AUTHENTIC VS. SYNTHETIC: A COMPARISON OF DIFFERENT METHODS FOR STUDYING TASK-BASED INFORMATION SEEKING

BY

YIWEI WANG

A dissertation submitted to the

School of Graduate Studies

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 Media

Written under the direction of

Chirag Shah

and approved by

____

____

____

____

New Brunswick, New Jersey

October 2020

ABSTRACT OF THE DISSERTATION

Authentic vs. Synthetic: A Comparison of Different Methods for Studying Task-based Information Seeking

by YIWEI WANG

Dissertation Director: Chirag Shah, Ph.D.

In task-based information seeking research, researchers often collect data about users' online behaviors to predict task characteristics and personalize information for users. User behavior may be directly influenced by the environment in which a study is conducted, and the tasks used. This dissertation investigates the impact of study setting and task authenticity on users' searching behaviors, perceived task characteristics, and search experiences. Thirty-six undergraduate participants finished one lab session and one remote session in which they completed one authentic and one simulated task. The findings demonstrate that the synthetic lab setting and simulated tasks had significant influences mostly on behaviors related to content pages, such as page dwell time and the number of pages visited per task. Meanwhile, first-query behaviors were less affected than whole-session behaviors, indicating the reliability of using first-query behaviors in task prediction. Subjective task characteristics—such as task motivation and importance—also varied in different settings and tasks. Interview data reveal why users were influenced. This dissertation addresses methodological limitations in existing research and provides new insights and implications for researchers who collect online user search behavioral data.

ACKNOWLEDGEMENTS

First and foremost, I am grateful for the guidance, help, and support I have received from my advisor and dissertation chair, Dr. Chirag Shah. I would not be able to finish my PhD studies without his invaluable support and encouragement throughout this journey. It is also my great privilege to learn from my dissertation committee members: Dr. Nicholas J. Belkin, Dr. Kaitlin L. Costello, and Dr. Diane Kelly. They offered me insightful comments and suggestions that tremendously strengthened my dissertation. Their own works also provided great sources of inspiration in my research. I owe earnest thankfulness to all of my study participants without whom this research would not be possible.

I would like to thank Dr. Charles Senteio for the opportunity to work in several of his research projects. He gave me training and insights that greatly influenced my interest in health information behavior and qualitative research. I would also like to thank Dr. Xiaomu Zhou for her mentorship early in the PhD program and Dr. Sunyoung Kim for her generous support in teaching.

I am thankful for my colleagues in the InfoSeeking Lab: Dr. Chathra Hendahewa, Dr. Long Le, Dr. Ziad Matni, Dr. Dongho Choi, Dr. Matthew Mitsui, Jiqun Liu, Shawon Sarkar, Souvick Ghosh, Soumik Mandal, Manasa Rath, Jonathan Pulliza, Ruoyuan Gao, Shannon Tomlinson, and Kevin Albertson. Especially, I want to thank Dr. Mitsui who provided enormous support in several of my projects. I am also very fortunate to have collaborated with Jiqun Liu, Shawon Sarkar, and Soumik Mandal in various research projects.

My dissertation is supported by the School of Communication and Information Dissertation Grant and Library and Information Science Area Dissertation Support Grant. I would like to thank the donors for their generosity.

Many thanks go to my mother Jisong Li and father Chaoguang Wang, who inspired me to pursue this degree and provided me continuous love, kindness, and support. My cousin Xiaodan Zhu and my best friend Yiran Kang have been consistent companions whenever I want to share my progress or frustration.

I am most grateful for the support I have received from my husband Lei Yang, who comforted me when I struggled and pushed me forward. Last but not the least, I want to thank my baby girl, Annie Yang. She has given me great joy in these last few months of my PhD studies.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1 : Introduction

1.1 Background

Tasks as a driving force of information seeking have been identified as essential elements of context affecting information seeking behaviors (Ingwersen & Järvelin, 2005). In recent decades, task-based information seeking (TBIS) has drawn considerable research attention. Accounting for users' tasks in the study and design of information systems assists researchers in finding out for what purposes the system is used (Saastamoinen & Järvelin, 2017), and thus informs system design to personalize information according to users' tasks at hand. One major focus of the research in TBIS is investigating the relationships between task characteristics and information seeking behaviors. The literature has not only provided theoretical frameworks for studying TBIS but has also presented empirical evidence of the systematic relationships between task characteristics and information seeking behaviors (e.g., Jiang, He, & Allan, 2014).

Researchers have adopted various methods to study TBIS. The design of each study involves several components – such as study setting and task – that can be authentic (i.e., part of a user's real life) or synthetic (i.e., something created and provided by the researcher). User studies may be conducted in a laboratory setting or at users' locations of choices, such as their workplace or home. Researchers could ask study participants to bring their own tasks (i.e., authentic tasks) or assign them simulated tasks, a concept initially proposed by Borlund (2000). A simulated work task situation provides a textual description of a realistic situation that motivates users to search for information. While researchers often advise participants to search as they usually would during a study in order to observe their natural behaviors, they cannot conduct all studies in naturalistic settings due to a number of factors (e.g., equipment in the lab, confounding factors in the field). In fact, the majority of TBIS studies have been conducted in a laboratory setting, assigning participants simulated search tasks. A few studies also took place in users' natural work or home environments and elicited their real work or everyday life tasks (e.g., He & Yilmaz, 2017; Saastamoinen & Järvelin, 2017).

Each type of study setting or task has its strengths and limitations. There has been constant debate about the strengths and limitations of each method and how real search behaviors and tasks could be studied. For example, laboratory setting is often criticized for being artificial, while field setting has brought other concerns such as lack of control. Although researchers have included authentic or synthetic components or both in their studies, they rarely systematically compare how different authentic and synthetic study components affect participation behaviors and the implications those differences have in our interpretation of study results.

This dissertation is an attempt to investigate the differences between authentic and synthetic study settings and tasks in respect to their influences on users' search behaviors (e.g., queries, page visits), subjective perceptions of task characteristics (e.g., task difficulty, task interest), and search experiences (e.g., engagement). The remainder of the dissertation is organized as follows. The next section of Chapter 1 provides a problem statement. Chapter 2 reviews relevant TBIS literature, existing comparison studies and a preliminary work that inspired this research. Chapter 3 introduces the conceptual framework that has guided this dissertation. It is followed by a description of the data collection and data analysis methods in Chapter 4 & 5. Chapter 6 reports the quantitative analysis results of this research. Chapter 7 discusses the quantitative analysis results with evidence found in the interview data. Chapter 8 concludes this dissertation by briefly summarizing the results, discussing practical implications, and proposing directions for future research.

1.2 Problem Statