, above. Chapter 7 presents results from the exploratory analysis that

answers question 4, above. Chapter 8 summarizes our findings and contributions, and

discusses the limitations of the study and future research.

4. RESEARCH METHOD

This chapter describes the details of our research method, focusing on the design of the factorial experiment. We begin by explaining controls for the large incidental effects we anticipate in our data. We then cover subject recruiting, the construction of search topics, and the protocol. The chapter concludes with a description of the experimental systems, including the interactive components and data collection.

4.1 DESIGN

4.1.a Designing for large incidental effects

One of the greatest challenges in the design of experiments in interactive information retrieval is the presence of incidental effects that are often larger than the effects of experimental factors (Banks, Over, & Zhang, 1999). *User effects* are generally the largest of these; different people have different habits, skills, and idiosyncrasies in their search behaviors. We assume that these characteristics are similar for every search conducted by a given user. Interactive search experiments generally require that subjects complete a series of searches where each search is on a different topic. *Topic effects* are produced by the subject matter of a search. Different topics often result in different levels of system performance or different levels of difficulty for searchers. Some topics may have many relevant information sources, while other topics may have very few. The vocabulary used to describe some topics may be very general and well known by searchers, while for other topics vocabulary may be highly specialized and unfamiliar to searchers. For any topic used in the experiment, we assume that effects of this type are similar for every search conducted on that topic. *Position effects* occur within the context of the experiment in which searches are conducted in a certain order: 1st, 2nd, 3rd, 4th, etc.. Searches conducted early in an experimental session (e.g., positions 1, 2 or 3) are likely

to be different from subsequent searches because subjects may grow tired or bored over the course of a session. We assume that the effects due to position are similar across subjects and topics.

Typically, these three effects cause high variability in measures of system performance, system response, or searcher behavior. Without a design that controls for these incidental effects, the system effects of interest may be undetectable. This experiment uses a diagram-balanced, mixed-model design, which permits estimation and isolation of the *main* effects of these incidental factors.

We know that there are also interaction effects between users and topics because users have different prior knowledge about a topic. It is not possible, however, to isolate these interactions in information retrieval research. As a subject searches on a topic, he or she learns about that topic. If a subject were to search again on the same topic, the prior search would affect his or her behavior. For this reason, we assign search topics to each subject only once and only one data point is collected for each subject/topic pair on any measure. Because each subject searches on each topic only once, we cannot eliminate the effects of subject/topic interaction. Thus, our experimental design must produce system effects large enough to be detectable within the noise of these interactions. For this reason, we administered our experimental treatments in a block of 4 consecutive searches.

4.1.b Blocked-sequential, mixed-model, diagram-balanced design

Our 3x3 mixed-model factorial design used 3 blocks and 3 groups. The design enables testing of between-group differences. Subjects completed 3 blocks of searches, each consisting of 4 topic searches, for a total of 12 searches (or trials). Block 1 was a
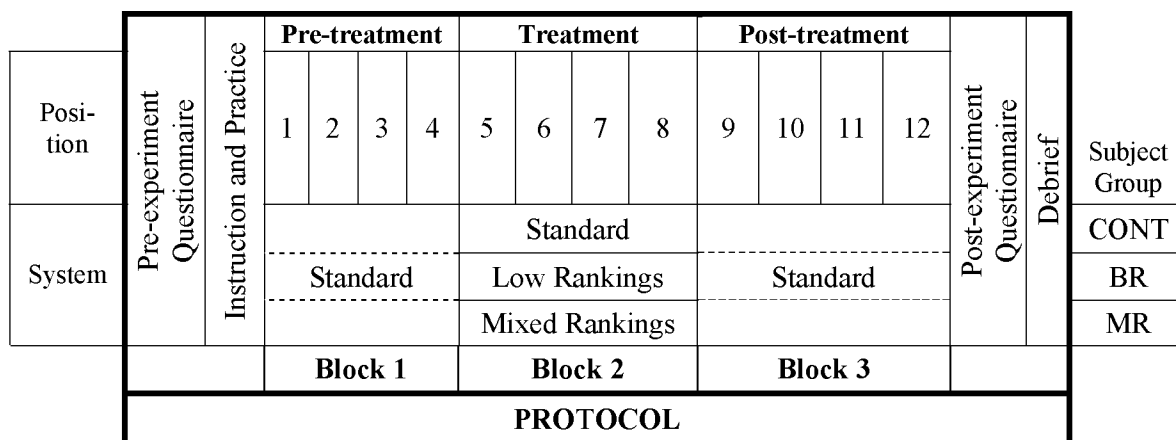
Posi-tion	Pre-experiment Questionnaire	Instruction and Practice	Pre-treatment				Treatment				Post-treatment				Post-experiment Questionnaire	Debrief	Subject Group
			1	2	3	4	5	6	7	8	9	10	11	12			
System							Standard										CONT
			Standard				Low Rankings				Standard						BR
							Mixed Rankings										MR
			Block 1				Block 2				Block 3						
	PROTOCOL																

Figure 4.1. Block design and protocol.

Table 4.1 The twelve topic search orders

	Pre-treatment				Treatment				Post-treatment			
Position	1	2	3	4	5	6	7	8	9	10	11	12
Order	Topic											
a	7	9	4	10	2	8	3	6	5	12	1	11
b	2	1	10	8	5	12	7	4	11	9	6	3
c	4	10	9	7	6	11	12	8	1	3	5	2
d	6	5	3	4	1	2	8	11	9	7	12	10
e	9	2	11	5	12	3	4	10	6	1	8	7
f	3	8	6	9	11	4	10	1	12	2	7	5
g	8	3	12	11	7	5	9	2	4	6	10	1
h	11	4	1	6	9	10	5	3	7	8	2	12
i	5	12	2	3	8	9	1	7	10	11	4	6
j	12	6	5	1	3	7	2	9	8	10	11	4
k	10	7	8	2	4	1	11	12	3	5	9	6
l	1	11	7	12	10	8	6	5	2	4	3	9
	Block 1				Block 2				Block 3			

pre-treatment block, in which all subjects searched using the standard system. During

Block 2, the *treatment block*, the control group continued to use the standard system

while subjects in the treatment groups used one of two degraded systems. In Block 3, all

subjects again searched using the standard system. We did not inform subjects of the

blocking, and no break was given between the blocks. Figure 4.1 (above) depicts the

design.

Topic order was diagram-balanced, with each subject assigned to one of 12 search orders (topic orders are depicted in Table 4.1, above). One subject in each group searched in each of the 12 order assignments, for a total 432 searches (3 groups x 12 subjects x 12 searches).

4.1.c Subject recruitment

We recruited our 36 subjects on the central New Jersey campuses of Rutgers University. Subjects were recruited using classroom posters and email sent to various school and department administrators. Undergraduate and graduate students, as well as non-students, were eligible to participate. They were paid $15 for their time. To motivate search effort, subjects were told that an additional $40 would be paid to the subject who "finds the most good information sources and the fewest bad sources." All sessions were conducted during one three-week period. Subjects were randomly assigned to one of the three groups: the control group, or one of two treatment groups. Subjects were told that they would search using the Google system, which was, in fact, the standard system behind the experimental interface.

4.1.d Experimental search topics

The 12 search topics (see Table 4.2, below) were presented as declarative statements. In order to make every search somewhat difficult, independent of system performance, we designed the statements to require disambiguation during query formulation. For this reason, each statement contained a subset of terms that could express one or more topics unrelated to the topic of the statement. For example, the topic statement "The option to purchase a bull is an investment alternative for farmers." is about a farmer's decision to purchase a bull. However, if a query on this topic was not

Table 4.2 The twelve topic statements

1	The option to purchase a bull is an investment alternative for farmers.
2	It is difficult to secure a mortgage or insurance for property directly on the bank of a river.
3	In some cultures it is common to hire a band for a wedding.
4	Conductors train for many years before reaching a professional level.
5	Fishermen find it difficult to earn a net profit.
6	Women boxers often receive a small purse.
7	It is easy to tire when driving a car.
8	For security, conductors carry radios as they move between stations.
9	Firemen can make progress on the ladder of the profession.
10	Mints and treats that look like coins are favorite holiday candies.
11	It is difficult to produce containers that maintain the freshness of vegetables during shipping.
12	Drinking water helps you to stay well.

sufficiently disambiguated, the system could easily return information sources about financial matters such as "bull markets," "alternative investments," "options," etc..

4.1.e Equipment and logistics

All sessions took place in the same quiet, isolated room. Only one subject participated at any one time. Two monitors were placed on a large table within reach of the subject; one displayed the experimental system and the other displayed any web pages or other documents "clicked open" by subjects. All subjects used the Firefox browser. Prior to beginning the experiment, each subject chose a familiar computer mouse and a chair from a small selection of each. Paper instruments and a complete printed copy of the protocol were bound in a three-ring binder and presented to each subject (see Appendix B.1). Subjects were encouraged to move the computer equipment, chair, and binder to remain comfortable throughout the session.

4.1.f Protocol

Introduction. The researcher greeted each subject in a waiting area and escorted

him or her to the inner office in which the experiment took place. After a brief

introduction, each subject signed an informed consent form and completed a pre-

experiment questionnaire, which collected demographic information and information

about prior search experience and attitudes (see Appendix B.3.a).

Search task assignment. Next, we gave subjects the details of their experimental

task in a mock "job description," which provided context for the activity (see Figure 4.2,

below). The subject's job was to find as many "good information sources" as possible for

a group of journalists who needed information about 12 topics. The researcher described

a good information source as one "*you* could and *would* use to get information about the

topic." Subjects were told twice that there was no time limit on searching, but that they

were expected to complete searching on all twelve topics in an hour or less. The

researcher also reminded the subjects orally "there may be some times when there is little

or no good information on a topic" and that they could "stop at any time." A printed copy

of the job description was accessible throughout the session.

Practice trials. Next, we showed subjects the features of the two interfaces they

would use during the experiment: an experiment control interface and a search interface

(we describe the interfaces below). The researcher explained the system and

demonstrated the interfaces by searching for an example topic. Before beginning the first

of the twelve experimental topics, subjects were required to practice using the interfaces

by searching on at least one practice topic, with the option to continue practicing until

ready to begin the experiment. The example topic and practice topics were the same for

TRAINEE JOB DESCRIPTION

In your job as a trainee, you support the journalists at the newspaper. Your responsibility is to find information about the journalists' article topics. Today you need to search for good sources of information about *twelve* different topics that the journalists are working on. You search by using Google.

The Google system you use looks slightly different from regular Google. As you search, the topic you are working on is displayed at the top of the screen. Just like with any Google search, you will see a list of websites, and you may visit those websites to see if they have good information about the topic. In order to tell the journalists about the good information sources, you simply check a box indicating that the site on the list is good. All the items you check as good will be automatically included in a list for the journalist working on the topic.

You won't be given any information about why a journalist is looking for information on a topic, or what about the topic is important. For this reason, any source with information that will inform the journalist on the topic can be considered a "good" source. You need to find as many "good" information sources as you can, but it is also important to avoid sending information sources that are not good.

At the newspaper, there is a bonus for finding only good information sources. The journalists judge whether the sources found by trainees are good. The five trainees who find the most good information sources and the fewest "bad" sources, are eligible to win a "bonus". The bonus is given to one trainee, who is selected by lottery.

There is no time limit on searching, but your boss expects that you will be able to finish searching for all twelve topics in an hour or less.

Figure 4.2. Trainee job description.

all subjects. Once subjects completed the practice topic(s), and any questions they had were answered, the experiment was started by clicking "first topic."

Experimental trials. At the start of each topic search, the experiment control interface displayed the topic statement and prompted subjects to complete a paper pre-search questionnaire (see Appendix B.3.b). Once the questionnaire was completed, subjects clicked "start topic search" and the search interface was displayed. Using the search interface, subjects submitted queries in the search box, received search results, browsed the results lists, clicked-through to inspect underlying information sources (website or other document form), and submitted additional queries, as needed. There was no limit on the number of queries submitted, nor on the type or number of captions that could be "clicked open" for display.

When the searcher *clicked* a caption to open it, the underlying information source was displayed on a second monitor, so that subjects could see the results page on one monitor, and simultaneously, the open information source on the second monitor. Subjects were able to open websites and files in various formats such as PDF, PowerPoint, and MSWord. They were also able to use the browser's within-page search function to navigate through open sources without restriction.

On every results list, a small checkbox was displayed next to each caption listed. Subjects used the box to "flag" the corresponding information source if it was judged a "good information source" for the current topic. Once a subject flagged a caption, if it was displayed on a subsequent list (for that subject), it appeared with the check already placed in the checkbox, indicating that the source had already been identified as *good.* Subjects were able to uncheck the checkbox. We based our record of the subject's

judgment on his or her *final* indication for each item. The only sources a subject could flag were those displayed on a results list. That is, if a subject found a good source by exploring a website, he or she could not flag that source unless the system could be induced to present it in a results list. The subject clicked a series of buttons to confirm completion of a search, and thereafter, could not continue to search on the topic.

After each topic was completed, the experiment control interface prompted the subject to complete a paper post-search questionnaire for that topic (see Appendix B.3.c). Once the questionnaire was completed, the subject clicked a "continue" button and the cycle of Topic-display→Pre-search-questionnaire→Topic-search→Post-search-questionnaire began anew. Subjects repeated this sequence until all twelve topics were completed or until they quit the experiment.

Debrief. Finally, after completing the 12th topic or quitting, subjects completed a post-experiment questionnaire (see Appendix B.3.d), and were debriefed regarding the deceptive aspects of the experiment. They then received the $15 participation payment. At a later date, we paid the $40 bonus to the winner by mail.

4.1.g Instruments

We used four questionnaires, all of which were administered on paper (see Appendix B.3). The pre-experiment questionnaire gathered demographic characteristics and asked subjects about prior experience with, and attitudes toward, web searching and Google. A post-experiment questionnaire asked subjects to define the phrase "good information source" in their own words in an open-ended, written response. After writing their definition, they turned the page to see a list of 12 possible attributes of a good information source (McInerney & Bird, 2005). They then indicated the importance of each of the 12 attributes using a 5-point Likert scale. The order of the attributes was

rotated between subjects with the same ordering scheme used for the search topics (see Table 4.1).

During the experiment subjects completed two different questionnaires for each search topic. A pre-search questionnaire was administered immediately after a subject read the search topic, but before the subject started searching on the topic. It measured 8 aspects of the subject's familiarity with the topic, his or her expectations, and confidence level. A post-search questionnaire was administered immediately after each search was completed and before the next topic was displayed. It measured 7 aspects of the subject's assessment of the immediately preceding search experience.

4.2 EXPERIMENTAL SYSTEMS

4.2.a Interactive component

Underlying system. Queries submitted by subjects were passed through a proxy server (e-kiwi), which stored the queries and other data collected. Queries were submitted to Google in real time with *url* parameters that requested a 20-caption list. The *html* code returned from each query was scraped, parsed, and manipulated. Prior to display in the search interface, the system stripped all advertising and sponsored captions from the list. The Google links "Cached - Similar pages" were also be removed. The *html* code for each caption was stored before display.

The standard system. The standard system displayed captions in the order returned by Google. The list always displayed the top-ranked caption first and subsequent captions in an unaltered order. Subjects in the control group continued to receive results from the standard system during the treatment block (searches 5 through 8 in the sequence of 12 searches).

The experimental systems. We created the two experimental conditions by manipulating both the queries submitted by subjects and the search results returned by Google. Starting ranks were altered according to the subject's assigned condition as follows: For the bottom-rankings (BR) condition, the query always requested a list starting at the 300th caption in Google's results set; this mimicked the failure of a system with little or no information in a topic domain. For the mixed-rankings (MR) condition, the starting point of the displayed list varied within a topic search, as indicated in Table 4.3; this mimicked a maladaptive mechanism such as an automatic query expansion that fails to converge correctly on the search topic. In both treatment conditions, the rank *order* of the Google results was not altered.

Table 4.3 Starting ranks for the Mixed-Rankings (MR) condition

Queries	Ranks Displayed (displayed as rankings 1 – 20)
First, Second	300 – 319
Third	120 – 139
4th-5th	300 – 319
6th	1 - 20
7th	300 – 319
8th	120 – 139
9th – 10th	300 – 319
11th	1 – 20
12th to last	300 – 319

Interfaces. The experimental system had two interfaces. The *control interface* displayed experimental instructions, including introductory text, requests to complete paper surveys, and the display of topic statements prior to search. The *search interface* was always reached through the control interface. It displayed two frames, with the topic statement always visible in the upper frame, and a modified Google interface in the larger lower frame (see Figure 4.3).

After subjects completed the first topic search, the upper frame of the search interface displayed a "reminder" box. This box reported the total elapsed search time since the start of the first search, the number of topics completed, and the number of topics not yet finished. The box was updated at the start of each topic. The standard navigational links usually appearing on the Google search interface were displayed but disabled. Every caption in the results list was left-aligned and displayed using the text and formatting obtained from Google. In addition, a single checkbox was displayed to the left of each caption listed; subjects used the boxes to "flag" good information sources. Each results list was limited to no more than twenty captions, with no option to continue to the next page of results ("next page" links were visible but disabled). Lists returned from Google with fewer than twenty captions displayed only the returned captions. For the two experimental systems, the standard Google results counts and timing text (e.g., *"Results 1 - 20 of about [number] for [query terms]. (0.xxx seconds)"*) were altered to indicate that the list started at rank 1. Any "did you mean…" links and "hint" messages returned by Google were also displayed. The "did you mean…" spelling correction link was "clickable," so that subjects could click the link to submit the suggested query. The link to each information source in the list was live and subjects were able to click those links to open information sources. Buttons were provided for confirming completion of each search.

4.2.b Data collection

As each search progressed, the system logged measures of search activity in a database. These measures include:

(a) the beginning timestamp for each search,

(b) each query submitted (with timestamp),

(c) codes for messages and query suggestions returned by Google,

(d) each caption displayed to the subject and its rank position in the display,

(e) a record of each caption flagged that remained flagged at the end of the topic

search, and

(f) the ending timestamp for each search (click of button indicating completion of

topic search).
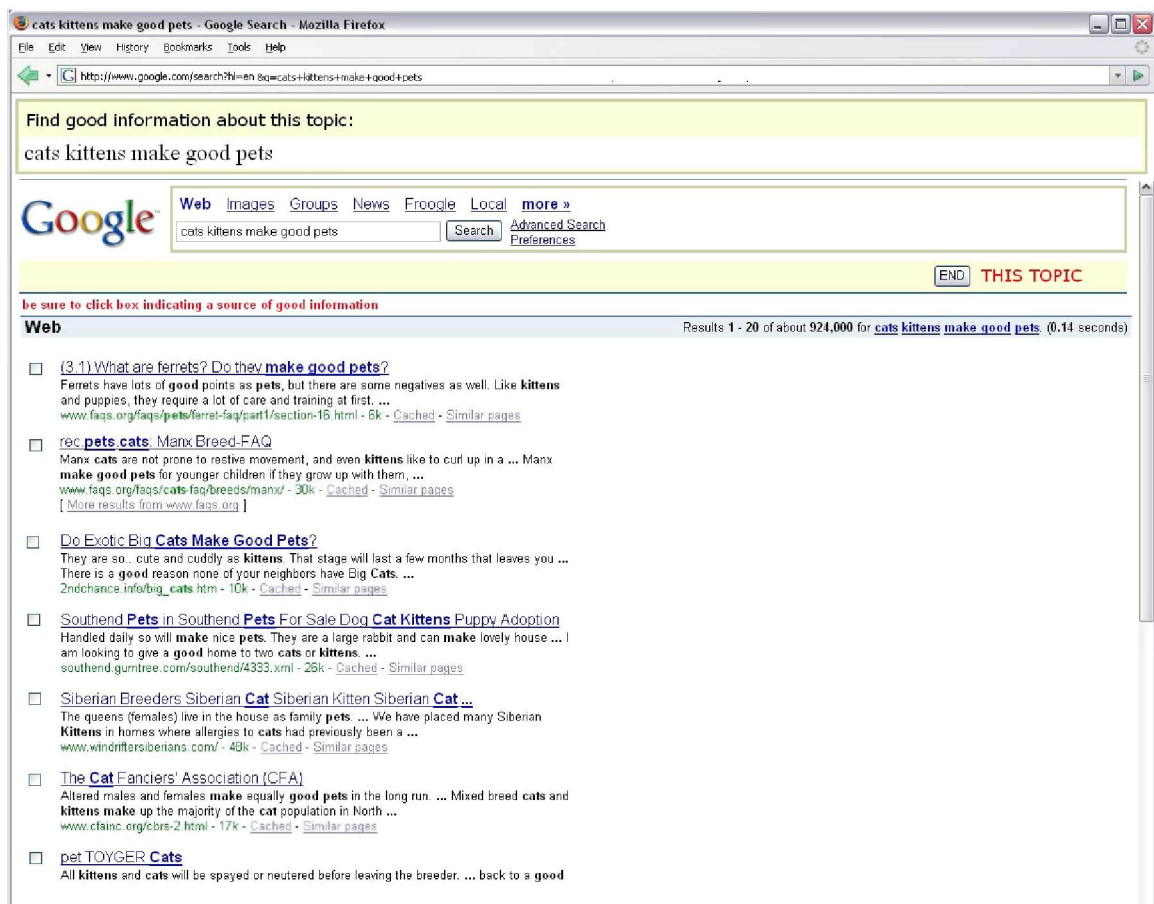
Figure 4.3. Experimental search interface.

5. DATA PREPARATION AND ANALYSIS

In this chapter, we present our analysis process, beginning with the characteristics of our research subjects. We then describe the assessment process used to rate each of the information sources flagged by subjects during the experiment. Next, we detail the derivation of our measures of system performance, system response, searcher productivity, and search behavior. Methods for extracting incidental effects are described next, followed by details of the analysis plan. The chapter concludes with an analysis confirming that our experimental manipulations of Google produced degraded performance.

5.1 SUBJECTS

5.1.a Subject characteristics

Pre-experiment measures revealed no significant differences among the three subject groups with regard to prior experience with, and attitudes about, web searching and Google. We find no significant differences in the demographic characteristics of subject groups. A summary of demographics and χ^2 tests of independence are in Appendix C.

We used Principal Components Analysis to examine the relationships between responses to questions 8 through 12 in the pre-experiment questionnaire. These questions covered prior experience with searching and prior experience with the Google system (see Appendix C for the correlation matrix). Varimax rotation with Kaiser normalization was used to simplify interpretation of the components. Two components had eigenvalues greater than 1.0 (see Table 5.1) and communalities greater than 0.70 (Stevens, 2002). We named these factors *F_self_assessment* (component 1) and *F_confidence_in_Google* (component 2) (see Table 5.2). We find no significant between-group differences in

Table 5.1 Eigenvalues and percentage variance explained before and after rotation

	Extraction sums of squared loadings		Rotation sums of squared loadings	
Component	Eigenvalue	% variance	Eigenvalue	% variance
1	3.038	60.8	2.924	58.5
2	1.123	22.5	1.238	24.8

Table 5.2 Rotated component matrix of factor weightings*

Question	Component 1	Component 2
(# 9) I am interested in online searching.	.920	-.054
(# 10) I enjoy trying new ways to use the Internet or World Wide Web.	.895	-.102
(# 11) I am familiar with Google searching.	.836	.283
(# 8) I usually find what I am looking for on the Internet or World Wide Web.	.759	.446
(# 12) Google can find anything I need.	.020	.972

*Varimax rotation method, with Kaiser normalization. Rotation converged in 3 iterations.

either F_self_assessment ($F(2,34)$=0.57, p=.57) or F_confidence_in_Google ($F(2,34)$ =0.75, p=.48).

5.1.b Persistence and attrition

Six subjects (3 control, 2 BR, 1 MR) quit the experiment before completing the third block, but after completing all the searches in the first two blocks. Table 5.3 details the incomplete sessions. In our subsequent analysis, we excluded data from incomplete searches and retained data from 416 completed topic searches.

We used a χ^2 test of independence to compare the demographics of subjects who quit and those who completed all 12 searches and found no significant differences in native language, gender, age, college enrollment status, educational background, educational level, and PC usage. No significant differences were found in pre-experiment beliefs (F_self_assessment and F_confidence_in_Google). We used an ANOVA to examine the main effect of completion status. No significant differences were found in

Table 5.3 Incomplete experimental sessions

SubjectID	Group	Number of searches completed	Number of incomplete searches
4	CONT	9	3
6	CONT	9	3
12	CONT	11	1
22	MR	9	3
26	BR	9	3
36	BR	9	3
Total incomplete searches			16

the number of queries submitted, the number of items flagged, the number of marginal items flagged, the number of bad items flagged, and the number of missing items flagged (refer to definitions in section 5.2). Subjects who quit took longer to search on each completed topic (an average of 9.5 minutes per search, vs. 6.5 minutes; $F(1,32)= 10.3$, p<.01), and found more good items during each completed search (an average of 2.3 items vs. 1.6 items; $F(1,32)=4.4$, p<.05).

5.2 *POST-HOC* JUDGMENT OF GOODNESS

After all 36 experimental sessions were completed, the researcher judged the *goodness* of each source that had been flagged during the experiment. All sources flagged for a topic comprised the "pool" for the topic. Sources were identified by the full *urls* the subjects clicked to open them. Within each topic pool, sources were judged in alphabetical order by *url*. While making the judgments, the researcher was blind to the search conditions under which each source had been flagged. All sources in a topic pool were judged in a single session. A 4-level scale was used: *good, marginal, bad,* or *missing* (link no longer viable).

If a source covered all aspects of the topic statement, the researcher judged it as *good.* If it covered only some but not all of the aspects of the topic, it was judged

marginal. Because we instructed subjects to search for "good information sources", not "good entry pages," the researcher used the following rule when judging websites: if the entry page was not good, but the needed information could be reached using one navigational link, or if one entry in the site's search mechanism could do so, the source was judged as *good*. If a source was not about the topic the researcher judged it as *bad*. For example, a source about investing in the stock market was judged a bad source for the topic "The option to purchase a bull is an investment alternative for farmers." The distribution of the researcher's judgments, including items found by more than one subject, was 51.8% good, 19.3% marginal, 24.4% bad, and 4.5% missing.

5.3 MEASURES

5.3.a Measures based on counts

Using the data described above, we computed the measures listed in Table 5.4 for each of the 416 completed topic searches.

5.3.b A measure of elapsed time

We also computed elapsed topic time (*ETTime*) for every completed search. This measures the minutes and seconds that elapsed from submission of the initial query in a search, to the click of the final "end" button at the conclusion of a search.

5.3.c Ratio variables

We used the measures listed in Table 5.4, and ETTime, to compute the ratio variables listed in Table 5.5 (System Variables) and Table 5.6 (Searcher Variables). Appendix E details descriptive statistics for each frequency and ratio measure, for each group, in each block. Ratios were computed for system performance, system responses, searcher productivity, and search behavior. The following sub-sections review each of these sets of variables.

Table 5.4 Variable names for measures based on counts
(for each topic search, unless otherwise noted)

Measurement / description	Variable			
	Item judgment			All
	Bad	Marginal	Good	
Item displays *(includes repeated displays of the same item)*				
# of item displays	--	*MIDs*	*GIDs*	*AIDs*
# of flagged item displays	--	*MFIDs*	*GFIDs*	*AFIDs*
Items displayed *(excludes repeated displays of the same item during topic search)*				
# of items	--	*MI*	*GI*	*AI*
# of flagged items	*BFI*	*MFI*	*GFI*	*AFI*
# of items in topic "pool"*	--	*MTI*	*GTI*	--
Other measures – system related				
# of empty lists received (*0 items*)				*EBRec*
# of short lists received (*1 through 19 items*)				*SBRec*
# of full lists received (*20 items*)				*FBRec*
Other measures – searcher related				
# of queries submitted				*QCount*
# of space-delimited query terms				*QTerms*
# of spelling messages received				*SpMess*

* for each topic

5.3.c.i Measures of system performance (see Table 5.5)

Measures of retrieval system performance generally use *relevance* judgments to assess the utility of retrieved items. We did not ask our subjects to assess the *relevance* of information sources. Rather, we asked them to indicate whether a source was one they "could and *would* use to get information about the topic"; that is, they were asked to indicate whether the source seemed good *to them*. For this reason, we measure the relative performance of the three systems using the presence of *good items*, not relevant items. A measure of precision, *GPrec*, is the fraction of all item displays that are good items. A measure of recall, *GRec*, is the fraction of all good items in the topic pool that are displayed at least once during the search.

Table 5.5. Ratio measures: system variables
(for each topic search; refer to Table 5.4 for acronyms under "Ratio")

Variable name	Ratio	Description
System Performance		
GPrec	*GIDs / AIDs*	fraction of item displays that are good items
GRec	*GI / GTI*	fraction of all known good items for the topic that are displayed during the search
System Response		
Average-list-length	*AIDs / QCount*	average length of a displayed list
Fraction-full-lists	*FBRec / QCount*	fraction of queries submitted returning a 20 item list
Fraction-empty-lists	*EBRec / QCount*	fraction of queries submitted returning a 0 item list
Fraction-short-lists	*SBRec / QCount*	fraction of queries submitted returning a 1 to 19 item list
Unique-items-per-query	*AI / QCount*	average number of unique items displayed per query submitted
Item-display-repetitions	*(AIDs – AI) / AIDs*	fraction of item displays that repeat a previously displayed item

5.3.c.ii Measures of system response (see Table 5.5, above)

We use "system response" to describe how our manipulations affected results

returned to searchers. All of these measures are observable by a user. Over the course of a

user's session, all are also observable by a suitably instrumented system.

Measures of system response include:
- the average length of lists returned to searchers (*average-list-length*),
- the fraction of queries submitted for which the system returns a 20 item list (*fraction-full-lists*),
- the fraction of queries submitted for which the system returns a page with no items listed (*fraction-empty-lists*),
- the fraction of queries submitted for which the system returns a list with 1 to 19 items (*fraction-short-lists*),
- the average number of unique items displayed per query submitted (*unique-items-per-query*), and
- the fraction of item displays that repeat a previously displayed item (*item-display-repetitions*).

Table 5.6 Ratio measures: searcher variables
(for each topic search; refer to Table 5.4 for acronyms under "Ratio")

Variable name	Ratio	Description
Searcher Productivity		
Good-item-ratio	GFI / AFI	fraction of flagged items that are good items
Marginal-item-ratio	MFI / AFI	fraction of flagged items that are marginal items
Bad-item-ratio	BFI / AFI	fraction of flagged items that are bad items
Good-item-detection-rate[1]	$GFIDs / GIDs$	fraction of the good source displays that are flagged
Search Behavior		
Query-rate	$QCount / ETTime$	number of queries submitted per minute of elapsed topic time
Flagging-rate	$AFIDs / AIDs$	fraction of item displays flagged by searcher
Average-query-length	$QTerms / QCount$	average number of space-delimited terms per query
Spell-message-per-query	$SpMess / QCount$	average number of spelling error messages received per query

5.3.c.iii Measures of searcher productivity (see Table 5.6, above)

Measures of searcher productivity characterize the searcher's success in finding as many good sources as possible, and as few bad sources as possible, during a topic search. The *good-item-ratio* is the fraction of flagged items that the researcher subsequently judged as "good" (see section 5.2). Similarly, the *marginal-item-ratio* and the *bad-item-ratio* are the fraction of flagged items assigned to each corresponding category. The *good item detection rate*[1] measures how often a searcher flags a good item that is displayed.

---

[1] Findings for *good-item-detection-rate* cannot be interpreted due to complexities in the experimental design.

5.3.c.iv Measures of search behavior (see Table 5.6, above)

Measures of search behavior describe a searcher's actions during a topic search.
The *query-rate*, the number of queries submitted per minute, measures the pace of the
searcher's query submissions. The *flagging-rate* is the fraction of item displays that a
searcher flags, which we use as an indicator of a searcher's propensity to flag items. The
average-query-length is the average number of words in the queries submitted during a
topic search, with words delimited by white space. We measure a searcher's typos and
misspellings, as detected by Google, as the *spelling-message-per-query* ratio. With the
exception of the flagging-rate, a suitably instrumented system may directly observe all of
these measures of behavior.

5.3.d Topic search averages

The 416 completed topic searches are not dissimilar from searches conducted in
other experimental settings. The average elapsed time for each topic was 6.5 minutes,
with the shortest search taking 1.5 minutes, and the longest taking 22.7 minutes. The
average number of queries submitted was 5.5, with the lowest being 1 and the highest
being 30.

5.4 ANALYSIS PLAN

Our data analysis has two phases, an initial planned phase, and an exploratory
phase. The first phase uses planned contrasts to examine effects due to system
performance. This involves three steps. First, incidental effects are removed from each
measurement and the remaining value is saved for analysis. In the second step, we test
the performance of the three systems (control, BR and MR) to confirm that our
experimental manipulations caused degraded performance. Having confirmed that our
manipulations produced their intended effect, we use contrast analysis to answer our first

three research questions: (1) How does search behavior change when system performance is degraded? (2) How does searcher productivity change when the system is degraded? and (3) How do system responses change as a result of the experimental manipulations? Details of these three steps are presented in the following sections of this chapter, and results of the contrast analysis are presented in Chapter 6.

The second phase of analysis uses the data prepared for the first analysis, and more detailed measures of query behavior and system responses. These data, and variables derived from them, are presented at the beginning of Chapter 7. The second analysis phase explores the relationships set out in our final research question: How are system performance, system response, and search behavior interrelated? More specifically, we use the General Linear Model (GLM) to explore questions raised in the first phase of analysis.

5.5 ANALYSIS PHASE 1: CONTRAST ANALYSIS

5.5.a Data preparation: extraction of incidental effects

In preparing our data for analysis, we extract three incidental effects from every measure, for each search, including: (1) the topic of the search, (2) the subject conducting the search (user), and (3) the position of the search in the set of 12 searches completed during the session. We use the GLM to estimate these effects (Sun & Kantor, 2006; Wacholder et al., 2007). The model relates any specific measure y, resulting from the actions of a user u, working on the pth search in the session, searching on a topic t, under a system treatment s. The equation is:

$$y_{upts} = \lambda_u^{(U)} + \lambda_p^{(P)} + \lambda_t^{(T)} + \lambda_s^{(S)} + \varepsilon \tag{1}$$

The λs represent the main effects, which we model as fixed. The term ε represents random variation not accounted for by the model.

To focus on system effects, we subtract the incidental effects of user, position, and topic from our data prior to analysis (see Appendix F for effect sizes for each incidental factor). The values remaining after extraction represent effects due to our manipulation of the system during the treatment block, and random error, as:

$$y_{upts} - \left(\lambda_u^{(U)} + \lambda_p^{(P)} + \lambda_t^{(T)} \right) = \lambda_s^{(S)} + \varepsilon \qquad (2)$$

We save these remaining values for analysis. Using P-P plots (Maxwell & Delaney, 2004) we inspected the distribution of these values for each measure and found all to be normally distributed, or nearly normally distributed in the case of average-list-length. We use a set of planned contrasts to analyze the saved values, adjusting the degrees of freedom to account for the variability removed from the data (Keppel & Wickens, 2004).

5.5.b Contrast analysis

The contrasts test a set of first order and second order differences for each measure of interest. The first order difference (d_v) is the within-group, block-to-block change in the average of a measure v for a group. The second order difference $\Delta\_v_{treatment}$ is the between-group measure of the difference between the first-order difference for the control group (d_v)$_{control}$, and the first-order difference for one of the two treatment groups (d_v)$_{treatment}$, given as:

$$\Delta\_v_{treatment} = (d\_v)_{treatment} - (d\_v)_{control} . \qquad (3)$$

The second-order difference $\Delta\_v_{compare}$ is the between-group difference for the two treatment groups, as:

$$\Delta\_v_{compare} = (d\_v)_{BR} - (d\_v)_{MR} . \qquad (4)$$

5.6 CONFIRMATION OF SYSTEM PERFORMANCE DEGRADATION

Contrast analysis confirms that system performance was degraded in both the MR and BR systems (see Table 5.7). For the BR group, relative to the pre-treatment block in which the standard system was used, GPrec decreased in the treatment block. The block-to-block change in GPrec is significantly different from the corresponding block-to-block change in GPrec for the control group (v=$GPrec$: Δ_v_{BR} = -0.044, $F(1,354)$=4.2, p<.05). Similarly, the manipulation in the MR system resulted in lower GPrec in the treatment block and the change is significantly different from the block-to-block change for the control group (v=$GPrec$: Δ_v_{MR} = -0.045, $F(1,354)$=4.3, p<.05). Results are similar for GRec, with lower GRec in the treatment block for both the BR group (v=$GRec$: Δ_v_{BR} = -0.080, $F(1,354)$=13.9, p<.001), and the MR group (v=$GRec$: Δ_v_{MR} = -0.091, $F(1,354)$=18.0, p<.001).

Table 5.7 Contrasts: system performance

	$d\_v \pm s.e.m$	$\Delta\_v \pm s.e.m$
	v=GPrec	
Control	$0.03 \pm .018$	-------
BR	$-0.014 \pm .012$	$-0.044 \pm .022$ *
MR	$-0.015 \pm .011$	$-0.045 \pm .022$ *
compare	-------	$0.001 \pm .017$
	v=GRec	
Control	$0.057 \pm .013$	-------
BR	$-0.023 \pm .013$	$-0.080 \pm .019$ ***
MR	$-0.034 \pm .014$	$-0.091 \pm .020$ ***
compare	-------	$0.011 \pm .019$

* α=.05, ** α=.01, *** α=.001

6. KEY FINDINGS: THE EFFECTS OF PERFORMANCE DEGRADATION

This chapter reports on our key findings from the first phase of analysis, including analyses of system responses, searcher productivity, and search behavior. The chapter concludes with a summary of these findings and general comments.

6.1 SYSTEM RESPONSE CHARACTERISTICS

The manipulations of the starting ranks resulted in degraded performance, however, we did not tell our subjects about the change in rankings or performance. Of course, subjects did receive information from the changes they observed in the system's responses. We explore three types of system response (see Table 6.1): (1) the length of results lists returned, (2) the relative frequency with which the same item appears in multiple lists, and (3) the number of unique items returned per query submitted.

6.1.a Length of results lists

The manipulation of the starting ranks resulted in significant differences in the length of results lists returned (see Table 6.1). For the BR group (rankings 300 – 319), the block-to-block change in *average-list-length* is significantly different from the block-to-block change for the control group (v= *average-list-length*: Δ_v_{BR} = -1.17, $F(1,354)$=4.4, p<.05). The difference is not significant for the MR group (mixed-rankings). For the BR group, a smaller fraction of the lists returned were "full" 20-item lists; the block-to-block change in *fraction-full-lists* is significantly different from the block-to-block change for the control group (v= *fraction-full-lists*: Δ_v_{BR} = -0.098, $F(1,354)$=6.8, p<.01). The difference is not significant for the MR group. For both the BR and MR systems, a larger fraction of lists returned were truncated (lists with at least one item, but fewer than 20). For both groups, the block-to-block change in *fraction-truncated-lists* is significantly different from the block-to-block change for the control group (v= *fraction-*

Table 6.1 Contrasts: system responses

	$d\_v \pm s.e.m$	$\Delta\_v \pm s.e.m$
v=average-list-length		
Control	$0.53 \pm .28$	-------
BR	$-0.64 \pm .43$	$-1.17 \pm .51*$
MR	$0.12 \pm .29$	$-0.41 \pm .40$
v=fraction-full-lists		
Control	$0.051 \pm .017$	-------
BR	$-0.046 \pm .031$	$-0.098 \pm .035**$
MR	$-0.005 \pm .021$	$-0.056 \pm .027$
v=fraction-truncated-lists		
Control	$-0.046 \pm .012$	-------
BR	$0.031 \pm .026$	$0.077 \pm .029**$
MR	$0.016 \pm .020$	$0.062 \pm .023*$
v=fraction-empty-lists		
Control	$-0.005 \pm .013$	-------
BR	$0.016 \pm .019$	$0.021 \pm .023$
MR	$-0.010 \pm .010$	$-0.005 \pm .016$
v=item-display-repetitions		
Control	$0.063 \pm .021$	-------
BR	$-0.022 \pm .018$	$-0.085 \pm .027**$
MR	$-0.041 \pm .016$	$-0.104 \pm .026***$
v=unique-items-per-query		
Control	$-0.78 \pm .46$	-------
BR	$-0.16 \pm .50$	$0.62 \pm .68$
MR	$0.93 \pm .40$	$1.71 \pm .61*$

* $\alpha=.05$, ** $\alpha=.01$, *** $\alpha=.001$

truncated-lists: $\Delta\_v_{BR} = 0.077$, $F(1,354)=7.0$, $p<.01$; $\Delta\_v_{MR} = 0.062$, $F(1,354)=4.5$, $p<.05$). For both treatment groups, there is no significant difference in the fraction of lists returned empty (BR: $F(1,354)=0.8$, $p>.35$; MR: $F(1,354)=0.05$, $p>.8$). For all four measures of list length, there are no significant differences between the block-to-block changes for the BR group and the block-to-block changes for the MR group.

6.1.b Item display repetitions

For both treatment groups, during the treatment block, there was a significant decrease in the fraction of item displays that repeat a previously displayed item (see

Table 6.1, above). For both treatment groups, the block-to-block change in item-display-repetitions is significantly different from the block-to-block change for the control group ($v=$ *item-display-repetitions*: Δ_v_{BR} = -0.085, $F(1,354)$=7.9, p<.01; Δ_v_{MR} = -0.104, $F(1,354)$=11.9, p=.001). For the MR group, the block-to-block change in the number of unique items displayed per query is significantly different from the block-to-block change for the control group ($v=$ *unique-items-per-query*: Δ_v_{MR} = 1.71, $F(1,354)$=5.2, p<.05); the difference is not significant for the BR group. For both measures, there are no significant differences between the block-to-block changes for the BR group and the block-to-block changes for the MR group.

6.2 SEARCHER PRODUCTIVITY

Four basic measures of searcher productivity were examined: the number of good sources flagged during a topic search (*good-flagged-items*), the number of bad sources flagged (*bad-flagged-items*), the number of marginal sources flagged (*marginal-flagged-items*) and time spent searching (*elapsed-topic-time*). For all four measures, there are no significant differences between the block-to-block changes for the control group and the block-to-block changes for either treatment group. Table 6.2 details these results.

We also explored three measures of searcher accuracy, an aspect of productivity, (see Table 6.2). We defined accuracy as the fraction of flagged items falling in each of the three "goodness" categories, as judged subsequently by the researcher: *good-item-ratio*, *marginal-item-ratio*, and *bad-item-ratio*. For all three ratios, there are no

Table 6.2 Contrasts: searcher productivity

	$d\_v \pm s.e.m$	$\Delta\_v \pm s.e.m$
v=good-flagged-items		
Control	$0.319 \pm .318$	-------
BR	$-0.285 \pm .254$	$-0.604 \pm .408$
MR	$-0.035 \pm .310$	$-0.354 \pm .444$
v=marginal-flagged-items		
Control	$0.062 \pm .132$	-------
BR	$-0.208 \pm .163$	$-.271 \pm .209$
MR	$0.146 \pm .149$	$0.083 \pm .199$
v=bad-flagged-items		
Control	$.028 \pm .213$	-------
BR	$-0.160 \pm .229$	$-0.188 \pm .313$
MR	$0.132 \pm .208$	$0.104 \pm .298$
v=elapsed-topic-time		
Control	$0.209 \pm .432$	-------
BR	$0.174 \pm .354$	$-0.034 \pm .559$
MR	$-0.383 \pm .393$	$-0.592 \pm .584$
v=good-item-ratio		
Control	$0.063 \pm .069$	-------
BR	$-0.085 \pm .065$	$-0.148 \pm .095$
MR	$0.014 \pm .073$	$-0.049 \pm .101$
v=marginal-item-ratio		
Control	$-0.065 \pm .051$	-------
BR	$0.013 \pm .056$	$0.078 \pm .076$
MR	$0.056 \pm .052$	$0.120 \pm .073$
v=bad-item-ratio[1]		
Control	$-0.026 \pm .058$	-------
BR	$0.044 \pm .058$	$0.070 \pm .083$
MR	$-0.015 \pm .063$	$0.012 \pm .086$
v=good-item-detection-rate		
Control	$-0.262 \pm .068$	-------
BR	$0.091 \pm .084$	$0.352 \pm .108$ ***
MR	$0.233 \pm .063$	$0.495 \pm .093$ ***

* α=.05, ** α=.01, *** α=.001

significant differences between the block-to-block changes for the control group and the block-to-block changes for the treatment groups[1].

---

[1] Smith & Kantor (2008) report an erroneous significant difference in *bad-item-ratio* for the BR group. After publishing the paper, we reran all computations and found the error that had occurred during the extraction of subject, topic, and position effects.

We examined the *good-item-detection-rate* of searchers (see Table 6.2, above). For both treatment groups, the block-to-block change in good-item-detection-rate is significantly different from the block-to-block change for the control group ($v=$ *good-item-detection-rate*: $\Delta\_v_{BR} = 0.352$, $F(1,259)=11.6$, p=.001; $\Delta\_v_{MR} = 0.495$, $F(1,259)=24.4$, p<.001). We reported this result in Smith & Kantor (2008), and Smith (2008). We cannot readily interpret this result due to complexities in the experimental design.

Finally, for all eight measures of searcher productivity, there are no significant differences between the block-to-block changes for the BR group and the block-to-block changes for the MR group.

6.3 SEARCH BEHAVIOR

We examined five measures of search behavior (see Table 6.3). During the treatment block, subjects in the BR group increased their pace of query submission (queries per minute); the block-to-block change in *query-rate* is significantly different from the block-to-block change for the control group ($v=$ *query-rate*: $\Delta\_v_{BR} = 0.306$, $F(1,354)=6.2$, p<.05). The difference is not significant for the MR group.

We also examined the *flagging-rate* and *average-query-length*, both of which show no significant differences between block-to-block changes for the control group and block-to-block changes for either treatment group. A measure of spelling or typing errors, *spell-message-per-query*, was also examined. For the MR group the block-to-block change in spelling-message-per-query is not significantly different from the change for control, however, for the BR group the difference is weakly significant at the .10 level ($v=$ *spell-message-per-query*: $\Delta\_v_{BR} = 0.306$, $F(1,354)=3.2$, p=.076). For all four

Table 6.3 Contrasts: search behavior

	$d\_v \pm s.e.m$	$\Delta\_v \pm s.e.m$
	v=query-rate	
Control	$-0.135 \pm .058$	-------
BR	$0.171 \pm .094$	$0.306 \pm .111 *$
MR	$-0.036 \pm .068$	$0.099 \pm .090$
	v=flagging-rate	
Control	$0.002 \pm .014$	-------
BR	$-0.008 \pm .010$	$-0.010 \pm .017$
MR	$0.006 \pm .007$	$0.004 \pm .015$
	v=average-query-length	
Control	$0.127 \pm .223$	-------
BR	$-0.339 \pm .206$	$-0.465 \pm .303$
MR	$0.212 \pm .176$	$0.085 \pm .284$
	v=spell-message-per-query	
Control	$0.041 \pm .015$	-------
BR	$-0.025 \pm .021$	$-0.066 \pm .026 \wedge$
MR	$-0.015 \pm .017$	$-0.056 \pm .023$

$\wedge$ α=.10, * α=.05, ** α=.01, *** α=.001

measures, there are no significant differences between the block-to-block changes for the BR group and the block-to-block changes for the MR group.

6.4 SUMMARY OF PHASE 1 RESULTS BY EXPERIMENTAL SYSTEM

6.4.a Bottom-rankings system

Compared to searchers using the standard system, searchers using the bottom-rankings system receive results lists that are shorter, on average; a larger fraction of the lists are truncated and a smaller fraction are full 20-item lists. There is no evidence that searchers are more likely to receive an empty list. Fewer item displays repeat previously displayed items. There is no evidence that the BR system affects searcher productivity. There is weak evidence that searchers receive fewer spelling error messages when using the BR system.

When using the BR system, subjects submit more queries per minute. Figure 6.1

charts the average query-rate for each of the 12 search positions in the experiment for

each group. We have removed only subject and topic effects from these measurements

and effects due to the position of the search remain. The chart shows that the pace of

query submission tends to increase over the course of the experiment; the trend is most

clear in the control group. Figure 6.2 shows the block-to-block change in query-rate for

the first two blocks, using the data from Figure 6.1. The increasing pace is evident for all

three groups. It is also clear that the block-to-block increase is larger for the BR group.

Figure 6.3 shows the block-to-block change *after position effects have been extracted*.

The contrast analysis tests this change. The increase in query-rate is evident in the chart.
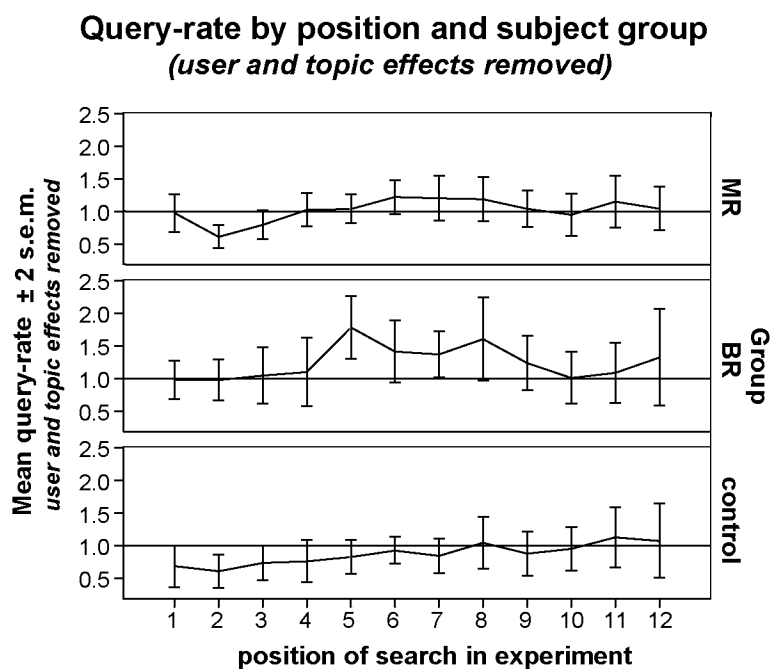
Figure 6.1. Query-rate by position and subject group: user and topic effects removed (n=416).

Query-rate by block and subject group
(user and topic effects removed)
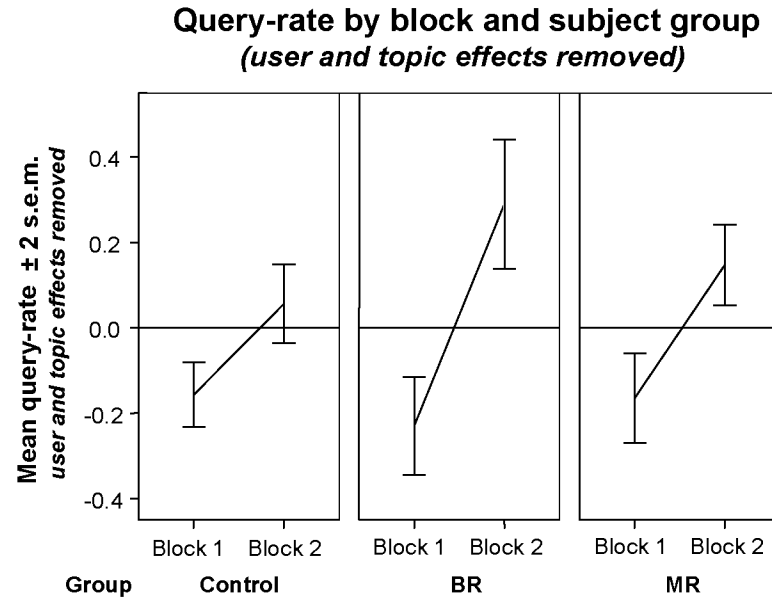
Figure 6.2. Query-rate by block and subject group: user and topic effects removed (n=288).

Query-rate by block and subject group
(user, topic and position effects removed)
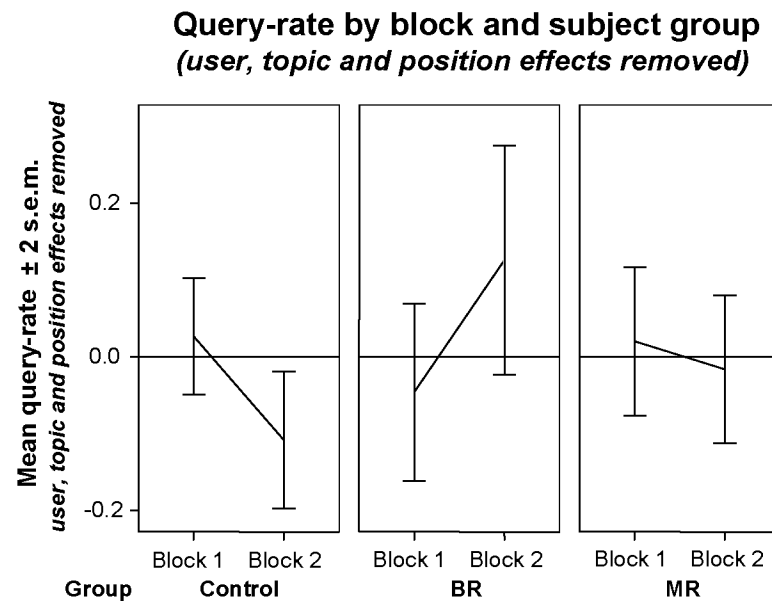
Figure 6.3. Query-rate by block and subject group: user, topic, and position effects removed (n=288).

6.4.b Mixed-rankings system

Compared to searchers using the standard system, searchers using the mixed-rankings system are more likely to receive results lists that are truncated. There is no evidence that the average length of results lists is different, or that the fraction of full or empty lists is different. Fewer item displays repeat previously displayed items. Each query is more likely to return an item that will be displayed only once during the search. There is no evidence that the MR system affects productivity. There is no evidence that the system affects query-rate, or the likelihood of receiving a spelling error message.

6.4.c General comments

One way in which a searcher using a poor system might remain productive, in the context of this experiment, is by simply flagging more items on each list, with the expectation that at least some will be good. This strategy would be apparent if flagging-rates increased during the second block, however, we find no evidence that subjects did this and conclude that they did not use this strategy. Alternatively, a searcher might remain productive by changing his or her judgment criteria for a good information source, and as a result, flag marginal or bad information sources more frequently. This strategy would be apparent if the fraction of marginal and/or bad items increased during the second block, but we also find no evidence that subjects used this strategy. To make these two strategies unappealing to subjects, we designed our experimental task with a disincentive for flagging bad items, as stated in the "job description" (see page 56):

> At the newspaper, there is a bonus for finding only good information sources. The journalists judge whether the sources found by trainees are good. The five trainees who find the most good information sources and the fewest "bad" sources, are eligible to win a "bonus."

This aspect of the protocol appears to have produced its intended effect.

7. RESULTS FROM PHASE 2 ANALYSIS: THE RELATIONSHIP BETWEEN LIST-LENGTH AND INTER-QUERY TIME INTERVAL

The results presented above describe what happened when we manipulated Google's performance by displaying results from low rankings. As planned, the manipulations caused degraded performance, however, we also found an unplanned effect; for both degraded systems, the results lists returned were more likely to be truncated, although there was no significant change in the probability of receiving an empty list. For those using the BR system, the effect was pronounced; lists were shorter on average, and were less likely to be full 20-item lists. Those using the BR system also increased the pace of query submissions. These findings raise three questions. (1) Are shorter lists an artifact of the manipulation of the starting ranks, or do they result from search behavior, or both? (2) If shorter lists do result from search behavior, what is the relationship between manipulation of the system and that behavior? (3) How do shorter lists affect the pace of query submission, if at all? This final chapter addresses these questions in three exploratory analyses.

We organize the chapter as follows. First, we present variables used in the analyses. Second, we examine the relationships among query behavior, system manipulation, and list-length, in answering the first question above. Third, we explore the relationships among system manipulation, query errors, and Google's error messages, in addressing the second question. Finally, we analyze inter-query time intervals (IQTI), in answering the third question. The chapter concludes with a summary of our answers.

7.1 QUERY DATA

This analysis uses measurements for all of the 2,295 queries submitted, and results lists received, during all 416 valid searches completed in the three blocks of the

experiment. We have *not* subtracted user, topic, and position effects from the data. The factors and measures used in our analysis include, for each query:

- *IQTI* (inter-query time interval, in minutes and seconds, for all queries except the 1st query in every topic search)

- *subject, topic* and *position* of the topic search (topic-search factors)

- *query-length* (number of space-delimited query terms)

- *query-position* (the position of the query in the sequence of queries submitted during the topic search, e.g., 1st, 2nd, 3rd, etc.)

- *spelling-error-flag* (a binary variable indicating whether a spelling error message was displayed on the results list)

- *empty-error-flag* (a binary variable indicating whether an error message was displayed on the results list)

- *rank-group* (two "bins" that group the queries by the starting rank of results displayed)
    - *top* = starting rank at 1 (n=1,691)
    - *low* = starting ranks 120 or 300 (n=604[1])

- *list-length* of the results list returned (0 through 20 items)

- *list-type* (a categorical variant of list-length, which groups the queries into three "bins" by the length of the results list)
    - *full* = a list with 20 items
    - *truncated* = a list with at least 1 but no more than 19 items
    - *empty* = a list with 0 items

- *flagged-item-displays* (number of flagged items displayed on the list[2])

---

[1] For the MR group, during the second block, lists were returned from rank 120 on any 3rd (n=41) or 8th (n=10) query submitted. In order to simplify the model, these 51 queries are combined with the 553 queries returned from rank 300 (total n=604). Collapsing across ranks does not affect the significance of the ANOVAs reported.

[2] It is important to note that flagged-item-displays includes *every* display of an item that was flagged and not only the displays on which it was first flagged by the subject. This is because we recorded, for any item, only the *last* flag status given by the subject, and not the flag status of every display.

**Average List-length
by Rank-group**
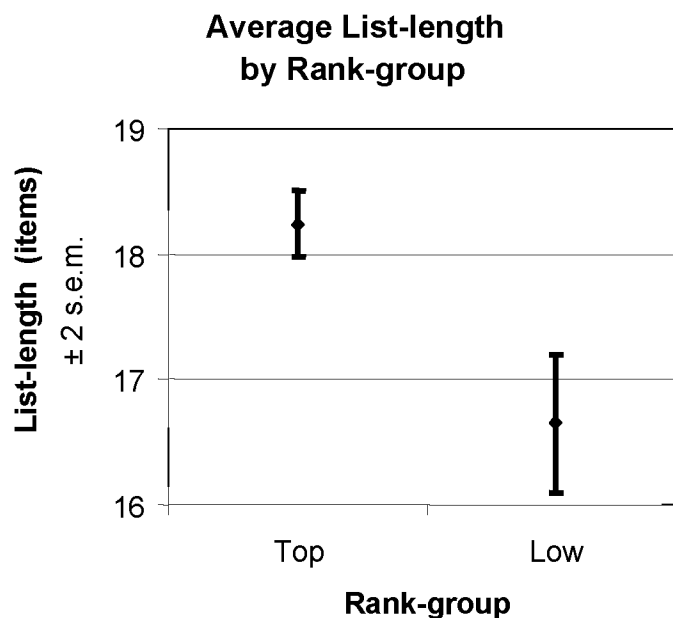
Figure 7.1. Average list-length by rank-group (n= 2,295).

7.2 EXPLORATORY ANALYSIS OF LIST-LENGTH

Figure 7.1 depicts the average length of results lists returned from low rankings

(mean = 16.65) and top rankings (mean = 18.24); lists returned from low rankings are

significantly shorter than lists returned from top rankings (ANOVA, $F(1,2293)$= 33.7,

p<.001). Because the shorter lists may have affected behavior, independent of system

performance, it is important to understand why the short lists occurred, and how list-

length and query behavior are related.

The shorter lists may have resulted from manipulation of the starting ranks, or

they may have occurred because of query behavior, such as spelling errors, typing errors,

or excessively long queries. If the shorter lists are a result of the experimental

manipulations, the increase in query-rate in the BR group may be a response to short lists,

and not a response to the performance of the system. If this is the case, we have reason to

question the generalizability of our finding that query-rates increase when search is

difficult. On the other hand, if shorter lists are a result of searchers' behavior, we should

expect to see them in searches conducted on the experimental systems *and* in searches conducted on the standard system. If short lists do occur on the standard system, we can examine the relationship between list-length and query-rate, independent of the effect of the experimental manipulations.

This section of the chapter explores factors associated with shorter lists, with the objective of answering the question: Are shorter lists an artifact of the manipulation of starting ranks, or do they result from search behavior, or both? We begin with examples of truncated lists returned by Google. We then develop a regression model for list-length, including four experimental factors (spelling-error messages, starting ranks, query length, and query position) and three incidental factors (subject, topic, and topic-search position). The model helps us understand how our experimental manipulations and searchers' behavior affect the length of results lists returned.

7.2.a Examples of short lists returned from Google

A simple test of the *standard* Google system shows that spelling and typing errors result in both truncated and empty lists. For example, the standard system returns a 2-
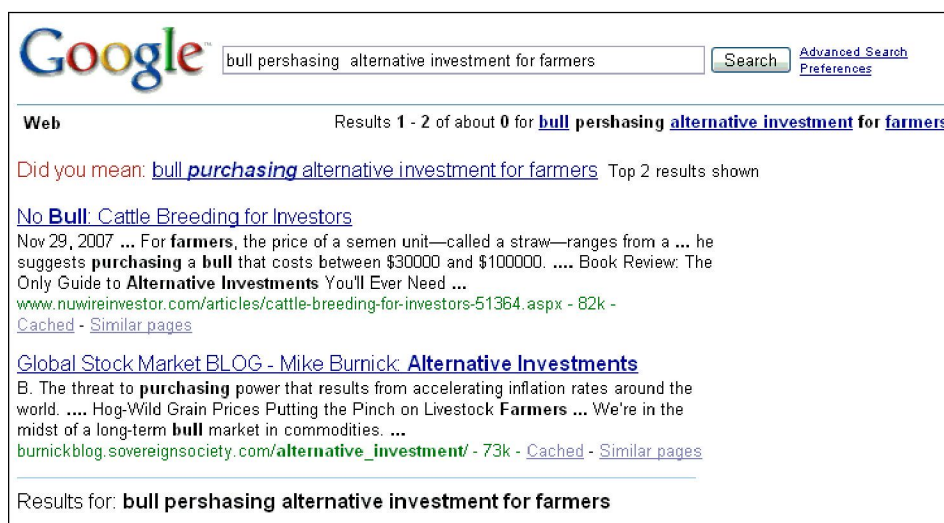
Figure 7.2. Truncated results list returned from top ranks - with spelling error.
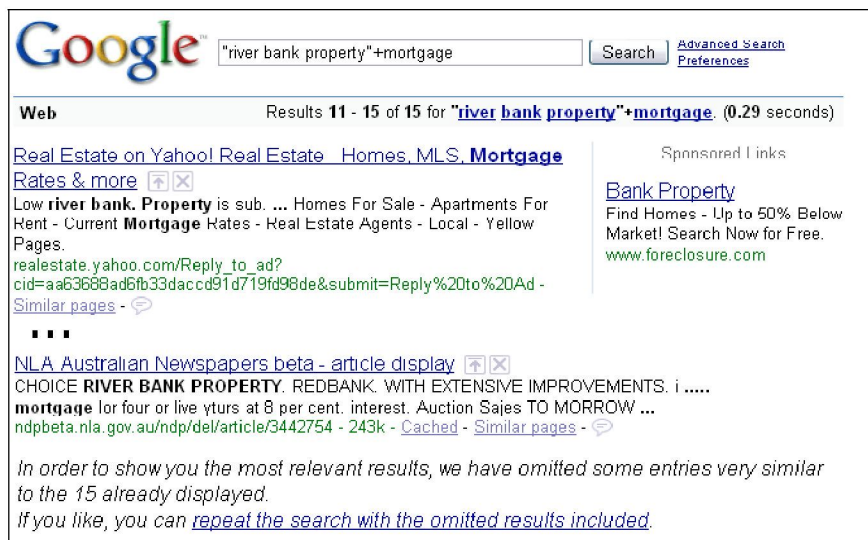
Figure 7.3. Truncated results list returned from top ranks - no spelling error.

item results list for the query "bull pershasing alternative investment for farmers" (a

query submitted by a control group subject; the results list is shown in Figure 7.2, above).

The list includes a spelling error message, and a message indicating that the two results

are a subset of results for the correct spelling of *purchasing*. A truncated list may also

occur in the standard system when there is no spelling error. The query '"river bank

property"+mortgage' (double-quotes in original, submitted by a control group subject)

returns the list depicted in Figure 7.3. While there is no spelling error, a message

indicates that there are more items available than the 15 items displayed. During the

experiment, subjects would not have seen this message at the bottom of the page.

Standard Google also returns *empty* lists. For example the spelling error in '"conductor

training" "professonal level"' (double-quotes in original, submitted by a control group

subject) returns an empty list from the standard system.

The experimental systems also return truncated lists, as in Figure 7.4. This list

was returned from rank 300 for the query '"iwbf" + "sexism"' (double-quotes in original,

Figure 7.4. Truncated results list returned from low rankings - no spelling error.

submitted by a BR group subject). When low rankings are requested from Google, the system may return a small set of items (7 in this example), along with the message about duplicate items. The message implies that Google has 26 total items, 19 of which are higher on the ranked list. We have no explanation for why Google returns this peculiar list in response to a request for the 300th item. During the experiment, subjects would have seen the 6-item list and a message indicating that the list started at the 1st item returned by Google, but *not* the message at the bottom of the list.

In summary, truncated and empty lists occur naturally in the standard system and when results are returned from low rankings. The conditions under which Google returns shorter lists are complex.

7.2.b Exploration of factors affecting list-length

In examining the lists returned during the experiment, we find that list-length has an unusual distribution: 86% of results lists are full 20-item lists. The remaining 14% comprise empty lists (7.5%) and truncated lists (6.5%). There is no discernable

**Histogram of List-length
(truncated lists)**
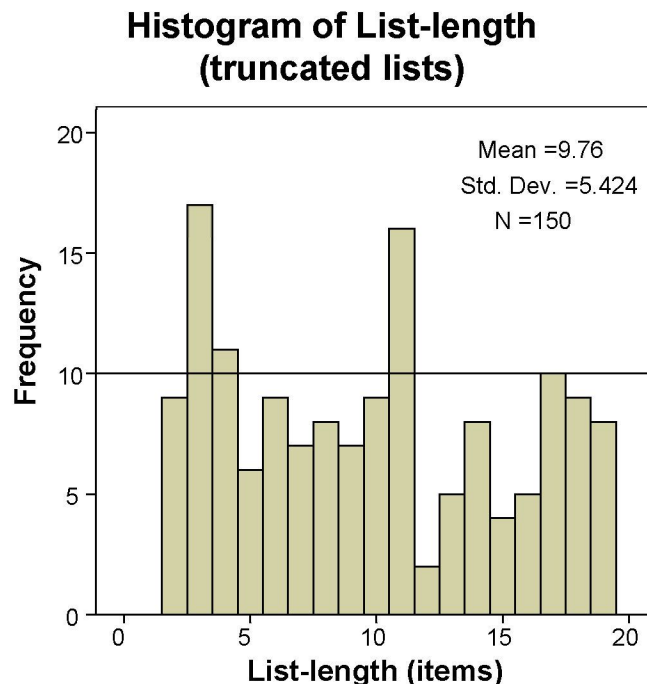
Figure 7.5. Histogram of truncated lists (length 1 through 19 items) (n= 150).

distribution pattern in the lengths of truncated lists (see Figure 7.5). As discussed above,

the conditions under which Google returns a short list (an empty or truncated list) are

complex[3]. The goal of the following analysis is not to develop a model capable of

predicting list-length, but to learn about possible causes of short lists. Specifically, the

objective is to understand how search behavior and the experimental manipulations affect

list-length.

Our exploration begins with the cross-tabulation analysis in Table 7.1. The table

separates results lists into three bins by list-length: full, truncated, and empty. It also

separates the lists by block, and within the second block, by rank-group (top vs. low). We

combine the treatment groups, MR and BR, in the table, and show the control group

separately. For each group and block, the table shows the number of lists in each bin and

the total number of lists returned. The table also shows the fraction of lists in each bin,

---

[3] Going forward, the term "short list" refers to non-full lists, those that are truncated or empty.

for each block, for each group. For example, during Block 1, the treatment groups received 632 lists; of those, 560 were full lists, or 89% of lists received by the group during the block.

For the control group, the distribution of lists into the three bins is relatively consistent over the three blocks; between 2% and 4 % of lists were truncated, and 4% of lists were empty. This is not true for the treatment groups. Two differences stand out: (1) During the second block, the lists returned from low rankings are much more likely to be truncated (14% in Block 2) than in the first block (4% in Block 1). This suggests a relationship between low rankings and truncated lists. (2) *In all three blocks*, empty lists are more likely for the treatment groups (8%, 10%, and 10% for each respective block) than for the control group (4% in each of the 3 blocks). This suggests that the relationship between group membership and empty lists is independent of rank-group.

Table 7.1 Cross-tab: fraction of results lists received in each length bin (by group, block, and starting rank)

Group		Length Bin	*rankings*				Total
			Block 1	**Block 2**		**Block 3**	
			top	*top*	*low*	*top*	**Total**
control (CONT)	# lists	*full*	230	207	--	183	620
		truncated	7	9	--	4	20
		empty	11	9	--	7	27
		total	248	225	--	194	667
	% total lists for group	*full*	93%	92%	--	94%	93%
		truncated	3%	4%	--	2%	3%
		empty	4%	4%	--	4%	4%
treatment (MR & BR)	# lists	*full*	560	23	459	310	1352
		truncated	24	1	84	21	130
		empty	48	0	61	37	146
		total	632	24	604	368	1628
	% total lists for group	*full*	89%	96%	76%	84%	83%
		truncated	4%	4%	14%	6%	8%
		empty	8%	0%	10%	10%	9%

We have two goals for the next step of our analysis: (1) to learn more about the effect of low rankings on the length of results lists, and (2) to understand why empty lists are more common among the treatment groups. We use the General Linear Model to explore factors affecting list-length, using the analysis-of-variance and model-parameters to explore the relative size and direction of effects. A fine-grained prediction of list-length is not our goal. As is demonstrated below, the analysis provides direction for further investigation.

Our analysis includes seven factors. We model the main effects of spelling-error-flag and rank-group as fixed, with query-length and query-position as covariates. The main effects of incidental factors (subject, topic, and position) we model as fixed. We do not model interactions. Results from the ANOVA are reported in Table 7.2 ($R^2 = .374$). The final model is presented in Equation 8, and parameters of the model are in Table 7.3.

As expected, the model is not a good predictor of list-length, as indicated by the scatter plot of list-length vs. predicted list-length (see Figure 7.6). A model capable of
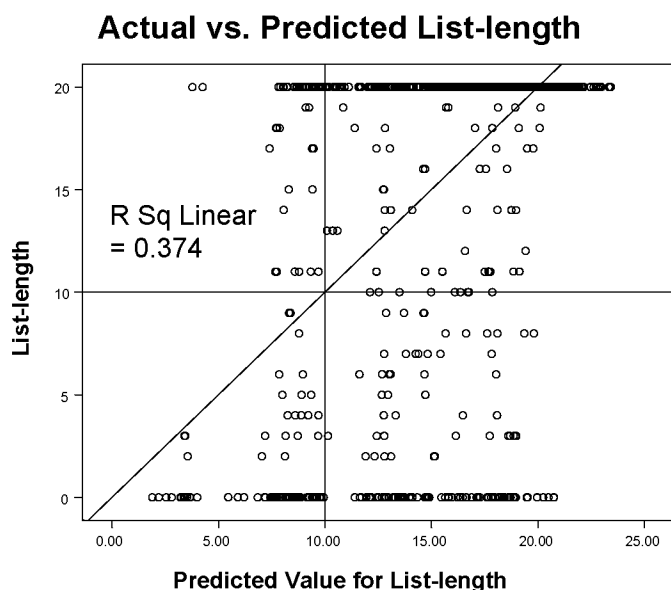
Figure 7.6. Scatter plot: predicted list-length vs. actual list-length (n=2,295).

Table 7.2 Analysis of variance for list-length (error df = 2,233; n=2,295)

variable	F	df	p	partial-η^2
spelling-error-flag	220.4	1	<.001	.090
query-length	19.6	1	<.001	.009
rank-group	11.5	1	<.001	.005
query-position	.2	1	>.600	.000
subject	23.2	35	<.001	.267
topic	3.6	11	<.001	.017
position	1.7	11	=.058	.009

Analytical Model for Length of Results Lists

$$ListLength(i, j) = 23.27 + SpellFlag * \lambda_{SpellFlag} + QueryLength * \lambda_{QueryLength} + \tag{8}$$
$$RankGroup * \lambda_{RankGroup} + \lambda_{Subject(i)} + \lambda_{Topic(j)} + error$$

Where:

ListLength(i,j)	= number of items in results list displayed for subject *i*, searching on topic *j*
23.27	= the constant in the model
SpellFlag	= 1 when spelling error message displayed, otherwise = 0
QueryLength	= number of space-delimited terms in query
RankGroup	= 1 when low rank results are displayed, otherwise = 0
Subject(i)	= subject conducting the search (i = 1 - 36)
Topic(j)	= topic of the search (j = 1 - 12)

Table 7.3 Parameters of linear model for list-length[4]

Experimental Effects	λ	Partial-η^2
SpellFlag	- 6.1 ***	.090
QueryLength	- 0.25 ***	.009
RankGroup	- 1.5 **	.005

Incidental Effects	λ lower-bounds	λ upper-bounds	Partial η^2
Subject #25	- 11.1	- 8.9	.119
Subject #27	- 5.9	- 3.3	.021
Subject # 22	- 4.4	- 1.1	.005
Topic # 11	- 4.0	- 1.9	.013
Topic # 7	- 2.5	- 0.3	.003

p* < .05 p** < .01 p*** < .001

λ is reported for significant experimental effects. For incidental factors, the upper and lower bounds of significant parameters are reported. For subject, all three significant parameters are reported. For topic, the parameters with the highest non-zero range (#7) and lowest non-zero range (#11) are reported.

---

[4] See Appendix H for subject and topic parameters.

predicting list-length would need to predict the classification of full and empty lists, while also predicting the number of items in a truncated list. Certainly, our relatively simple linear model does not accomplish this objective.

The main effect of spelling-error-flag is significant and accounts for 9% of the variance in list-length. There are also significant but smaller effects due to query-length and rank-group. Query-position has no significant effect on list-length and is not included in the model. Of the incidental factors, both subject and topic are significant. Subject is the largest effect in the model, accounting for over 26% of variance. Position is not significant and is not included in the model.

The model suggests that results lists are shorter when they are returned from low rankings ($\lambda_{RankGroup}$ = - 1.5 items, partial-η^2 = .005); this is consistent with results from the contrast analysis (see Section 6.1.a). The model also suggests that lists are shorter when queries are longer ($\lambda_{QueryLength}$= - .25 items, partial-η^2 =.009). The specific value of this parameter suggests that for every 4 terms in a query, a results list will be 1 item shorter. While this ratio is likely to be imprecise, it may be correct directionally. A one-way ANOVA comparing query-length for each list-type (full, truncated, empty) indicates that queries are longer when an empty list is returned ($F(2,2292)$=12.53, p<.001). Scheffe's *post-hoc* test shows that a query that returns an empty list is significantly longer (mean query-length$_{empty}$ = 5.06) than a query that returns full or truncated list (mean query-length$_{full}$=4.29, p<.001; mean query-length$_{truncated}$=4.07, p<.001). The difference between the query-lengths for full and truncated lists is not significant.

Importantly, the model indicates that lists are shorter when a spelling error message is returned ($\lambda_{SpellFlag}$= - 6.1 items, partial-η^2 = .090). While the specific value of

**Average List-length
(by spelling-error-flag and rank-group)**
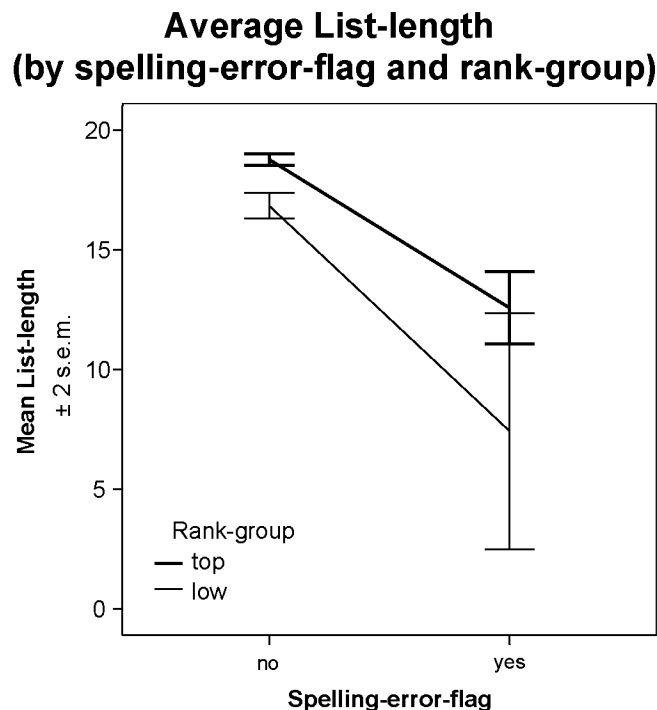
Figure 7.7. Average list-length by spelling-error-flag and rank-group (n=2,295).

this parameter may be imprecise, it may be correct directionally. Figure 7.7 (above)

shows the effects of the spelling-error-flag on list-length for each rank-group. The chart

hints at a possible interaction between spelling-error-flag and rank-group and an effect on

list-length. Indeed, when we add the interaction between spelling-error-flag and rank-

group to the model, we find a significant interaction with a very small effect

(F(2,2232)=5.5, p<.05, partial η^2=.002). However, it is important to note the large error

bar for lists returned from low-rankings when a spelling error message is returned

(misspell = *yes*). The large variability reflects the fact that *very few* of the lists returned

from low rankings contained a spelling error message. We investigate this further in

section 7.3, below.

Differences between subjects accounts for a large portion of the variance in list-

length, however, parameters for only 3 of 36 subjects are significantly different from

**Average List-length
(by subject and block)**

Mean List-length ± 2 s.e.m.

pre-treatment

block treatment

post-treatment

Subject #

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
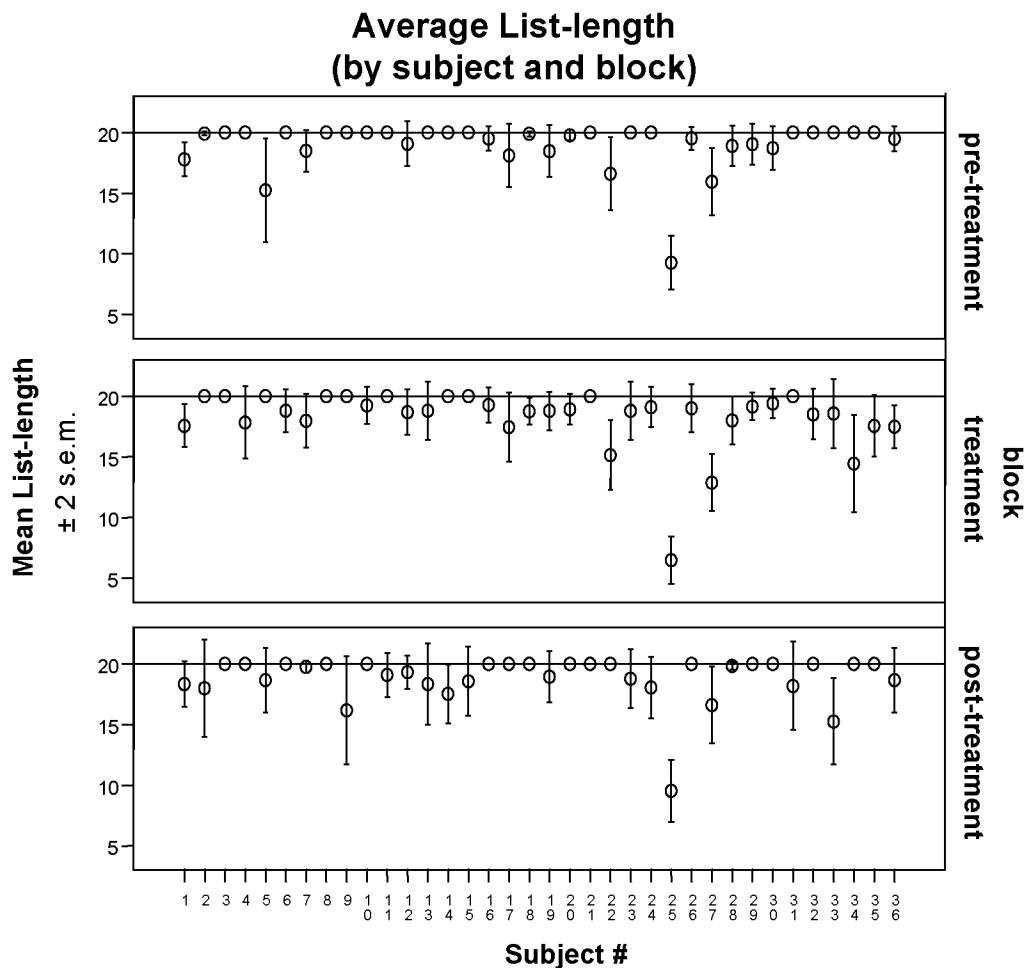
Figure 7.8. Average list-length by subject and block.

subject number 36, which has been selected by SPSS[5] as the 0-point for the factor (see Appendix H for subject parameters). These include subject number 22 from the MR group, and subjects 25 and 27, both from the BR group (see Table 7.3, above).

Subject number 25 alone accounts for 12% of the variance in list-length. Figure 7.8 (above) shows the mean and standard error of list-length for the 36 subjects for each block of the experiment. As can be seen in the chart, throughout the experiment the lists received by subject 25 are consistently very short relative to other subjects. Inspection of

---

[5] SPSS selects as the 0-point the "last" factor level in its ascending ordered list. For levels named numerically, the last factor has the highest number. For levels named with text strings, it is the last level alphabetically.

subject 25's queries reveals highly idiosyncratic query formulations. Throughout the experiment subject 25 used double-quotes in *every* query submitted. Ostensibly, the quotes delimit phrases, however, subject 25 used quotes even when a single word was submitted (e.g., "aiff"). A plus-sign (+) was also used as a concatenation device for adjacent words and phrases (e.g., "conductors" + "carry radios" + "safety"). Subject 25's queries account for fully 60% of empty lists returned to subjects in the treatment groups (see Table 7.4). If we exclude queries submitted by subject 25, the fraction of empty lists returned to the treatment groups is 3% to 4% in all three blocks, the same fraction experienced by the control group.

Subjects 22 and 27 also used idiosyncratic punctuation in their queries. Subject 22 used commas, double-quotes, or both in *every* query. Subject 27 used double-quotes and the plus-sign, although less consistently than subjects 22 and 27. Other subjects also used idiosyncratic punctuation occasionally. Punctuation, and other features of queries not explored here, may explain much of the variance in list length. Further analysis of query features would clarify these factors.

As modeled, the topic of a search has a significant effect on list-length (see Appendix H for topic parameters). All 11 topic parameters are significant because SPSS selected topic 12, "Drinking water helps you to stay well," as the 0-point in the model. Coincidentally, it is the only topic that resulted in no truncated or empty lists.

The above results suggest that *both* the characteristics of queries (a behavioral factor) and manipulation of starting ranks (the experimental treatments) affect the length of results lists. Anecdotal evidence suggests that Google's internal processing determines

**Table 7.4 Analysis of empty and truncated lists by block
for Subject #25 and treatment group**

List Bin	Block 1	Block 2	Block 3	Total
Treatment Group				
# lists				
full	560	459	310	1329
truncated	24	84	21	129
empty	48	61	37	146
total	632	604	368	1604
% lists				
truncated	4%	14%	6%	8%
empty	8%	10%	10%	9%
Subject 25				
# lists				
truncated	15	21	11	47
empty	30	36	23	89
% Treatment Group Total				
truncated	63%	25%	52%	36%
empty	63%	59%	62%	61%
Treatment Group without Subject 25				
# lists				
truncated	9	63	10	82
empty	18	25	14	57
total	561	535	313	1409
% lists				
truncated	2%	12%	3%	6%
empty	3%	5%	4%	4%

the specific length of truncated lists, and that characteristics of queries such as punctuation, affects that processing.

7.3 EXPLORATION OF ERROR MESSAGES

The above analysis shows that shorter lists do result from search behavior. This leads to our second question: What is the relationship between manipulation of the system and the behavior of searchers? The model of list-length suggests that results lists are shorter when they contain spelling error messages. Figure 7.7 (above) suggests that lists are shortest when results are returned from low-rankings and a spelling error

message is returned. We note, however, that only a small fraction of lists returned from low rankings contain spelling error messages. Of course, it is reasonable to associate spelling error messages with search behaviors such as spelling or typing errors. It is less reasonable that searchers will make *fewer* errors when the system is performing poorly and the pace of query submission is faster, as is suggested by the contrast analysis (see Section 6.3). In this section, we use cross-tabulation analysis to explore the relationship between manipulation of the system, spelling error messages, and query errors.

We have not used the chi-square test here because the queries in our dataset are not independent; the number of queries contributed to the data by each subject would bias the test. For example, the query dataset contains 195 queries from subject 25 (8.5% of all queries), and 159 queries from subject 1 (7%), but only 22 queries from subject 8 (1%) and 24 from subject 34 (1%). With the caveat that this analysis is exploratory, we make the observations below.

Table 7.5 is a cross-tabulation showing the fraction of results lists that contain spelling and empty-list messages from Google. The table separates the lists by block, and within the second block, by rank-group (top vs. low). The table also separates the lists received by the control group (CONT) from those received by the combined treatment groups.

In the first block, 8% of lists received by the two treatment groups contained a spelling error message, and the control group received the message in a comparable 9% of lists. In contrast, during the second block, the treatment groups received the message in only 2% of lists returned from low rankings, while the control group received the message in 10% of lists received during the block. Surprisingly, searchers were *less*

Table 7.5 Cross-tab: queries receiving Google error messages
(by block, starting rank, and group)

Group		Message Type	*rankings*				Total
			Block 1	**Block 2**		**Block 3**	
			top	*top*	*low*	*top*	
control (CONT)	# queries	*spelling*	21	23	--	21	65
		empty	11	6	--	4	21
		both	5	5	--	3	13
	# queries in block		248	225	--	194	667
	% queries in block	*spelling*	9%	10%	--	11%	10%
		empty	4%	3%	--	2%	3%
		both	2%	2%	--	2%	2%
treatment (MR & BR)	# queries	*spelling*	50	3	12	23	88
		empty	36	0	44	30	110
		both	9	0	0	7	16
	# queries in block		632	24	604	368	1628
	% queries in block	*spelling*	8%	13%	2%	6%	5%
		empty	6%	0%	7%	8%	7%
		both	1%	0%	0%	2%	1%

likely to receive a spelling error message for results returned from low rankings. This is consistent with the contrast analysis. Subjects in the BR group had lower spelling error rates in the second block (see Section 6.3; the difference from the control group is significant at $\alpha=.10$; $p=.076$).

There are two possible explanations for the lower rate of spelling error messages in the second block. One is that searchers adapted their query behavior when the system performed poorly and made fewer errors. For example, they may have paid more attention while formulating queries. If this is what happened, and the model for list-length is correct, the lower rate of spelling errors actually suppressed the number of truncated lists that would otherwise have been returned in the low rankings condition. This explanation is implausible, however, because truncated lists were *more* likely in the low rankings condition.

A second explanation is that Google processes spelling error messages differently for results returned from low rankings. For example, Google might return empty-list messages rather than spelling-error messages, however, this specific hypothesis does not appear to be true. In the second block, for the treatment groups, 7% of lists returned from low rankings received the empty-list message, which is consistent with the 6% and 8% of lists that received the message in the first and third blocks, respectively. In addition, the contrast analysis shows no significant block-to-block change in the frequency of empty lists for either treatment group. It is also possible that Google is simply less likely to return a spelling error message with results returned from low rankings, even when a spelling error occurs. This is the more plausible explanation. Truncated lists were *more* likely during the second block, and the model for list-length indicates that spelling error messages are associated with shorter lists, independent of rank-group.

If this second explanation is correct, it affects the model for list-length as follows. The model for list-length assumes that a spelling-error-flag accurately represents a query error, that is, a query behavior. If Google's messages do not accurately identify spelling errors made during the second block, the model attributes differences in list-length to the rank-group. If the differences are actually the result of query errors, the model underestimates the effect of query behavior, and overestimates the effect of rank-group. Further experiments are required to investigate the cause of the small number of spelling error messages received from the manipulated systems.

7.4 EXPLORATORY ANALYSIS OF INTER-QUERY TIME INTERVALS (IQTI)

The analysis in this chapter is motivated by the finding that subjects using the BR system increased their query submission rate *and* received shorter results lists. In this final section, we address our third question: How do shorter lists affect the pace of query

submission, if at all? We report on results from our exploratory analysis of the relationships between system performance, search behavior, and inter-query time intervals (IQTI).

It is important to note that IQTI includes any time spent after a click-through, when a searcher is examining an information source, therefore it is not a measure of the time spent scanning a results list. Inter-query time intervals occur only when more than one query is used during a search; for this reason, this analysis excludes 47 of the 416 searches completed during the experiment. Table 7.6 shows the fraction of total searches that we have excluded from the analysis, for each group and block.

Table 7.6 Distribution of 1-query searches excluded from the analysis

Group		Block			Total
		Pre-treatment	Treatment	Post-treatment	
Control	total completed searches	48	48	41	137
	# of searches excluded	10	6	7	23
	% of searches excluded	21%	13%	17%	17%
BR	total completed searches	48	48	42	138
	# of searches excluded	3	2	10	15
	% of searches excluded	6%	4%	24%	11%
MR	total completed searches	48	48	45	141
	# of searches excluded	0	1	8	9
	% of searches excluded	0%	2%	18%	6%
Total	total completed searches	144	144	128	416
	# of searches excluded	13	9	25	47
	% of searches excluded	9%	6%	20%	11%

Over the 369 searches in the analysis, mean IQTI is 0.745 minutes (s.d. = 0.83 minutes, minimum = 0, maximum = 7.8). As is common for time interval data (Hill & Lewicki, 2006), IQTI was found to have an exponential distribution. We transformed IQTI using $\log_{10}$ and saved the variable *t_IQTI*. We refer to untransformed IQTI as *u_IQTI*.

We use the General Linear Model to investigate five possible factors affecting t_IQTI. We model the main effects of spelling-error-flag and rank-group as fixed, and the main effects of list-length, query-position, and flagged-item-displays as covariates. The main effects of subject and topic we also model as fixed, with position modeled as a covariate. Interactions are not modeled. Results from the ANOVA for t_IQTI are reported in Table 7.7 (R^2 = .449). Two scatter plots show the fit of the model: Figure 7.9 plots actual vs. predicted t_IQTI, and Figure 7.10 shows the standardized residuals vs. predicted t_IQTI.

Table 7.7 Analysis of variance for t_IQTI
(n = 1,878; error df = 1,824)

variable	F	df	p	partial-η^2
flagged-item-displays	197.1	1	<.001	.098
spelling-error-flag	191.0	1	<.001	.095
list-length	141.7	1	<.001	.072
query-position	9.4	1	<.01	.005
rank-group	.4	1	>.5	.000
subject	9.6	35	<.001	.155
position	31.3	1	<.001	.017
topic	2.0	11	<.05	.012

t_IQTI by Predicted t_IQTI
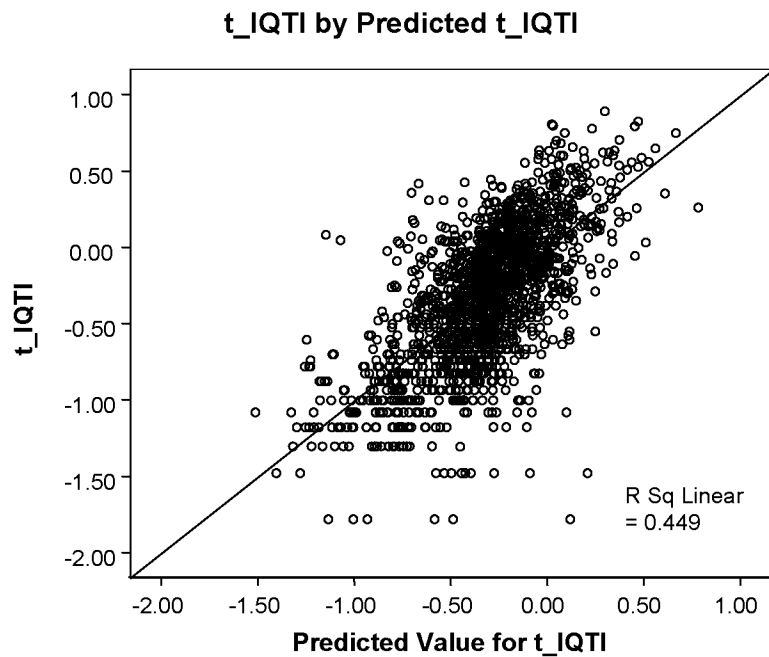
Figure 7.9. Scatter plot: predicted t_IQTI vs. t_IQTI (n=1,878).

Standardized Residual by Predicted t_IQTI
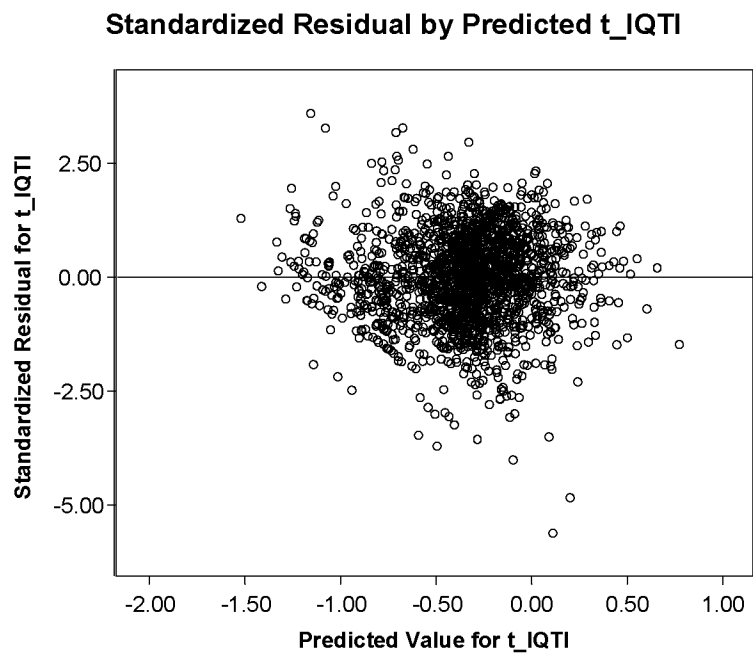
Figure 7.10. Scatter plot: predicted t_IQTI vs. standardized residuals (n=1,878).

Analytical Model for t_IQTI

$$t\_IQTI(i,j,k) = -0.601 + FlaggedDisplays\lambda_{FlaggedDisplays} + SpellFlag\lambda_{SpellFlag} +$$

$$ListLength\lambda_{ListLength} + QueryPosition\lambda_{QueryPosition} + \tag{9}$$

$$Subject(i)\lambda_{Subject(i)} + Topic(j)\lambda_{Topic(j)} + Position\lambda_{Position} + error$$

t_IQTI(i,j,k)	= log 10 of the number of minutes between query submissions for subject *i*, searching on topic *j*, in position *k*
- 0.601	= the constant in the model
FlaggedDisplays	= number of flagged item displays on list
SpellFlag	= 1 when spelling error message displayed, otherwise = 0
ListLength	= number of items on list
QueryPosition	= position of query in sequence during search (1, 2, 3, 4,...)
Subject(i)	= subject conducting the search (*i* = 1 - 36)
Topic(j)	= topic of the search (*j* = 1 - 12)
Position	= position of search in sequence of 12

Table 7.8 Parameters of linear model of t_IQTI[6]

Experimental Effects	λ	Partial-η^2
FlaggedDisplays	+ 0.115 ***	.097
SpellFlag	- 0.449***	.096
ListLength	+ 0.019 ***	.071
QueryPosition	-0.007**	.004

Incidental Effects	lower-bounds	upper-bounds	
Subject # 34	+ 0.33	+ 0.74	.014
Subject # 29	- 0.33	- 0.11	.008
Topic # 10	- 0.17	- 0.001	.002
Position	- 0.02	- 0.01	.017

p* < .05 p** < .01 p*** < .001

λ is reported for significant experimental effects. For incidental factors, the upper and lower bounds of significant parameters are reported. The one significant parameter for topic is reported. For subject, the parameters with the highest non-zero range and lowest non-zero range are reported. n= 369.

---

[6] See Appendix H for subject, topic, and position parameters.

The main effect of flagged-item-displays and spelling-error-flag are significant, with each accounting for almost 10% of the variance in t_IQTI. There is also a significant but smaller effect due to list-length, which accounts for 7% of the variance. The effect of query-position is significant, but very small. Rank-group has no significant effect. All three incidental effects are significant. Subject is the largest effect in the model, accounting for about 16% of variance. The effects of position and topic are significant but smaller. The model is presented in Equation 9, and parameters are in Table 7.8 above. See Appendix H for the full set of subject and topic parameters.

Not surprisingly, the parameters suggest that when a results list contains items good enough to warrant a flag, searchers take more time before submitting the next query ($\lambda_{\text{FlaggedDisplays}} = + 0.115 \log_{10}$ units, partial-$\eta^2 = .097$). We show this relationship in a scatter plot of t_IQTI vs. flagged-item-displays (Figure 7.11). The parameter for flagged-item-displays represents 0.115 units of $\log_{10}$ IQTI for each flagged item, as $10^{0.115} \approx 1.30 = 130\%$. Each flagged item adds about 30% to the inter-query time interval.

t_IQTI by Flagged-item-displays
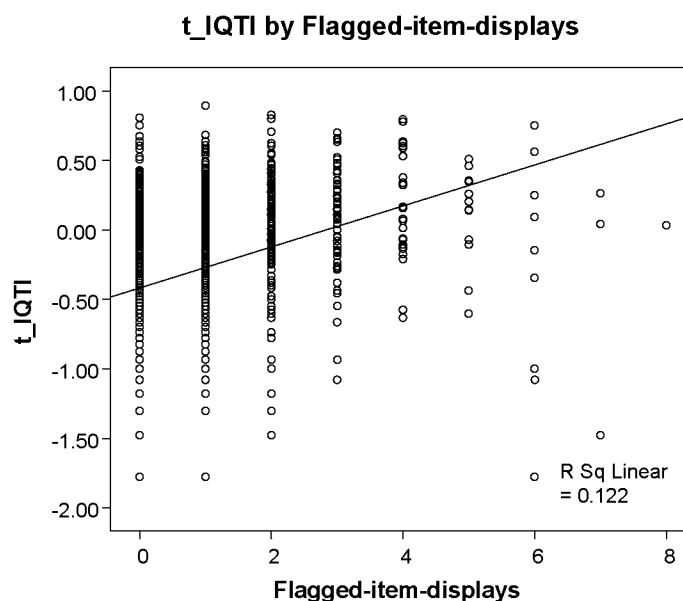
Figure 7.11. Scatter plot: flagged_item_displays vs. t_IQTI (n=1,878).

The model also suggests that when a spelling error message is returned, searchers spend less time on the list before submitting the next query ($\lambda_{SpellFlag}$= - .449 $\log_{10}$ units, partial-η^2 = .096). When a spelling error message is present, the IQTI is reduced by about 64% ($10^{-0.449} \approx 0.36 = 36\%$). More time is spent on longer lists ($\lambda_{ListLength}$= 0.019 $\log_{10}$ units, partial-η^2 =.071). The position of a query during a search has a small effect ($\lambda_{QueryPosition}$= -0.007 $\log_{10}$ units, partial-η^2 =.004). The time spent between queries tends to decrease with each query submission. Rank-group has no significant effect on t_IQTI.

As expected, effects due to differences between individual subjects are large. While the omnibus ANOVA indicates a small but significant effect due to topic, it appears that topic is not a significant factor in t_IQTI; only the parameter of Topic 5 (Fishermen find it difficult to earn a net profit) is significantly different from the 0-point. The position of a search in the order of 12 has a small but significant effect on t_IQTI ($\lambda_{Postiion}$= - 0.015 $\log_{10}$ units, partial-η^2 =.017). The time spent between queries tends to decrease over the course of the experiment.

The above results suggest that three factors have the largest effect on inter-query time intervals. The presence of an information source that is good enough to be flagged has a large effect. This makes sense intuitively; if an item is good enough to be flagged, the searcher has probably spent time clicking the item and examining it. Conversely, when there are few valuable items on a list, the searcher uses less time to examine the list before submitting the next query. Of course, these effects occur within the context of an individual searcher's average query-rate. The presence of a spelling message also has a large effect. Searchers spend relatively little time examining a list that the system has identified as a query failure. The length of a list also affects the interval; searchers spend

less time on shorter lists. Further experiments would be required to understand the details of how these factors interact to influence a searcher's behavior.

7.5 GENERAL DISCUSSION

The above results have answered two of the three questions posed at the outset of this chapter:

(1) Are shorter lists an artifact of the manipulation of the starting ranks, or do they result from search behavior, or both?

Short lists occur in the standard system. Anecdotal evidence shows that when a query contains punctuation there is a strong effect on list-length. The use of punctuation appears to be a personal preference or "query-style" for some individuals. Much of the variance in list-length is explained by differences between individual searchers. Two characteristics of queries, spelling/typing errors and the length of a query, also affect the length of a results list. Not surprisingly, query errors have a relatively large effect. Clearly, short lists do result from search behavior. On the other hand, the manipulation of the starting ranks also affected the length of results lists, but the effect is relatively small. In sum, the difference in list-length, as shown in Figure 7.1, results from *both* search behavior and the manipulation of the system, but it appears that search behavior is the largest factor.

(2) If shorter lists do result from search behavior, what is the relationship between manipulation of the system and that behavior?

This question cannot be answered with the data collected in this experiment. It appears that when the starting ranks were low, the query-error-flag may have become an unreliable measure of query errors, that is, query errors may have occurred without being identified by an error message. If this is the case, then the model for list-length underestimates the affect of query behavior. In addition, this would suggest that query-

error-flag is a poor measure of behavior because it has the potential to confound behavior and system response. It is also possible, although less plausible, that the query-error-flag is a reliable measure of query errors, and that spelling and typing improves when a searcher is faced with poor performance and increases the pace of query submission. If this is the case, it is likely that the change in behavior suppressed the number of short lists that would otherwise have been returned by the degraded systems. This would imply that the model for list-length underestimates the effect of the manipulations.

(3) How do shorter lists affect the pace of query submission, if at all?

When a results list contains items good enough to be flagged, searchers take more time before entering a new query. It makes sense that searchers allocate their time to valuable lists and to investigating good items. It follows that less time is spent on a short list, on average; a truncated list has less potential value, and there is certainly no value in an empty list. It also makes sense that this has a smaller effect on the time interval than does the presence of a flagged item; truncated lists as short at two items do sometimes have value. Figure 7.12 shows the number of flagged-item-displays relative to the length of a list, by block and rank-group.

It is likely that inter-query time intervals have been affected to some extent by the shorter lists resulting from our manipulation of the system. However, there is no way to separate the effect of shorter lists from the effect of system performance using the data collected in this experiment. Understanding the effect of list-length on IQTI would require new experiments.

Taken together, the above results suggest that *when the performance of the system is sub-standard, inter-query time intervals will be shorter*. This supports results from the

contrast analysis, which found a significant increase in query submission rates among subjects using the BR system. In addition, these results suggest that searchers use cues such as spelling error messages, or the length of a list, to make rapid judgments about the potential value of results lists.
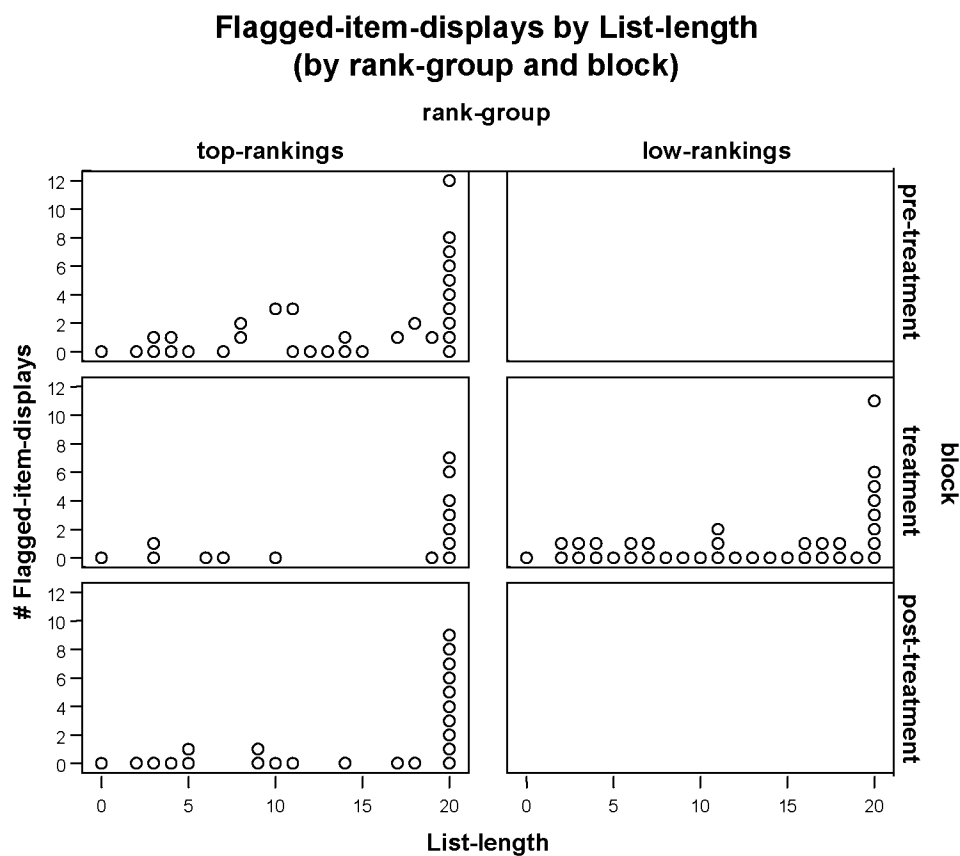
Figure 7.12. Flagged-item displays by list-length, rank-group, and block (n=416).

8. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

8.1 GENERAL DISCUSSION

As discussed in the initial chapters of this dissertation, search engine users often face query failure, particularly during complex searches. Users have learned to judge the potential value of a results list by scanning the top two or three items. Judgments of value occur very quickly. As a result, searchers often fail to recognize valuable information lower on a list. Users have also learned to overcome query failure by submitting a revised query. Because the scanning and judgment process is fast, very little time passes between receipt of the results list and submission of the next query.

As discussed in Chapter 1, searchers often have little information about the specific cause of a query failure. Generally, all that a searcher may know is that the words submitted did not work. It appears that when the system provides few or no cues about how to optimize the query, searchers can generate and submit the next query using a process as simple as rapid word association to "repair" the query. When the system communicates about the cause of a failure, as in a "clickable" spelling error message, a user can repair the query as quickly as a mouse-click. Theoretically, searchers can respond rapidly to query failure when the system presents results in the familiar best-on-top order, with a familiar snippet layout, and a "clickable" query repair. This implies that when searchers use a familiar search interface, they *speed up* when search is difficult, and productivity may remain robust in the face of poor performance.

This dissertation has explored how searchers respond to query failure. We studied the question in the context of the Google search engine, with users searching on behalf of another (the journalists), while focused on complex information needs. The study involved a single factorial experiment in which some searchers experienced particularly

difficult searches. We produced the difficulties intentionally by degrading the performance of the Google system over four consecutive searches. We examined two types of degradation. One system returned poor results consistently, while the other system returned poor results with better results presented occasionally. Both systems were the same, however, in that we did not alter the best-on-top order within results lists, nor the formatting of snippets. This preserved the informativeness of the rankings and snippet content, and did not interfere with the utility of searchers' previously learned scanning behavior. Both systems also provided clickable spelling suggestions for rapid query repair.

8.2 SUMMARY OF RESULTS

The main finding of our study is that the pace of query submission increases when system performance is consistently poor. Our detailed analysis of inter-query time intervals suggests that the increase in query-rate is due primarily to the degraded performance of the system. This finding is a major contribution to research on search behavior and its relationship to system responses. Specifically, it adds to our understanding of how users adapt their behavior to overcome query failure and remain productive searchers. Tentatively, this finding provides insight into the relationship between users' experience of the system and the predictive features found in the *SAMLight* model. The finding suggests that inter-query time intervals are predictive because they represent an adaptive response to query failure. Finally, the finding suggests that a system might be designed to monitor changes in inter-query time intervals for indications of repeated query failure. Ideally, this would provide a system with a mechanism for detecting its own performance.

We obtained these results in a carefully controlled factorial experiment. The design of the experiment allowed us to measure and detect the relatively small effect of system performance within the larger effects of search topics and individual users. Our findings were revealed in a series of analyses. We used planned contrasts to examine changes in system performance and the effect of those changes on system responses, searcher productivity, and search behavior. We used the General Linear Model to explore effects associated with the length of results lists and inter-query time intervals. The research demonstrates the use of experimental controls in the study of small effects on search behavior.

We summarize our results below:

Relative to the *standard system*, the *bottom-rankings (BR) system*:

- was degraded with respect to both precision and recall,

- returned lists that were shorter on average,

- was less likely to return a full list and was more likely to return a truncated list,

- was less likely to display an item more than once during a search.

Relative to the *control group, users of the BR system*:

- experienced no significant difference in productivity,

- increased the pace of query submissions,

- were less likely to receive a spelling error message from Google.

Relative to the *standard system*, the *mixed-rankings (MR) system*:

- was degraded with respect to both precision and recall,

- was more likely to return a truncated list,

- was less likely to display an item more than once during a search.

Relative to the *control group, users of the MR system*:

- experienced no significant difference in productivity.

Analysis of *list-length* shows that:

- Punctuation in a query affects the length of a Google results list.

- When Google returns a spelling error message, results lists are shorter.

- Longer queries are more likely to result in empty lists.

Analysis of *inter-query-time-intervals (IQTI)* shows that:

- The shorter a results list, the less time a searcher takes before submitting the next query.

- The fewer items the searcher flags on a list, the less time a searcher takes before submitting the next query.

- When Google returns a spelling error message, searchers take less time before submitting the next query.

- Inter-query-time-intervals vary widely among searchers.

8.3 CONCLUSIONS AND QUESTIONS RAISED

We draw the following conclusions from the above results:

- When a query fails, searchers quickly submit a new query. By rapidly abandoning lists with little potential value, a searcher is able to be productive in the face of poor system performance. When searching is difficult, as when queries failed repeatedly in the bottom-rankings system, query submissions occur at a faster rate.

- Inter-query-time-intervals vary considerably among searchers, and for a given searcher, IQTIs change over the course of several consecutive searches. If a

search system is to interpret query-rate as an indicator of system performance, only the *change in the time interval*, and not the size of the interval, is likely to be meaningful.

- The use of punctuation in queries is a preference, or "query style," for searchers. Different query styles may result in very different system responses, hence, very different experiences of the system. For example, punctuation affects the length of results lists and the frequency of empty lists.

Our results raise the following questions:

- Independent of the performance of the system, do short lists or error messages affect the rate of query submission?

- Do searchers make fewer typing and spelling errors when system performance is degraded?

8.4 LIMITATIONS AND FUTURE WORK

This study reports on a single exploratory experiment. Of necessity, trade-offs were made in the design of the experiment and the experimental systems. We conclude by outlining the resulting limitations and suggest remedies in future work.

- The length of results lists is a significant factor in inter-query-time-intervals. In order to understand how list-length affects behavior, independent of system performance, we must control this factor in future experiments.

- The effect of system performance was tested at three levels, however, performance was not controlled, per se. We tested in a narrow range of performance conditions. It is quite possible that the effects observed in this range are not present at more extreme levels of performance degradation, or performance enhancement. In order to

gain a deeper understanding of the relationship between behavior and system performance, it will be important to test over a larger range and steeper gradient of performance levels. Because the topic of a search has such a large effect on performance, it will be essential to develop methods for the constructing performance levels that are equivalent across topics. In addition, it will be important to investigate effects due to different *types* of performance degradation.

- The topics used in this experiment were generally all of the same kind: complex and informational. The topics had very little effect on inter-query-time-intervals. Of course, not all search topics are of this type; many are simple or navigational. A more varied set of topics may produce different effects on inter-query-time intervals. In order to investigate these effects future experiments should include different types of topics.

- In this experiment, subjects experienced a single treatment condition in a series of four searches. We detected treatment effects by analyzing differences in block averages, which combine all the searches conducted during a block, over all subjects in a group. The design increased the likelihood of discovering system effects in the noisy data of interactive search. However, outside the laboratory search systems usually do not change their performance consistently over a set of consecutive searches. Typically, performance is associated with the topic of a search (Lagergren & Over, 1998). In order to examine behavior in more realistic conditions, future experiments should use intermixed-treatment designs, interleaving treatment conditions with control conditions.

- Prior to beginning each search, subjects completed a pre-search questionnaire, which asked for an estimate of the number of good information sources they expected to find, and the number of minutes they expected the search would take (see Appendix B.3.b). This may have caused subjects to focus on their stated expectations, with two possible consequences. One, the pre-search questions may have caused subjects using a degraded system to become more persistent and willing to expend greater effort to meet the stated expectations. Two, for searches conducted using the standard system, the pre-search questions may have motivated *satisficing*; this would cause subjects to be satisfied with meeting the stated expectations with little motivation to exceed expectations by maximizing the number of sources found. The net effect of the pre-experiment questions would be to keep productivity relatively stable between conditions. Future experiments should examine the effect of questions about expectations.

- We used the chance to win a $40 bonus to motivate earnest effort from our subjects. As discussed above, the questions about expectations may also have motivated greater persistence and effort among those using a degraded system. These two aspects of the protocol may have resulted in persistence that we would not find in search behavior outside the laboratory. Future experiments should examine how the protocol affects effort, with the goal of understanding the external validity of laboratory results.

- Subjects received a maximum of 20 items on each results page returned, with no option to continue to the next page of results. This forced subjects to submit a new query when, if using the "real" Google system, they might have continued to the next

page of results without changing their query. Several studies have found that searchers rarely request the "next" results page, however, the relationship between this behavior and system performance has not been examined. Future experiments should eliminate this potential confound by allowing subjects to examine multiple pages returned from a query.

- In this experiment, we assessed the value (*goodness*) of information sources using the researcher's *post-hoc* judgments. Ideally, at least one other person would make these judgments, preferably someone completely blind to the objectives of the study. Because of funding limitations, we did not collect independent judgments. It is important, however, to examine whether the *post-hoc* judgments unintentionally biased the results. We addressed this issue by re-running the contrast analyses reported in Chapters 5 and 6. In these additional analyses we combined "good" and "marginal" sources, treating both as "good" sources in each applicable variable. With respect to the significance of each contrast, none of the results changed. Future experiments should use at least two assessors.

- It would also be ideal if we could compare inter-rater agreement on *goodness* within and between subject groups. Unfortunately, due to a problem with the experimental design, this was not possible for this study. For subjects using either degraded system (BR or MR), relative to the standard system, very few items were received by more than one subject. While 44% of items for the standard system *overlapped*[1] between subjects, only 10% of items overlapped between subjects for the degraded systems (see Appendix D). For queries submitted to the BR system, of the 471 total items received by more than one subject, only 22 were flagged by any subject, and of those,

---

[1] "Overlap" is the fraction of all items that were received by more than one subject (see Appendix D).

only 5 were flagged by more than one subject. The ratios are similar for the MR system. As a result, there are not enough examples from the degraded systems to produce an analysis of inter-rater agreement. Future experimental systems should control the frequency of item displays within each group.

APPENDIX A – Behavioral features in the SAMLight model (Downey, Dumais, & Horvitz, 2007b)

Feature type	*# of Features*	*Feature*	*1 = most predictive feature relationship to p(click)*
Temporal/ transition	4	r(SearchAct): elapsed time between two search actions	**1.** $p(click)$ decreases for longer interval between actions
		DayOfWeek: session day of week	
		TimeOfDay: 1 of 3 8-hour windows	
		qq(WordDelta): word-length change between queries	
Query	24	q(FirstResult rank of first result [on list] requested	**2.** $p(click)$ decreases for queries that request results lists starting at lower rank positions
		q(HasSuggestion) query has spelling suggestion	**3.** $p(click)$ decreases for query with spelling suggestion
		q(HasDefinitive) query has definitive result (e.g. navigation)	**5.** $p(click)$ increases for queries with definitive result (e.g. amazon.com)
		$q(c_r$Prob) probability of a click for the query	**6.** $p(click)$ increases for queries likely to result in a click
		q(WordLen): number words in query	
		q(CharLen): number characters in query	
		q(Freq): number times the query is entered	
		q(AvgCrPos): avg. results position clicked	
		q(AvgCrDelay): avg. time between query submission and click	
		q(AvgPathSec): avg. elapsed time on click-path -*PathDwellSec*	
		q(AvgPathPages): avg. pages in click-path - *PathPageLength*	
		q(AvgAfterPathSec): avg. time to next search action - *AfterPathSec*	
		q(Distinct$U$): number unique users submitting query	
		q(AdImpressions): number times advertisement is displayed	
		q(AvgNumAds): avg. number of advertisements displayed	
		q(AdBid): avg. bids on advertisements for this query	
		q(MinWordFreq): web-frequency of least frequent word	
		q(MaxWordFreq): web-frequency of most frequent word	
		q(GeoMeanFreq): geometric mean of web-frequencies for words	
		q(AvgWordFreq): avg. web-frequencies for words	
		q(MaxColloqQuot): maximum bi-gram collocation quotient	
		q(IsAdvanced): query contains advanced features (e.g. Boolean)	
		q(ContainsName): query contains a person name	
		q(ContainsLoc): query contains a location name	

Feature type	# of Features	Feature	1 = most predictive feature relationship to p(click)
Search session	5	$S(q$Frac) ratio queries / search actions	**4** $p(click)$ decreases as more search actions are queries
		$S($Num$q)$ number queries entered in session	**7.** $p(click)$ decreases as more queries are entered
		$S($Max$q$Words) number of words in longest query submitted	**8.** $p(click)$ increases as query length increases over the session
		$S($DurationSec): duration of session	
		$S($Min$q$Words): number of words in shortest query submitted	

Feature type	# of Features	Feature
User	11	$U($Avg$S$Sec): avg. elapsed time per session
		$U($AvgSecToC_r): avg.
		$U($qPerSecInS): avg. queries per second in session
		$U($qRepeatRate): fraction queries that are repeats
		$U($qPerDay): avg. queries per day
		$U($AvgC_rPos): avg. rank clicked results
		$U($Avg$q$WordLen): avg. query length
		$U($C$_r$Prob): ratio of click-through to queries
		$U($PrefEngine): engine queried most frequently
		$U($PrefEngFreq): fraction of queries on preferred engine
		$U($AvgFirstResult): avg. rank of starting result requested
Results click	4	$C_r($position): results rank of item clicked
		C_r (DwellSec): elapsed time on the page opened after click
		C_r (IsAd): click is on advertisement
		$q\mid C_r$ (engine): search engine used
Non-action features	3	PathPageLength: number of pages in click-path after click to open page where click is on link or back-button
		PathDwellSec: total elapsed time on the click-path
		AfterPathSec: total elapsed time after path ends to next search action
TOTAL	51	

APPENDIX B – Protocol and Instruments

B.1 Protocol

[**RESEARCHER:** Greet subject in outer office. Show to room and have subject select most comfortable chair and a familiar type of mouse. After seating, give printed copy of experimental materials to subject and ask subject to follow along as you read text aloud.]

Thank you for your assistance in our research today.

The activity you are about to undertake should take no longer than 1 hour and 30 minutes.

The objective of this study is to see how you use an experimental searching system to look for good information.

There are no right or wrong answers to the questions you will be asked. We want you to tell us your honest reactions and opinions.

Many of the questions you may have about this study will be answered in this package of information, however, you may ask the researcher any questions at any time.

As a thank you for your efforts today, we will give you a $15 Knight Express Card.

You can stop at any time for any reason.

If you would like to stop the study, please tell the researcher.

[**RESEARCHER:** Explain to subject that they need to read the consent form and sign 2 copies if they agree.]

Informed Consent Form

Thank you for volunteering for this study of how people use information systems. The research will lead to better designs for online systems for finding information. Better designs will make it easier to find information quickly.

During the study, you will be asked to find information using an experimental computer system. When you use the system, a record will be made of the screens you see, the keys you press, and clicks you make with the mouse. It will take about 1 hour and 30 minutes to complete the experiment. Approximately 36 people will participate in the study. The risk from participating is no greater than normal everyday activity. You are volunteering to participate. You may change your mind and stop working on the experiment at any time. You don't need to explain if you decide to stop. There is no penalty for stopping.

The information you provide for the study will remain confidential. Only combined statistics will be reported in any published reports. Only trained researchers will work with the recorded information. No one will be able to identify you or your work with your name. You will have a chance to ask for a copy of any reports written from the data gathered today.

If you want a copy of this form, one will be given to you. If you have any questions about your rights as a research participant, you may contact the Sponsored Programs Administrator at Rutgers University at (732) 932-0150 ext. 2104 or at
Rutgers University Institutional Review Board for the Protection of Human Subjects
Office of Research and Sponsored Programs
3 Rutgers Plaza
New Brunswick, NJ 08901-8559
Email: humansubjects@orsp.rutgers.edu

In case you have any questions related to the research project, the principal investigator may be reached at (978) 337-6425 or by email at csmith@scils.rutgers.edu. The investigator's full address is:

Catherine L. Smith, Doctoral Student
Department of Library and Information Science
School of Communication, Information and Library Studies
Rutgers, The State University of New Jersey
4 Huntington Street
New Brunswick, New Jersey 08901-1071

My signature below indicates that I have read the information above and have decided to participate. I realize that I may withdraw without prejudice at any time after signing this form.

Participant's signature _____ Date _____

Participant's name *(please print)* _____

Investigator's signature _____ Date _____

Investigator's name *(please print)* _____Catherine L. Smith_____

This informed consent form was approved by the Rutgers University Institutional Review Board for the Protection of Human Subjects on 4/24/06; approval of this form expires on 4/23/07.

[RESEACHER: take the forms, sign both copies, and continue]

There are five steps to the activity you will complete today.

1. Complete questionnaire (12 questions)
2. Introduction
3. Practice
4. Searching (12 topics)
5. Final Step (brief set of questions)

Please:
- do not take any of the pages out of the package
- read the pages in this package in order in the package
- do not turn to the next page until instructed to do so in the package
- do not turn back to a prior page once you have gone forward

Your task today will be to look for information using a computer system.

Before continuing to the introduction, we would like you to complete the questionnaire.

1. COMPLETE QUESTIONNAIRE
.... [INSERT PRE-EXPERIMENT QUESTIONNAIRE] ....

During the introduction, you will learn more about the task you will be doing today. Before you start the task, you will see the system and examples of the type of topics you will be searching for. The researcher will explain the system to you. You will then get a chance to practice with the system, and you will have a chance to ask questions. Once you are done practicing you will start the task. Once you have completed the task you will be asked to answer a small set of questions and then you will be done.

2. INTRODUCTION

Today, while you are working, please pretend you are a trainee at a newspaper. Here is some information about your role as a trainee.

.... INSERT JOB DESCRIPTION ....

A copy of this "job description" is available on a card the researcher has for you. If you like, you may refer to it during your work.

Just to be clear, the "bonus" is a chance in a drawing for a $40.00 Knight Express card. The five volunteers for this study who find the most good information sources and the fewest bad information sources, as judged by the researcher, will be entered in the

drawing for the card. Also, while there really is no time limit on the searching, we really do hope that you will be able to finish searching for all twelve topics in an hour or less.

As you work, you will be asked to do the following steps:

- Before you begin work on a topic, you will be asked complete a short questionnaire, which is printed in this package.

- You will use Google to enter search words, browse through the list of sites, and indicate the sites you think have good information. If you want to visit any of the sites, you can. You can enter as many searches as you want.

- Once you have stopped working on a topic, you will be asked to complete another short questionnaire, also printed in this package.

PLEASE NOTE: There are no questionnaires for the example and practice topics.

As with any situation in which you are looking for information, some may be easy to find, and some may be difficult to find.

You can stop searching at any time to go on to the next topic. We will now look at an example topic.

Please click the button on the screen that says **START EXAMPLE** -- .

To see the first topic, you will click the button that says **FIRST TOPIC** -- .

Please click the **FIRST TOPIC** button now.

You will see a screen displaying the topic, which is presented as a statement. The screen also asks you to complete the questions you will find in the package. You will always answer these same questions before you start searching for a topic. When you have finished answering the questions, you will click **START TOPIC SEARCH** -- .

Please click **START TOPIC SEARCH** now.

You will see a screen with two parts. In the top part of the screen, the topic you are searching for is displayed on the screen. Here you can see the example topic *"Cats and kittens make good pets"*. Your task is to search for good information about this topic.

The bottom part of the screen contains a special version of the Google system. Please notice that you can only use Google's simple search. All the other functions are shut off. As you normally would with Google, enter the words you want to search on in the box, and click **SEARCH** --.

Please enter "cats and kittens" in the search box and click $\boxed{\text{SEARCH}}$ now.

As you would expect with Google, the results of the search will appear in the lower part of the screen. As with any Google search, you can repeat your search as many times as you want, changing the words you enter in the search box. You can also click the links to the web pages in order to explore the sites.

Please click a link to one of the web pages in the list.

Notice that the website you open appears in the screen to the right. You can use the mouse to investigate the site if you wish. Simply move the mouse pointer all the way to the right and it will appear in the right-hand screen. You may now explore the site by clicking links.

Looking back at the list found by Google, you can see that each website in the list has a checkbox next to it. If you think the item in the list is a good information source for the topic, click the box to add the source to the list sent to the journalist.

Please click some of the checkboxes.

When you enter a new search, if Google finds a site you have already checked as "good" the site will appear with a check in the checkbox, indicating that you have already selected that site. When you are done finding information sources for the topic, click $\boxed{\text{END}}$ -- .

Please click $\boxed{\text{END}}$ --

You will then be asked to confirm that you are done with the topic, or you can continue searching using the search box. For now, we will continue with the example.

Please click $\boxed{\text{CONFIRM}}$ --

You will be given one more chance to search again if you want to. All you would need to do is use the search box again. Once you have clicked the second confirmation, $\boxed{\text{I AM SURE}}$, you will not be able to go back to the topic.

Please click $\boxed{\text{I AM SURE}}$ now.

You will then see a screen asking you to complete the next set of questions in the package. As with the first set of questions, you will complete these questions every time you finish searching on a topic. When you have finished the questions, you will click $\boxed{\text{CONTINUE}}$ --.

Please click $\boxed{\text{CONTINUE}}$ now.

You will then be asked to click | CONTINUE TO NEXT TOPIC | when you are ready to start the next topic.

Please click | CONTINUE TO NEXT TOPIC | now.

You will then see the screen displaying the new topic statement, and asking you to complete the questions before beginning the search.

That is all there is to the system.

Some things to remember:
- Please don't use the back key. If you do accidentally use it, you will receive a message.
- If you get a message you don't understand, or if you have a problem with the system, please let me know.
- If you can't find any good information sources you don't need to click any checkboxes. Just like with any search, there may be some times when there is little or no good information on a topic.

Remember, you are looking for good information. A source of good information is a site *you* could and *would* use to get information about the topic. Also, you can stop looking for information at any time during the search by clicking | END | and going on to the next topic.

3. PRACTICE
You now have a chance to try the system before you start your work. While you practice, if you have questions, just ask. One important point: There are no question sets for the practice topics. When you see the request to complete the questions you can continue. After trying the system with the first practice topic, you can practice on up to five more topics if you want to, or you can start your task immediately. When you are ready to start searching for the practice topic, click | START TOPIC SEARCH | --.

When you have confirmed that you are ready to start the next topic, you will see | DONE WITH PRACTICE | on the screen. When you are done practicing, click that button to start your task of searching for the twelve topics.

To begin the practice topic, please click | START TOPIC SEARCH | now.

When you are done practicing, please click | DONE WITH PRACTICE | --.

After clicking | DONE WITH PRACTICE |, you may turn to the next page.

When you are ready to start working on the twelve topics, please click | FIRST TOPIC | --

4. SEARCHING

....
[SUBJECT USES SYSTEM AND COMPLETES PRE- and POST-SEARCH QUESTIONNAIRES]
.....

[**RESEARCHER:** upon completion of the final topic, continue]

Congratulations!

You have completed the task!

5. FINAL STEP

We have just a few more questions

…………….. [INSERT POST-EXPERIMENT QUESTIONNAIRE] ……………..

Thank you very much for your assistance in this research. Your work is very valuable in the study of how people use information systems. The researcher has your $15 Knight Express card to give you before you go.

Now that you are done with the task, we want you to know that some of your search tasks may have been more or less difficult than others. This was intended to help us understand some of the factors that influence how people look for information.

In order to give everyone the same chance at the $40 Knight Express card, in determining which five volunteers provided the most net good information, we will not count the parts of the task that were intentionally made more or less difficult.

If you would like to know more about this research, or would like to receive a copy of any published reports that use the information gathered today, please let the researcher know.

You may take the last page of this document with you. It contains contact information and a copy of the informed consent form.

If at any time now or in the future, you should have any questions or concerns about any aspect of the activities you just completed, please contact us. We are happy to answer any questions you may have.

Thank you again for your time and efforts today.

B.2 Subject Task Assignment: Mock Job Description

TRAINEE JOB DESCRIPTION

In your job as a trainee, you support the journalists at the newspaper. Your responsibility is to find information about the journalists' article topics. Today you need to search for good sources of information about *twelve* different topics that the journalists are working on. You search by using Google.

The Google system you use looks slightly different from regular Google. As you search, the topic you are working on is displayed at the top of the screen. Just like with any Google search, you will see a list of websites, and you may visit those websites to see if they have good information about the topic. In order to tell the journalists about the good information sources, you simply check a box indicating that the site on the list is good. All the items you check as good will be automatically included in a list for the journalist working on the topic.

You won't be given any information about why a journalist is looking for information on a topic, or what about the topic is important. For this reason, any source with information that will inform the journalist on the topic can be considered a "good" source. You need to find as many "good" information sources as you can, but it is also important to avoid sending information sources that are not good.

At the newspaper, there is a bonus for finding only good information sources. The journalists judge whether the sources found by trainees are good. The five trainees who find the most good information sources and the fewest "bad" sources, are eligible to win a "bonus". The bonus is given to one trainee, who is selected by lottery.

There is no time limit on searching, but your boss expects that you will be able to finish searching for all twelve topics in an hour or less.

B.3.a Pre-experiment Questionnaire

The information below is being collected for statistical purposes only. After you complete today's activities, it will not be kept with any personally identifying information about you. It will never be reported except in aggregate. Please complete the questions to the best of your ability. If you do not know an answer, or do not want to provide an answer, please leave the question blank. There are no right or wrong answers to these questions.

1. What is your gender?
   (please mark one)

   [ ] MALE   [ ] FEMALE

---

2. What is your native language?
   (please mark one)

   [ ] English   [ ] Non-English

---

3. What is your academic background?
   (please mark all that apply)

   [ ] Library Science   [ ] Information Science
   [ ] Computer Science   [ ] Other

---

4. What is your highest level of education?
   (please mark one box)

   [ ] grade school   [ ] associates degree   [ ] some graduate school

   [ ] some high school   [ ] trade school   [ ] Masters Degree

   [ ] graduated high school [ ] on the job training   [ ] some doctoral school

   [ ] some college   [ ] Bachelors degree   [ ] Doctoral degree

---

5. For how many hours did you use a personal computer yesterday?
   (please mark one box)

   [ ] did not use one yesterday   [ ] one hour to five hours

   [ ] less than one hour   [ ] over five hours

---

6. Are you currently a registered student at Rutgers University?
   (please mark one box)

   [ ] YES   [ ] NO

Please continue on the next page.

7. What is your age? *(please mark one box)*

[ ] younger than 18   [ ] 35 or older and not yet 50

[ ] 18 or older and not yet 25   [ ] 50 or older and not yet 80

[ ] 25 or older and not yet 35   [ ] 80 or older

For questions 8 through 10, please mark the number closest to your agreement with the statement.

8. I usually find what I am looking for on the Internet or World Wide Web.
(please mark one number)

1	2	3	4	5	6
Strongly disagree					Strongly agree

9. I am interested in online searching.
(please mark one number)

1	2	3	4	5	6
Strongly disagree					Strongly agree

10. I enjoy trying new ways to use the Internet or World Wide Web.
(please mark one number)

1	2	3	4	5	6
Strongly disagree					Strongly agree

11. I am familiar with Google searching.
(please mark one number)

1	2	3	4	5	6
Strongly disagree					Strongly agree

12. Google can find anything I need.
(please mark one number)

1	2	3	4	5	6
Strongly disagree					Strongly agree

B.3.b Pre-search Questionnaire

Please complete the questions below before starting your search.

Topic 2 statement:
Fishermen find it difficult to earn a net profit.

1. I am familiar with this topic.

1	2	3	4	5	6

Strongly disagree   Strongly agree
Please mark the number closest to your agreement with the statement

2. I expect that Google will have a lot of good information about this topic.

1	2	3	4	5	6

Strongly disagree   Strongly agree

3. I am confident Google will work as well as I expect it to.

1	2	3	4	5	6

Strongly disagree   Strongly agree

4. I expect to find _____ sites with good information on this topic.
                          number
           Please complete the sentence with a number

5. I am confident I will find that many sites.

1	2	3	4	5	6

Strongly disagree   Strongly agree
Please mark the number closest to your agreement with the statement

6. I expect it will take about _____ minutes to find good information on this topic.
                                   number
           Please complete the sentence with a number

7. I am confident it will take that long to find good information.

1	2	3	4	5	6

Strongly disagree   Strongly agree
Please mark the number closest to your agreement with the statement

8. It will be easy to find good information on this topic.

1	2	3	4	5	6

Strongly disagree   Strongly agree

B.3.c Post-search Questionnaire

Thinking about the search you just completed, please complete the questions below.

Topic 2 statement:
Fishermen find it difficult to earn a net profit.

For the questions below, please mark the number closest to your agreement with the statement.

1. Overall, I think my estimates for what I would find and how long it would take were:

1	2	3	4	5	6

Very Optimistic   Very Pessimistic
Please mark the number closest to your opinion.

2. Google had a lot of good information about the topic.

1	2	3	4	5	6

Strongly disagree   Strongly agree
Please mark the number closest to your agreement with the statement

3. Google worked as well as I expected.

1	2	3	4	5	6

Strongly disagree   Strongly agree

4. I was able to find the amount of good information I expected to find on the topic.

1	2	3	4	5	6

Strongly disagree   Strongly agree

5. It was easy to find good information.

1	2	3	4	5	6

Strongly disagree   Strongly agree

6. The search took as long as I expected.

1	2	3	4	5	6

Strongly disagree   Strongly agree

7. I was as successful as I expected to be.

1	2	3	4	5	6

Strongly disagree   Strongly agree

B.3.d Post-experiment Questionnaire

Please define "good information source" in your own words:

Please continue on the next page.

A good information source may have certain characteristics.

Please mark the number indicating the importance to you of each characteristic listed below.

	Unimportant	Little Importance	Neither Important or Unimportant	Important	Highly Important
1. well written	1	2	3	4	5
2. objective sources	1	2	3	4	5
3. general information	1	2	3	4	5
4. factual	1	2	3	4	5
5. easy to find	1	2	3	4	5
6. good graphics	1	2	3	4	5
7. detailed information	1	2	3	4	5
8. expert authors	1	2	3	4	5
9. links to other sources	1	2	3	4	5
10. both overview and detail together	1	2	3	4	5
11. easy to read	1	2	3	4	5
12. opinions	1	2	3	4	5

B.4 Changes made to the protocol during data collection

After the 3rd subject session was completed, the following changes were made to the protocol:

1) A clock was placed next to the computer screen. After a subject completed the final practice topic, but before the first experimental topic was started, the researcher pointed out the clock saying: "Oh, there's a clock here so you can keep track of time".

2) After a subject completed the first topic questionnaire, but before he or she started the second topic, the researcher pointed out the timer and counter on the computer screen, saying: "Oh, I just want to make sure you see the timer and the counter here on your screen."

After the 24th experimental session was completed, the protocol was changed as follows:

After a subject had searched for 80 minutes, he or she was allowed to finish the current topic and post-experiment questionnaire. Before the subject started the next topic, the researcher said: "I want to tell you that you have been working for over 80 minutes now. You can stop if you want to. If you stop, it won't make any difference in your chance of winning the $40.00 card."

Of the 12 subject sessions run after this change, 3 searched for over 80 minutes and were reminded that they could quit without penalty. All 3 subjects quit before completing the next topic.

APPENDIX C - Subjects
C.1 – Subject demographics and tests of group independence (n=36)

		Control	BR	MR	All Groups	χ^2 test of indepen-dence
			number of subjects			
Native Language	English	8	8	8	24	*n.s.*
	Not English	4	4	4	12	
Gender	Male	3	3	2	8	*n.s.*
	Female	9	9	10	28	
Use of PC yesterday	<1 hour	1	0	2	3	*n.s.*
	1 to 5 hours	8	7	7	22	
	> 5 hours	3	5	3	11	
Age	18 through 24	6	9	10	25	*n.s.*
	25 through 34	4	2	0	6	
	35 or older	2	1	2	5	
Major	Other	6	8	8	22	*n.s.*
	IS or LS or both	5	3	4	12	
	CS	1	1	0	2	
Student Status	registered	10	12	11	33	*n.s.*
	not registered	2	0	1	3	
Education Level	no college	2	2	2	6	*n.s.*
	some college	5	7	8	20	
	college graduate	5	3	2	10	

C.2 – Correlation matrix for responses to 5 questions about search experience
(n=36 for all items; significant correlations are **bold**)

Pre-experiment Questionnaire *(6-point Likert scale)*		Question #				
		#8	#9	#10	#11	#12
I usually find what I am looking for on the Internet or World Wide Web. (#8)	Pearson Correlation	1	**.682**	**.532**	**.660**	**.365**
	Sig. (2-tailed)		**.000**	**.001**	**.000**	**.029**
I am interested in online searching (#9)	Pearson Correlation	**.682**	1	**.753**	**.663**	-.011
	Sig. (2-tailed)	**.000**		**.000**	**.000**	.950
I enjoy trying new ways to use the Internet or World Wide Web (#10)	Pearson Correlation	**.532**	**.753**	1	**.693**	-.011
	Sig. (2-tailed)	**.001**	**.000**		**.000**	.949
I am familiar with Google searching (#11)	Pearson Correlation	**.660**	**.663**	**.693**	1	.266
	Sig. (2-tailed)	**.000**	**.000**	**.000**		.117
Google can find anything I need (#12)	Pearson Correlation	**.365**	-.011	-.011	.266	1
	Sig. (2-tailed)	**.029**	.950	.949	.117	

APPENDIX D – Overlap of displayed items by system

Number of items	System standard	BR	MR	combined BR + MR	all 3 systems
A (tagged and displayed to > 1)	773	22	27	49	822
B (tagged and displayed to 1)	385	75	100	175	560
C (not tagged displayed to > 1)	10,390	449	475	924	11,314
D (not tagged displayed to 1)	14,130	4,645	4,248	8,893	23,023
Total (A+B+C+D)	25,678	5,191	4,850	10,041	35,719
A + B (total tagged)	1,158	97	127	224	1,382
A + C (total displayed to > 1)	11,163	471	502	973	12,136
C + D (total not tagged)	24,520	5,094	4,723	9,817	34,337
B + D (total displayed to 1)	14,515	4,720	4,348	9,068	23,583
% items tagged (A+B) / (A+B+C+D)	4.5%	1.9%	2.6%	2.2%	3.9%
OVERLAP (A+C) / (A+B+C+D)	**43.5%**	**9.1%**	**10.4%**	**9.7%**	**34.0%**

How to read this table.
There were 35,719 items displayed to subjects. Of these, 1,382 items were tagged by subjects, or 3.9% of those displayed.

Of the 35,719 items displayed, 12,136 were displayed to more than one subject. Across all systems, the average *overlap* of displayed items is 12,136 / 35,719 = 34%. For the combined experimental systems (BR + MR) only 9.7% of items were displayed to more than one subject. For the standard system, 43.5% of items were displayed to more than one subject.

APPENDIX E – Descriptive statistics by group and block
E.1.a.i – Raw Data - Average per Completed Search - Group by Block
Block 1 : Control Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrecision)	48	.10	.02	.00	.75
% good pool items displayed during search (GRecall)	48	.12	.01	.00	.32
System response					
average length of displayed lists (average list length)	48	19.4	.24	11.4	20
% queries returning 20 item list (fraction-full-lists)	48	.96	.01	.57	1.0
% queries returning 0 item list (fraction-empty-lists)	48	.02	.01	.00	.43
% queries returning 1 -19 item list (fraction-short-lists)	48	.02	.01	.00	.25
# of unique items returned per query submitted (unique items per query)	48	17.4	.43	10.4	20
% item displays that repeat previously dis. item (item display repetitions)	48	.10	.02	.00	.48
Items					
# of items displayed (AI)	48	83.2	11.9	20	448
# of good items displayed (GI)	48	4.7	.63	0	19
# of marginal items displayed (MI)	48	1.9	.27	0	8
# of flagged items (AFI)	48	3.9	.37	0	12
# of good flagged items (GFI)	48	2.3	.35	0	12
# of marginal flagged items (MFI)	48	.7	.09	0	2
# of bad flagged items (BFI)	48	.8	.21	0	7
# of flagged items invalid at judgment (XFI)	48	.1	.04	0	1
Item displays					
# of item displays (AIDs)	48	96.9	14.3	20	565
# of good item displays (GIDs)	48	5.8	.72	0	19
# of marginal item displays (MIDs)	48	2.4	.41	0	14
# of flagged item displays (AFIDs)	48	5.0	.43	0	13
# of good flagged item displays (GFIDs)	48	2.9	.46	0	16
# of marginal flagged item displays (MFIDs)	48	1.0	.16	0	5
Searcher productivity					
% flagged items that are good (good item ratio)	47	.51	.05	.00	1.0
% flagged items that are marginal (marginal item ratio)	47	.24	.04	.00	1.0
% flagged items that are bad (bad item ratio)	47	.23	.04	.00	1.0
% good item displays flagged (good item detection rate)	43	.56	.05	.00	1.1
% marginal items displays flagged (marg. item det. rate)	38	.52	.07	.00	1.0
Queries					
# of queries submitted (query count)	48	5.2	.79	1	30
elapsed topic time (minutes) (ETTime)	48	7.22	.57	2.8	20.8
queries submitted per minute (query rate)	48	.70	.07	.10	2.12
% of item displays flagged by searcher (flagging rate)	48	.10	.02	.00	.6
average # of terms per query (average query length)	48	4.1	.28	1.0	9.9
avg. # of spelling mess. per query (spelling message per query)	48	.04	.01	.00	.38
% query submissions that repeat previously submitted query (query repetitions)	48	.011	.008	.00	.33

E.1.a.ii – Raw Data - Average per Completed Search - Group by Block
Block 2 : Control Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	48	.08	.02	.00	.75
% good pool items displayed during search (GRec)	48	.10	.01	.00	.39
System response					
average length of displayed lists (average list length)	48	19.3	.19	15.7	20
% queries returning 20 item list (fraction-full-lists)	48	.96	.01	.75	1.0
% queries returning 0 item list (fraction-empty-lists)	48	.02	.01	.00	.20
% queries returning 1 -19 item list (fraction-short-lists)	48	.02	.01	.00	.25
# of unique items returned per query submitted (unique items per query)	48	17.5	.36	10	20
% item displays that repeat previously dis. item (item display repetitions)	48	.09	.02	.00	.50
Items					
# of items displayed (AI)	48	77.5	7.1	20	244
# of good items displayed (GI)	48	3.9	.62	0	18
# of marginal items displayed (MI)	48	1.6	.26	0	7
# of flagged items (AFI)	48	3.0	.32	0	9
# of good flagged items (GFI)	48	1.4	.29	0	8
# of marginal flagged items (MFI)	48	.5	.11	0	3
# of bad flagged items (BFI)	48	.8	.13	0	3
# of flagged items invalid at judgment (XFI)	48	.3	.08	0	3
Item displays					
# of item displays (AIDs)	48	87.8	8.5	20	283
# of good item displays (GIDs)	48	4.9	.74	0	22
# of marginal item displays (MIDs)	48	2.1	.38	0	14
# of flagged item displays (AFIDs)	48	3.9	.43	0	13
# of good flagged item displays (GFIDs)	48	1.9	.39	0	12
# of marginal flagged item displays (MFIDs)	48	.6	.13	0	3
Searcher productivity					
% flagged items that are good (good item ratio)	44	.41	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	44	.17	.04	.00	1.0
% flagged items that are bad (bad item ratio)	44	.33	.06	.00	1.0
% good item displays flagged (good item detection rate)	36	.40	.06	.00	1.0
% marginal items displays flagged (marg. item det. rate)	33	.32	.07	.00	1.0
Queries					
# of queries submitted (query count)	48	4.7	.50	1	18
elapsed topic time (minutes) (ETTime)	48	5.1	.29	1.9	10
queries submitted per minute (query rate)	48	.91	.07	.10	2.7
% of item displays flagged by searcher (flagging rate)	48	.06	.01	.00	.3
average # of terms per query (average query length)	48	4.2	.25	2.0	9.3
avg. # of spelling mess. per query (spelling message per query)	48	.06	.01	.00	.36
% query submissions that repeat previously submitted query (query repetitions)	48	.007	.005	.00	.20

E.1.a.iii – Raw Data - Average per Completed Search - Group by Block
Block 3 : Control Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	41	.07	.01	.00	.45
% good pool items displayed during search (GRec)	41	.10	.01	.00	.32
System response					
average length of displayed lists (average list length)	41	19	.37	10	20
% queries returning 20 item list (fraction-full-lists)	41	.94	.02	.50	1.0
% queries returning 0 item list (fraction-empty-lists)	41	.04	.02	.00	.5
% queries returning 1 -19 item list (fraction-short-lists)	41	.02	.01	.00	.5
# of unique items returned per query submitted (unique items per query)	41	17	.45	10	20
% item displays that repeat previously dis. item (item display repetitions)	41	.10	.02	.00	.36
Items					
# of items displayed (AI)	41	76.8	8.8	20	238
# of good items displayed (GI)	41	3.6	.47	0	10
# of marginal items displayed (MI)	41	1.7	.25	0	7
# of flagged items (AFI)	41	3.1	.37	0	9
# of good flagged items (GFI)	41	1.3	.26	0	8
# of marginal flagged items (MFI)	41	.7	.15	0	4
# of bad flagged items (BFI)	41	.8	.16	0	4
# of flagged items invalid at judgment (XFI)	41	.2	.06	0	1
Item displays					
# of item displays (AIDs)	41	90.2	11.2	20	300
# of good item displays (GIDs)	41	5.3	.98	0	35
# of marginal item displays (MIDs)	41	2.4	.41	0	11
# of flagged item displays (AFIDs)	41	4.3	.73	0	27
# of good flagged item displays (GFIDs)	41	2.2	.76	0	30
# of marginal flagged item displays (MFIDs)	41	1.0	.21	0	6
Searcher productivity					
% flagged items that are good (good item ratio)	36	.46	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	36	.24	.04	.00	1.0
% flagged items that are bad (bad item ratio)	36	.22	.04	.00	.75
% good item displays flagged (good item detection rate)	33	.47	.06	.00	1.0
% marginal items displays flagged (marg. item det. rate)	29	.42	.07	.00	1.0
Queries					
# of queries submitted (query count)	41	4.7	.57	1	15
elapsed topic time (minutes) (ETTime)	41	5.2	.53	1.7	17.4
queries submitted per minute (query rate)	41	1.0	.10	.10	2.8
% of item displays flagged by searcher (flagging rate)	41	.07	.01	.00	.25
average # of terms per query (average query length)	41	4.6	.23	2.0	8.4
avg. # of spelling mess. per query (spelling message per query)	41	.10	.02	.00	.50
% query submissions that repeat previously submitted query (query repetitions)	48	.024	.011	.00	.33

E.1.b.i – Raw Data - Average per Completed Search - Group by Block
Block 1 : Bottom Rankings Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	48	.08	.01	.00	.60
% good pool items displayed during search (GRec)	48	.12	.01	.00	.42
System response					
average length of displayed lists (average list length)	48	18.4	.49	7.2	20
% queries returning 20 item list (fraction-full-lists)	48	.90	.03	.27	1.0
% queries returning 0 item list (fraction-empty-lists)	48	.07	.02	.00	.6
% queries returning 1 -19 item list (fraction-short-lists)	48	.03	.01	.00	.6
# of unique items returned per query submitted (unique items per query)	48	15.7	.53	5.2	20
% item displays that repeat previously dis. item (item display repetitions)	48	.15	.02	.00	.55
Items					
# of items displayed (AI)	48	94.5	7.8	20	222
# of good items displayed (GI)	48	5.2	.77	0	25
# of marginal items displayed (MI)	48	2.2	.3	0	9
# of flagged items (AFI)	48	4.2	.41	0	12
# of good flagged items (GFI)	48	2.3	.35	0	12
# of marginal flagged items (MFI)	48	.9	.16	0	4
# of bad flagged items (BFI)	48	.9	.27	0	11
# of flagged items invalid at judgment (XFI)	48	.1	.05	0	1
Item displays					
# of item displays (AIDs)	48	117.3	10.5	20	280
# of good item displays (GIDs)	48	7.9	1.5	0	46
# of marginal item displays (MIDs)	48	3.0	.46	0	15
# of flagged item displays (AFIDs)	48	6.6	1.0	0	44
# of good flagged item displays (GFIDs)	48	3.3	.58	0	20
# of marginal flagged item displays (MFIDs)	48	1.2	.24	0	8
Searcher productivity					
% flagged items that are good (good item ratio)	46	.52	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	46	.25	.05	.00	1.0
% flagged items that are bad (bad item ratio)	46	.20	.05	.00	1.0
% item displays that are flagged (flagging rate)	48	.07	.01	.00	.35
% good item displays flagged (good item detection rate)	41	.48	.05	.00	1.0
% marginal items displays flagged (marg. item det. rate)	38	.33	.05	.00	1.0
Queries					
# of queries submitted (query count)	48	7.0	.76	1	26
elapsed topic time (minutes) (ETTime)	48	6.9	.37	2.2	12.3
queries submitted per minute (query rate)	48	1.0	.10	.16	3.4
% of item displays flagged by searcher (flagging rate)	48	.07	.01	.00	.4
average # of terms per query (average query length)	48	4.6	.26	2.0	11.0
avg. # of spelling mess. per query (spelling message per query)	48	.06	.02	.00	.57
% query submissions that repeat previously submitted query (query repetitions)	48	.021	.007	.00	.17

E.1.b.ii – Raw Data - Average per Completed Search - Group by Block
Block 2 : Bottom Rankings Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	48	.01	.003	.00	.10
% good pool items displayed during search (GRec)	48	.02	.004	.00	.14
System response					
average length of displayed lists (average list length)	48	17.1	.63	4.5	20
% queries returning 20 item list (fraction-full-lists)	48	.80	.04	.08	1.0
% queries returning 0 item list (fraction-empty-lists)	48	.09	.02	.00	.7
% queries returning 1 -19 item list (fraction-short-lists)	48	.11	.03	.00	.8
# of unique items returned per query submitted (unique items per query)	48	16.5	.68	3.2	20
% item displays that repeat previously dis. item (item display repetitions)	48	.05	.01	.00	.31
Items					
# of items displayed (AI)	48	108.2	10.2	20	399
# of good items displayed (GI)	48	1.0	.21	0	5
# of marginal items displayed (MI)	48	.5	.10	0	2
# of flagged items (AFI)	48	2.0	.27	0	7
# of good flagged items (GFI)	48	.8	.18	0	5
# of marginal flagged items (MFI)	48	.3	.08	0	2
# of bad flagged items (BFI)	48	.7	.14	0	4
# of flagged items invalid at judgment (XFI)	48	.2	.05	0	1
Item displays					
# of item displays (AIDs)	48	115.5	11.0	20	419
# of good item displays (GIDs)	48	1.2	.24	0	7
# of marginal item displays (MIDs)	48	.5	.10	0	2
# of flagged item displays (AFIDs)	48	2.2	.30	0	7
# of good flagged item displays (GFIDs)	48	.9	.21	0	7
# of marginal flagged item displays (MFIDs)	48	.4	.09	0	2
Searcher productivity					
% flagged items that are good (good item ratio)	36	.32	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	36	.22	.06	.00	1.0
% flagged items that are bad (bad item ratio)	36	.36	.06	.00	1.0
% good item displays flagged (good item detection rate)	23	.76	.08	.00	1.0
% marginal items displays flagged (marg. item det. rate)	19	.71	.10	.00	1.0
Queries					
# of queries submitted (query count)	48	7.5	.84	1	29
elapsed topic time (minutes) (ETTime)	48	4.7	.24	1.6	9.3
queries submitted per minute (query rate)	48	1.5	.12	.27	4.4
% of item displays flagged by searcher (flagging rate)	48	.03	.01	.00	.3
average # of terms per query (average query length)	48	4.2	.20	2.1	9.0
avg. # of spelling mess. per query (spelling message per query)	48	.01	.01	.00	.20
% query submissions that repeat previously submitted query (query repetitions)	48	.028	.009	.00	.33

E.1.b.iii – Raw Data - Average per Completed Search - Group by Block
Block 3 : Bottom Rankings Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	42	.09	.02	.00	.55
% good pool items displayed during search (GRec)	42	.09	.01	.00	.46
System response					
average length of displayed lists (average list length)	42	18.3	.57	6.3	20
% queries returning 20 item list (fraction-full-lists)	42	.9	.03	.21	1.0
% queries returning 0 item list (fraction-empty-lists)	42	.06	.02	.00	.5
% queries returning 1 -19 item list (fraction-short-lists)	42	.04	.01	.00	.3
# of unique items returned per query submitted (unique items per query)	42	16.2	.72	5.0	20
% item displays that repeat previously dis. item (item display repetitions)	42	.12	.02	.00	.51
Items					
# of items displayed (AI)	42	60.4	6.6	20	221
# of good items displayed (GI)	42	3.6	.57	0	17
# of marginal items displayed (MI)	42	1.6	.29	0	8
# of flagged items (AFI)	42	3.8	.48	0	11
# of good flagged items (GFI)	42	1.8	.34	0	8
# of marginal flagged items (MFI)	42	.9	.21	0	6
# of bad flagged items (BFI)	42	1.0	.28	0	8
# of flagged items invalid at judgment (XFI)	42	.2	.06	0	1
Item displays					
# of item displays (AIDs)	42	73.0	8.6	20	280
# of good item displays (GIDs)	42	4.5	.85	0	29
# of marginal item displays (MIDs)	42	2.2	.55	0	21
# of flagged item displays (AFIDs)	42	4.8	.66	0	18
# of good flagged item displays (GFIDs)	42	2.2	.49	0	16
# of marginal flagged item displays (MFIDs)	42	1.1	.31	0	10
Searcher productivity					
% flagged items that are good (good item ratio)	39	.42	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	39	.25	.05	.00	1.0
% flagged items that are bad (bad item ratio)	39	.25	.06	.00	1.0
% good item displays flagged (good item detection rate)	36	.50	.06	.00	1,0
% marginal items displays flagged (marg. item det. rate)	28	.49	.08	.00	1.0
Queries					
# of queries submitted (query count)	42	4.6	.69	1	19
elapsed topic time (minutes) (ETTime)	42	4.0	.41	1.4	15.6
queries submitted per minute (query rate)	42	1.2	.12	.17	3.5
% of item displays flagged by searcher (flagging rate)	42	.10	.02	.00	.5
average # of terms per query (average query length)	42	4.1	.23	1.4	9.3
avg. # of spelling mess. per query (spelling message per query)	42	.10	.04	.00	1.0
% query submissions that repeat previously submitted query (query repetitions)	48	.017	.007	.00	.20

E.1.c.i – Raw Data - Average per Completed Search - Group by Block
Block 1: Mixed Rankings Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	48	.09	.02	.00	.52
% good pool items displayed during search (GRec)	48	.15	.02	.00	.42
System response					
average length of displayed lists (average list length)	48	19.5	.22	13.6	20
% queries returning 20 item list (fraction-full-lists)	48	.96	.01	.56	1.0
% queries returning 0 item list (fraction-empty-lists)	48	.02	.01	.00	.3
% queries returning 1 -19 item list (fraction-short-lists)	48	.02	.01	.00	.3
# of unique items returned per query submitted (unique items per query)	48	16.8	.36	11.7	20
% item displays that repeat previously dis. item (item display repetitions)	48	.14	.02	.00	.41
Items					
# of items displayed (AI)	48	99.1	7.4	39	250
# of good items displayed (GI)	48	6.1	.73	0	19
# of marginal items displayed (MI)	48	1.9	.27	0	8
# of flagged items (AFI)	48	3.9	.38	0	10
# of good flagged items (GFI)	48	2.3	.37	0	9
# of marginal flagged items (MFI)	48	.7	.14	0	4
# of bad flagged items (BFI)	48	.8	.18	0	6
# of flagged items invalid at judgment (XFI)	48	.2	.05	0	1
Item displays					
# of item displays (AIDs)	48	120.8	11.3	40	420
# of good item displays (GIDs)	48	9.0	1.3	0	39
# of marginal item displays (MIDs)	48	3.0	.58	0	23
# of flagged item displays (AFIDs)	48	5.9	.70	0	24
# of good flagged item displays (GFIDs)	48	3.5	.58	0	18
# of marginal flagged item displays (MFIDs)	48	1.1	.33	0	14
Searcher productivity					
% flagged items that are good (good item ratio)	46	.56	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	46	.18	.04	.00	1.0
% flagged items that are bad (bad item ratio)	46	.21	.04	.00	1.0
% good item displays flagged (good item detection rate)	41	.41	.05	.00	1.0
% marginal items displays flagged (marg. item det. rate)	35	.37	.07	.00	1.0
Queries					
# of queries submitted (query count)	48	6.23	.57	2	21
elapsed topic time (minutes) (ETTime)	48	7.6	.41	2.7	16
queries submitted per minute (query rate)	48	.85	.06	.26	1.8
% of item displays flagged by searcher (flagging rate)	48	.07	.01	.00	.3
average # of terms per query (average query length)	48	4.0	.21	1.5	9.0
avg. # of spelling mess. per query (spelling message per query)	48	.06	.02	.00	.50
% query submissions that repeat previously submitted query (query repetitions)	48	.016	.007	.00	.19

E.1.c.ii – Raw Data - Average per Completed Search - Group by Block
Block 2 : Mixed Rankings Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	48	.02	.01	.00	.13
% good pool items displayed during search (GRec)	48	.04	.01	.00	.19
System response					
average length of displayed lists (average list length)	48	19.0	.26	12.7	20
% queries returning 20 item list (fraction-full-lists)	48	.9	.02	.40	1.0
% queries returning 0 item list (fraction-empty-lists)	48	.01	.01	.00	.3
% queries returning 1 -19 item list (fraction-short-lists)	48	.08	.02	.00	.6
# of unique items returned per query submitted (unique items per query)	48	18.7	.32	9.4	20
% item displays that repeat previously dis. item (item display repetitions)	48	.02	.01	.00	.26
Items					
# of items displayed (AI)	48	101	8.4	20	290
# of good items displayed (GI)	48	1.7	.32	0	11
# of marginal items displayed (MI)	48	.8	.14	0	4
# of flagged items (AFI)	48	2.7	.36	0	13
# of good flagged items (GFI)	48	1.2	.23	0	7
# of marginal flagged items (MFI)	48	.5	.11	0	4
# of bad flagged items (BFI)	48	.9	.17	0	5
# of flagged items invalid at judgment (XFI)	48	.1	.05	0	1
Item displays					
# of item displays (AIDs)	48	103.6	8.9	20	322
# of good item displays (GIDs)	48	1.7	.34	0	11
# of marginal item displays (MIDs)	48	.8	.14	0	4
# of flagged item displays (AFIDs)	48	2.7	.36	0	13
# of good flagged item displays (GFIDs)	48	1.2	.23	0	6
# of marginal flagged item displays (MFIDs)	48	.5	.11	0	4
Searcher productivity					
% flagged items that are good (good item ratio)	41	.44	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	41	.21	.05	.00	1.0
% flagged items that are bad (bad item ratio)	41	.32	.05	.00	1.0
% good item displays flagged (good item detection rate)	28	.77	.06	.00	1.0
% marginal items displays flagged (marg. item det. rate)	24	.65	.09	.00	1.0
Queries					
# of queries submitted (query count)	48	5.5	.49	1	18
elapsed topic time (minutes) (ETTime)	48	4.9	.32	1.5	11.6
queries submitted per minute (query rate)	48	1.2	.07	.35	2.4
% of item displays flagged by searcher (flagging rate)	48	.04	.01	.00	.2
average # of terms per query (average query length)	48	4.2	.20	1.7	7.5
avg. # of spelling mess. per query (spelling message per query)	48	.02	.01	.00	.25
% query submissions that repeat previously submitted query (query repetitions)	48	.014	.007	.00	.29

E.1.c.iii – Raw Data - Average per Completed Search - Group by Block
Block 3 : Mixed Rankings Group

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good (GPrec)	45	.11	.01	.00	.35
% good pool items displayed during search (GRec)	45	.13	.02	.00	.71
System response					
average length of displayed lists (average list length)	45	19.4	.23	13	20
% queries returning 20 item list (fraction-full-lists)	45	.96	.02	.57	1.0
% queries returning 0 item list (fraction-empty-lists)	45	.02	.01	.00	.3
% queries returning 1 -19 item list (fraction-short-lists)	45	.02	.01	.00	.3
# of unique items returned per query submitted (unique items per query)	45	17.1	.46	8.7	20
% item displays that repeat previously dis. item (item display repetitions)	45	.12	.02	.00	.55
Items					
# of items displayed (AI)	45	62.7	5.1	20	152
# of good items displayed (GI)	45	5.0	.68	0	20
# of marginal items displayed (MI)	45	1.6	.22	0	5
# of flagged items (AFI)	45	3.5	.36	0	13
# of good flagged items (GFI)	45	2.1	.30	0	8
# of marginal flagged items (MFI)	45	.8	.15	0	4
# of bad flagged items (BFI)	45	.6	.14	0	4
# of flagged items invalid at judgment (XFI)	45	.1	.04	0	1
Item displays					
# of item displays (AIDs)	45	74.0	6.2	20	180
# of good item displays (GIDs)	45	7.0	1.1	0	37
# of marginal item displays (MIDs)	45	2.2	.32	0	7
# of flagged item displays (AFIDs)	45	4.8	.58	0	19
# of good flagged item displays (GFIDs)	45	2.9	.47	0	16
# of marginal flagged item displays (MFIDs)	45	1.0	.19	0	5
Searcher productivity					
% flagged items that are good (good item ratio)	44	.53	.06	.00	1.0
% flagged items that are marginal (marginal item ratio)	44	.26	.05	.00	1.0
% flagged items that are bad (bad item ratio)	44	.19	.05	.00	1.0
% good item displays flagged (good item detection rate)	40	.50	.05	.00	1.0
% marginal items displays flagged (marg. item det. rate)	32	.47	.07	.00	1.0
Queries					
# of queries submitted (query count)	45	3.9	.34	1	10
elapsed topic time (minutes) (ETTime)	45	4.3	.51	.93	21.5
queries submitted per minute (query rate)	45	1.0	.08	.29	2.3
% of item displays flagged by searcher (flagging rate)	45	.08	.01	.00	.3
average # of terms per query (average query length)	45	4.7	.27	2.0	10.0
avg. # of spelling mess. per query (spelling message per query)	45	.05	.02	.00	.50
% query submissions that repeat previously submitted query (query repetitions)	48	.006	.004	.00	.14

E.2.a.i – Raw Data - Average per Completed Search - Group by Block
Block 1 : Control Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	48	.14	.02	.01	.75
% good or marginal pool items dis. during search (GpM_Recall)	48	.19	.02	.03	.71
Items					
# of good + marginal items displayed (GpMI)	48	6.5	.67	1	20
# of good + marginal flagged items (GpMFI)	48	3.0	.36	0	12
Item displays					
# of good + marginal item displays (GpMIDs)	48	8.2	.83	1	20
# of good + marginal flagged item displays (GpMFIDs)	48	3.9	.50	0	19
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	47	.75	.05	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	48	.54	.05	.00	1.1

E.2.a.ii – Raw Data - Average per Completed Search - Group by Block
Block 2 : Control Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	48	.10	.02	.00	.75
% good or marginal pool items dis. during search (GpM_Recall)	48	.16	.02	.00	.79
Items					
# of good + marginal items displayed (GpMI)	48	5.5	.65	0	18
# of good + marginal flagged items (GpMFI)	48	1.9	.29	0	8
Item displays					
# of good + marginal item displays (GpMIDs)	48	6.9	.82	0	22
# of good + marginal flagged item displays (GpMFIDs)	48	2.5	.39	0	12
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	44	.58	.06	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	42	.38	.05	.00	1.0

E.2.a.iii – Raw Data - Average per Completed Search - Group by Block: Control Group
Block 3 : Control Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	41	.11	.01	.00	.45
% good or marginal pool items dis. during search (GpM_Recall)	41	.15	.02	.00	.39
Items					
# of good + marginal items displayed (GpMI)	41	5.2	.54	0	12
# of good + marginal flagged items (GpMFI)	41	2.1	.28	0	8
Item displays					
# of good + marginal item displays (GpMIDs)	41	7.7	1.1	0	36
# of good + marginal flagged item displays (GpMFIDs)	41	3.2	.75	0	30
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	36	.70	.04	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	38	.48	.05	.00	1.0

E.2.b.i – Raw Data - Average per Completed Search - Group by Block
Block 1: Bottom Rankings Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	48	.12	.02	.00	.65
% good or marginal pool items dis. during search (GpM_Recall)	48	.21	.02	.00	.71
Items					
# of good + marginal items displayed (GpMI)	48	7.4	.76	0	25
# of good + marginal flagged items (GpMFI)	48	3.1	.36	0	12
Item displays					
# of good + marginal item displays (GpMIDs)	48	10.9	1.5	0	46
# of good + marginal flagged item displays (GpMFIDs)	48	4.5	.59	0	20
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	46	.77	.05	.00	1
% good or marginal item dis. flagged (GpM item det. rate)	46	.44	.04	.00	1.0

E.2.b.ii – Raw Data - Average per Completed Search - Group by Block
Block 2 : Bottom Rankings Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	48	.02	.004	.00	.15
% good or marginal pool items dis. during search (GpM_Recall)	48	.04	.005	.00	.14
Items					
# of good + marginal items displayed (GpMI)	48	1.5	.23	0	6
# of good + marginal flagged items (GpMFI)	48	1.2	.21	0	6
Item displays					
# of good + marginal item displays (GpMIDs)	48	1.7	.26	0	7
# of good + marginal flagged item displays (GpMFIDs)	48	1.3	.24	0	7
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	36	.54	.06	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	35	.73	.07	.00	1.0

E.2.b.iii – Raw Data - Average per Completed Search - Group by Block
Block 3: Bottom Rankings Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	42	.13	.02	.00	.55
% good or marginal pool items dis. during search (GpM_Recall)	42	.15	.02	.00	.79
Items					
# of good + marginal items displayed (GpMI)	42	5.1	.59	0	17
# of good + marginal flagged items (GpMFI)	42	2.7	.39	0	9
Item displays					
# of good + marginal item displays (GpMIDs)	42	6.7	1.0	0	35
# of good + marginal flagged item displays (GpMFIDs)	42	3.3	.58	0	18
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	39	.68	.06	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	39	.53	.05	.00	1.0

E.2.c.i – Raw Data - Average per Completed Search - Group by Block
Block 1: Mixed Rankings Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	48	.12	.02	.00	.52
% good or marginal pool items dis. during search (GpM_Recall)	48	.22	.02	.00	.71
Items					
# of good + marginal items displayed (GpMI)	48	8.0	.72	0	20
# of good + marginal flagged items (GpMFI)	48	3.0	.36	0	10
Item displays					
# of good + marginal item displays (GpMIDs)	48	12.0	1.4	0	48
# of good + marginal flagged item displays (GpMFIDs)	48	4.5	.58	0	18
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	46	.74	.05	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	44	.44	.03	.00	1.0

E.2.c.ii – Raw Data - Average per Completed Search - Group by Block
Block 2 : Mixed Rankings Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	48	.03	.01	.00	.17
% good or marginal pool items dis. during search (GpM_Recall)	48	.07	.01	.00	.29
Items					
# of good + marginal items displayed (GpMI)	48	2.4	.37	0	12
# of good + marginal flagged items (GpMFI)	48	1.6	.28	0	11
Item displays					
# of good + marginal item displays (GpMIDs)	48	2.5	.39	0	12
# of good + marginal flagged item displays (GpMFIDs)	48	1.6	.27	0	10
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	41	.65	.06	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	39	.72	.05	.00	1.0

E.2.c.iii – Raw Data - Average per Completed Search - Group by Block
Block 3 : Mixed Rankings Group (good plus marginal)

	N	Mean	s.e.m.	min.	max.
System performance					
% item displays that are good or marginal (GpM _Precision)	45	.14	.01	.01	.35
% good or marginal pool items dis. during search (GpM_Recall)	45	.19	.02	.02	.82
Items					
# of good + marginal items displayed (GpMI)	45	6.6	.70	1	23
# of good + marginal flagged items (GpMFI)	45	2.8	.31	0	9
Item displays					
# of good + marginal item displays (GpMIDs)	45	9.2	1.2	1	44
# of good + marginal flagged item displays (GpMFIDs)	45	3.8	.48	0	16
Searcher productivity					
% flagged items that are good or marginal (GpM item ratio)	44	.79	.05	.00	1.0
% good or marginal item dis. flagged (GpM item det. rate)	45	.49	.04	.00	1.0

APPENDIX F – Extraction of incidental effects: effect sizes of extracted factors

	N	Partial-η^2		
		Topic	Subject	Pos.
System performance				
% item displays that are good (GPrec)	416	.43	.14	.13
% good pool items displayed during search (GRec)	416	.25	.24	.19
System response				
average length of displayed lists (average list length)	416	.07	.52	.05
% queries returning 20 item list (fraction-full-lists)	416	.08	.48	.05
% queries returning 0 item list (fraction-empty-lists)	416	.04	.48	.02
% queries returning 1 -19 item list (fraction-short-lists)	416	.05	.23	.06
# of unique items returned per query submitted (unique items per query)	416	.13	.44	.06
% item displays that repeat previously dis. item (item display repetitions)	416	.10	.19	.13
Items				
# of items displayed (AI)	416	.12	.49	.10
# of good items displayed (GI)	416	.46	.23	.20
# of marginal items displayed (MI)	416	.27	.16	.11
# of flagged items (AFI)	416	.29	.39	.13
# of good flagged items (GFI)	416	.46	.20	.11
# of marginal flagged items (MFI)	416	.13	.16	.06
# of bad flagged items (BFI)	416	.07	.28	.04
# of flagged items invalid at judgment (XFI)	416	.06	.07	.01
Item displays				
# of item displays (AIDs)	416	.13	.50	.09
# of good item displays (GIDs)	416	.29	.19	.17
# of marginal item displays (MIDs)	416	.26	.16	.13
# of flagged item displays (AFIDs)	416	.09	.31	.14
# of good flagged item displays (GFIDs)	416	.23	.16	.11
# of marginal flagged item displays (MFIDs)	416	.16	.14	.07
Searcher productivity				
% flagged items that are good (good item ratio)	379	.26	.17	.05
% flagged items that are marginal (marginal item ratio)	379	.21	.12	.02
% flagged items that are bad (bad item ratio)	379	.12	.23	.07
% item displays that are flagged (flagging rate)	416	.27	.30	.11
% good item displays flagged (good item detection rate)	321	.18	.20	.05
% marginal items displays flagged (marg. item det. rate)	276	.19	.17	.03
Queries				
# of queries submitted (query count)	416	.15	.59	.14
elapsed topic time (minutes) (ETTime)	416	.07	.46	.28
queries submitted per minute (query rate)	416	.13	.58	.10
average # of terms per query (average query length)	416	.32	.51	.03
avg. # of spelling mess. per query (spelling message per query)	67	.04	.14	.05
% query sub. repeat previously submitted (query repetitions)	416	.03	.15	.02

F.2 – Effect sizes of extracted factors (good plus marginal)

	N	Partial-η^2		
		Topic	Subject	Pos.
System performance				
% item displays that are good or marginal (GpMPrec)	416	.34	.18	.18
% good or marginal pool items dis. during search (GpM Rec)	416	.31	.24	.21
Items				
# of good + marginal items displayed (GpMI)	416	.31	.27	.25
# of good + marginal flagged items (GpMFI)	416	.36	.26	.15
Item displays				
# of good + marginal item displays (GpMIDs)	416	.19	.23	.21
# of good + marginal flagged item displays (GpMFIDs)	416	.16	.21	.15
Searcher productivity				
% flagged items that are good or marginal (GpM item ratio)	379	.13	.22	.09
% good or marginal item dis. flagged (GpM item detection rate)	376	.17	.20	.05

APPENDIX G - Parameters for models of list-length and IQTI

G.1 – Subject, topic, and position parameters for model of list-length

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial-η^2
					Lower Bound	Upper Bound	
subject = 1	-0.206	0.616	-0.334	0.739	-1.414	1.003	0.000
subject = 2	0.264	0.840	0.315	0.753	-1.383	1.912	0.000
subject = 3	0.237	0.969	0.245	0.807	-1.664	2.138	0.000
subject = 4	-0.923	1.060	-0.870	0.384	-3.002	1.156	0.000
subject = 5	-1.089	0.861	-1.265	0.206	-2.777	0.599	0.001
subject = 6	0.678	0.815	0.832	0.406	-0.920	2.275	0.000
subject = 7	-0.198	0.673	-0.294	0.769	-1.518	1.122	0.000
subject = 8	0.097	1.131	0.086	0.932	-2.120	2.314	0.000
subject = 9	-0.596	1.010	-0.590	0.555	-2.577	1.385	0.000
subject = 10	0.521	0.855	0.610	0.542	-1.155	2.197	0.000
subject = 11	1.050	1.002	1.049	0.294	-0.914	3.014	0.000
subject = 12	-0.041	0.793	-0.052	0.959	-1.596	1.515	0.000
subject = 13	0.964	0.918	1.050	0.294	-0.837	2.766	0.000
subject = 14	0.422	0.759	0.555	0.579	-1.067	1.911	0.000
subject = 15	0.983	0.848	1.159	0.247	-0.680	2.645	0.001
subject = 16	0.611	0.709	0.862	0.389	-0.778	2.000	0.000
subject = 17	-0.184	0.820	-0.224	0.823	-1.791	1.424	0.000
subject = 18	0.436	0.633	0.689	0.491	-0.805	1.678	0.000
subject = 19	0.041	0.730	0.056	0.955	-1.391	1.473	0.000
subject = 20	0.612	0.730	0.837	0.402	-0.821	2.044	0.000
subject = 21	0.807	0.803	1.005	0.315	-0.768	2.383	0.000
subject = 22	-2.756	0.856	-3.221	0.001	-4.434	-1.078	0.005
subject = 23	0.951	0.898	1.059	0.290	-0.809	2.711	0.001
subject = 24	-0.130	0.782	-0.166	0.868	-1.664	1.405	0.000
subject = 25	-9.976	0.574	-17.387	0.000	-11.102	-8.851	0.119
subject = 26	0.101	0.748	0.136	0.892	-1.366	1.569	0.000
subject = 27	-4.574	0.663	-6.895	0.000	-5.874	-3.273	0.021
subject = 28	-0.053	0.713	-0.074	0.941	-1.450	1.345	0.000
subject = 29	0.210	0.721	0.291	0.771	-1.204	1.624	0.000
subject = 30	1.140	0.766	1.488	0.137	-0.362	2.641	0.001
subject = 31	1.048	0.977	1.072	0.284	-0.869	2.964	0.001
subject = 32	1.149	0.764	1.504	0.133	-0.349	2.647	0.001
subject = 33	-1.358	0.827	-1.642	0.101	-2.981	0.264	0.001
subject = 34	-1.945	1.080	-1.801	0.072	-4.063	0.173	0.001
subject = 35	0.074	0.930	0.080	0.936	-1.750	1.898	0.000
subject = 36	0	.	.	.	.	.	.

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial-η^2
					Lower Bound	Upper Bound	
topic = 1	-1.726	.563	-3.065	.002	-2.830	-.622	.004
topic = 2	-1.776	0.530	-3.352	0.001	-2.816	-0.737	0.005
topic = 3	-1.981	0.548	-3.615	0.000	-3.055	-0.906	0.006
topic = 4	-1.840	0.540	-3.407	0.001	-2.900	-0.781	0.005
topic = 5	-2.269	0.563	-4.031	0.000	-3.372	-1.165	0.007
topic = 6	-2.552	0.541	-4.720	0.000	-3.612	-1.492	0.010
topic = 7	-1.391	0.578	-2.405	0.016	-2.526	-0.257	0.003
topic = 8	-1.811	0.552	-3.284	0.001	-2.893	-0.730	0.005
topic = 9	-2.180	0.567	-3.847	0.000	-3.291	-1.069	0.007
topic = 10	-2.503	0.517	-4.839	0.000	-3.517	-1.488	0.010
topic = 11	-2.923	0.550	-5.318	0.000	-4.001	-1.845	0.013
topic = 12	0	.	.	.	.	.	.
position = 1	-0.587	0.517	-1.137	0.256	-1.601	0.426	0.001
position = 2	-0.275	0.542	-0.507	0.612	-1.338	0.788	0.000
position = 3	-0.784	0.530	-1.480	0.139	-1.822	0.255	0.001
position = 4	-0.963	0.523	-1.841	0.066	-1.990	0.063	0.002
position = 5	-1.188	0.601	-1.976	0.048	-2.367	-0.009	0.002
position = 6	-1.138	0.614	-1.853	0.064	-2.342	0.066	0.002
position = 7	-0.126	0.620	-0.203	0.839	-1.342	1.090	0.000
position = 8	-0.344	0.595	-0.578	0.564	-1.510	0.823	0.000
position = 9	-1.313	0.549	-2.392	0.017	-2.390	-0.236	0.003
position = 10	-0.007	0.605	-0.011	0.991	-1.194	1.180	0.000
position = 11	-1.100	0.580	-1.896	0.058	-2.238	0.038	0.002
position = 12	0	.	.	.	.	.	.

G.2 – Subject, topic, and position parameters for model of inter-query time interval

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial-η^2
					Lower Bound	Upper Bound	
subject = 1	.074	.046	1.628	.104	-.015	.164	.001
subject = 2	-.070	.068	-1.039	.299	-.203	.062	.001
subject = 3	.019	.084	.230	.818	-.145	.184	.000
subject = 4	.070	.087	.805	.421	-.101	.242	.000
subject = 5	.045	.070	.643	.520	-.092	.181	.000
subject = 6	.122	.062	1.956	.051	.000	.244	.002
subject = 7	-.146	.051	-2.869	.004	-.246	-.046	.004
subject = 8	.340	.115	2.959	.003	.114	.565	.005
subject = 9	.526	.091	5.772	.000	.347	.705	.018
subject = 10	.364	.069	5.308	.000	.230	.499	.015
subject = 11	.125	.089	1.407	.160	-.049	.300	.001
subject = 12	.176	.061	2.885	.004	.056	.296	.005
subject = 13	.273	.079	3.474	.001	.119	.427	.007
subject = 14	-.041	.060	-.673	.501	-.159	.078	.000
subject = 15	.362	.070	5.201	.000	.225	.498	.015
subject = 16	.046	.055	.836	.403	-.062	.154	.000
subject = 17	.094	.066	1.424	.155	-.035	.223	.001
subject = 18	-.094	.048	-1.975	.048	-.187	-.001	.002
subject = 19	.136	.057	2.404	.016	.025	.247	.003
subject = 20	-.018	.056	-.323	.747	-.128	.092	.000
subject = 21	.071	.064	1.102	.271	-.055	.196	.001
subject = 22	.256	.068	3.758	.000	.122	.390	.008
subject = 23	.017	.075	.223	.824	-.131	.164	.000
subject = 24	.312	.062	5.075	.000	.192	.433	.014
subject = 25	-.009	.046	-.194	.846	-.099	.081	.000
subject = 26	.071	.057	1.234	.217	-.042	.184	.001
subject = 27	-.186	.050	-3.686	.000	-.285	-.087	.007
subject = 28	.035	.055	.638	.523	-.073	.143	.000
subject = 29	-.217	.056	-3.885	.000	-.327	-.108	.008
subject = 30	.024	.060	.397	.691	-.094	.143	.000
subject = 31	.344	.086	3.980	.000	.175	.514	.009
subject = 32	.077	.060	1.272	.203	-.041	.195	.001
subject = 33	.013	.067	.194	.846	-.119	.145	.000
subject = 34	.535	.106	5.049	.000	.327	.742	.014
subject = 35	.191	.080	2.389	.017	.034	.348	.003
subject = 36	0	.	.	.	.	.	.

155

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial-η^2
					Lower Bound	Upper Bound	
topic = 1	-.022	.047	-.455	.649	-.114	.071	.000
topic = 2	-.080	.043	-1.835	.067	-.165	.005	.002
topic = 3	-.014	.046	-.305	.761	-.104	.076	.000
topic = 4	.001	.045	.016	.987	-.088	.089	.000
topic = 5	.038	.047	.809	.419	-.054	.130	.000
topic = 6	-.028	.045	-.627	.531	-.116	.060	.000
topic = 7	-.087	.049	-1.792	.073	-.183	.008	.002
topic = 8	-.037	.046	-.809	.419	-.127	.053	.000
topic = 9	-.044	.048	-.909	.363	-.137	.050	.000
topic = 10	-.085	.043	-1.974	.049	-.170	-.001	.002
topic = 11	-.001	.046	-.011	.991	-.091	.090	.000
topic = 12	0	.	.	.	.	.	.
position	-.015	.047	-.455	.649	-.114	.071	.000

REFERENCES

Allan, J., Carterette, B., & Lewis, J. (2005). When will information retrieval be 'good enough'? *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil*, 433-440.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum.

Banks, D., Over, P., & Zhang, N.-F. (1999). Blind men and elephants: Six approaches to TREC data. *Information Retrieval, 1*(1-2), 7-34.

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

Bruza, P., & Dennis, S. (1997). Query re-formulation on the internet: Empirical data and the hyperindex search engine. *Proceedings of the 5th International Conference Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) (RAIO 1997), Montreal, Quebec, Canada*, 488-499.

Card, S., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Carterette, B., & Bennett, P. (2008). Evaluation measures for preference judgments. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08), Singapore*, 685-686.

Carterette, B., Bennett, P., Chickering, D. M., & Dumais, S. (2008). Here or there: Preference judgments for relevance. *Proceedings of the 30th European Conference on IR Research (ECIR 2008), Glasgow, UK*, 16–27.

Clarke, C., Agichtein, E., Dumais, S., & White, R. (2007). The influence of caption features on click through patterns in web search. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), Amsterdam, The Netherlands*, 135-142.

Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), San Jose, California, USA*, 407-416.

Downey, D., Dumais, S., & Horvitz, E. (2007a). Heads and tails: Studies of web search with common and rare queries. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), Amsterdam, The Netherlands*, 847-848.

Downey, D., Dumais, S., & Horvitz, E. (2007b). Models of searching and browsing: Languages, studies, and application. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India,* 2740-2747.

Downey, D., Dumais, S., Liebling, D. J., & Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. *Proceedings of the 17th Conference on Information and Knowledge Management, Napa Valley, CA, USA,* 449-458.

Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01), Seattle, WA, USA,* 277-284.

e-kiwi, L. L. C. Screen-scraper software Retrieved October, 23, 2008, from http://www.screen-scraper.com/

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems, 23*(2), 147-168.

Goncalves, M. A., Fox, E., Krowne, A., Calado, P., Laender, A., da Silva, A., et al. (2004). The effectiveness of automatically structured queries in digital libraries. *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tucson, AZ, USA,* 98-107.

Google. Retrieved October 10, 2008, from www.google.com

Granka, L., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in www search. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04), Sheffield, UK,* 478-479.

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07), San Jose, CA, USA.*

Hearst, M. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern information retrieval* (pp. 257-340). New York: ACM Press.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K. P. (2002). Finding the flow in web site research: Designing a search system and interface may be best served (and executed) by scrutinizing usability studies. *Communications of the ACM, 43*(9), 42-49.

Hill, T., & Lewicki, P. (2006). *Statistics methods and applications: A comprehensive reference for science, industry, and data mining.* Tulsa, OK: StatSoft.

Hunter, R. (1991). Successes and failures of patrons searching the online category at a large academic library: A transaction log analysis. *RQ, 30*, 395-402.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada*, 133-142.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil*, 154-161.

Kantor, P. B. (1987). A model for the stopping behavior of users of online systems. *Journal of the American Society for Information Science and Technology, 38*(3), 211-214.

Kelly, D., & Belkin, N. (2001). Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01), New Orleans, LA, USA*, 408-409.

Kelly, D., & Belkin, N. (2004). Display time as implicit feedback: Understanding task effects. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04), Sheffield, UK*, 377-384.

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *ACM SIGIR Forum, 37*(2), 18-28.

Kemp, C., & Ramamaohanarao, K. (2002). Long-term learning for web search engines. *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland*, 263-274.

Keppel, G., & Wickens, T. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson Prentice Hall.

Kim, J., Oard, D., & Romanik, K. (2001). *User modeling for information access based on implicit feedback*. Poster/short paper presented at the Third ISKO Workshop on Information Filtering, Paris, France.

Klockner, K., Wirschum, N., & Jameson, A. (2004). Depth- and breadth-first processing of search result lists. *Proceedings of the CHI '04 Extended Abstracts on Human Factors in Computing Systems, Vienna, Austria*.

Konstan, J., Miller, B., Maltz, D., Herlock, J., Gordon, L., & Riedl, J. (1997). Applying collaborative filtering to Usenet news. *Communications of the ACM, 40*(3), 77-87.

Kural, Y., Robertson, S., & Jones, S. (2001). Deciphering cluster representations. *Information Processing & Management, 37*(4), 593-601.

Lagergren, E., & Over, P. (1998). Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 98), Melbourne, Australia,* 164-172.

Lau, T., & Horvitz, E. (1999). Patterns of search: Analyzing and modeling web query refinement. *Proceedings of the Seventh International Conference on User Modeling, Banff, Canada,* 119 - 128.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th ACM International Conference on Design of Communication (SIGDOC '86), Toronto, Ontario, Canada,* 24-26.

Leuski, A., & Allan, J. (2004). Interactive information retrieval using clustering and spatial proximity. *User Modeling and User-Adapted Interaction, 14*(2-3), 259-288.

Liu, Y.-S., & Wacholder, N. (2008). Do human-developed index terms help users? An experimental study of mesh terms in biomedical searching. *American Society for Information Science and Technology, 45*(1). Retrieved from http://dx.doi.org/10.1002/meet.2008.1450450284

Lorigo, L., Haridassan, H., Brynjarsdottir, H., Xia, L., Joachims, T., Gay, G., et al. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology, 59*(7), 1041-1052.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using Google. *Information Processing & Management, 42,* 1123-1131.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ: Lawrence Earlbaum Associates.

McInerney, C., & Bird, N. (2005). Assessing website quality in context. *Information Research, 10*(2), 213. Retrieved from http://InformationR.net/ir/10-2/paper213.html.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94), Dublin, Ireland,* 272-281.

Muller, H.-M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology, 2*(11), e309. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pbio.0020309

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Oard, D., & Kim, J. (1998). *Implicit feedback for recommender systems.* Poster/short paper presented at the AAAI Workshop on Recommender Systems, Madison, WI, USA.

Phan, N., Bailey, P., & Wilkinson, R. (2007). Understanding the relationship of information need specificity to search query length. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), Amsterdam, The Netherlands,* 709-710.

Rieh, S. Y., & Xie, H. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing and Management, 42,* 751-768.

Rosenthal, R. (1969). Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press.

Sanderson, M., & Dumais, S. (2007). Examining repetition in user search behavior. *Proceedings of the 29th European Conference on IR Research (ECIR 2007),* 597-604.

Smith, C. L. (2008). Searcher adaptation: A response to topic difficulty. *American Society for Information Science and Technology, 45*(1). Retrieved from http://dx.doi.org/10.1002/meet.2008.1450450381

Smith, C. L., & Kantor, P. B. (2008). User adaptation: Good results from poor systems. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore,* 147-154.

Spoerri, A. (2006). Visualizing meta search results: Evaluating the Metacrystal toolset. *American Society for Information Science and Technology, 43*(1). Retrieved from http://dx.doi.org/10.1002/meet.1450430174

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Earlbaum Associates.

Sun, Y., & Kantor, P. B. (2006). Cross-evaluation: A new model for information system evaluation. *Journal of the American Society for Information Science and Technology, 57*(5), 614-628.

Turpin, A., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01), New Orleans, LA, USA*, 225-231.

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06), Seattle, WA, USA*, 11-18.

Voorhees, E., & Buckland, L. (2008). Appendix: Common evaluation measures. *NIST Special Publication 500-274: The Sixteenth Text REtrieval Conference Proceedings (TREC 2007)* Retrieved from http://trec.nist.gov/pubs/trec16/appendices/measures.pdf

Wacholder, N., Kelly, D., Kantor, P., Rittman, R., Sun, Y., Bai, B., et al. (2007). A model for quantitative evaluation of an end-to-end question-answering system. *Journal of the American Society for Information Science and Technology, 58*(8), 1082-1099.

White, R., Jose, J., & Ruthven, I. (2001). Comparing explicit and implicit feedback techniques for web retrieval: TREC-10 interactive track report. *Proceedings of the Text REtrieval Conference (TREC), Gaithersburg, MD, USA*, 534-538.

White, R., & Ruthven, I. (2006). A study of interface support mechanisms for interactive information retrieval. *Journal of the American Society for Information Science, 57*(7), 933-948.

White, R., Ruthven, I., & Jose, J. (2002). The use of implicit evidence for relevance feedback in web retrieval. *Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research, Glasgow, UK*, 93-109.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96), Zurich, Switzerland*, 4-11.

Zhang, J., & Marchionini, G. (2005). Evaluation and evolution of a browse and search interface: Relation browser++. *Proceedings of the 2005 National Conference on Digital Government Research, Atlanta, GA, USA*.

CURRICULUM VITAE

CATHERINE L. SMITH

2010   Ph.D., Information Science, Rutgers University

2009   Certificate in Cognitive Science, Rutgers University

1987   M.B.A., Simmons College

1977   B.A, Bard College

RELEVANT PUBLICATIONS

Smith, C.L. (2009). Sensitivity to the results list: A response to poor system performance. Poster presented at the *Association for Library and Information Science Education Annual Conference (ALISE '09)*, Denver, CO.

Smith, C.L. (2008). Searcher adaptation: A response to topic difficulty. *Proceedings of the Annual Conference of the American Society for Information Science and Technology (ASIST '08)*, Columbus, OH.

Smith, C.L. (2008). What might users be learning from the system?. Workshop paper/poster presented at the *Workshop on Human-Computer Interaction and Information Retrieval (HCIR '08)*, Seattle, WA.

Smith, C.L. & Kantor, P.B. (2008). User adaptation: Good results from poor systems. *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*, Singapore, 147-154 (17% acceptance rate)

Muresan, G., Smith, C. L., Cole, M., Liu, L., & Belkin, N. (2006). Detecting document genre for personalization of information retrieval. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS '06)*, Kauai, HI.

Muresan, G., Cole, M., Smith, C. L., Liu, L., & Belkin, N. (2006). Does familiarity breed content? Taking account of familiarity with a topic in personalizing information retrieval. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS '06)*, Kauai, HI.

Muresan, G., Liu, L., Cole, M., Smith, C. L., & Belkin, N. (2005). The effect of document readability on perceived familiarity and relevance. *Proceedings of the Annual Conference of the American Society for Information Science and Technology (ASIST '05)*, Charlotte, NC.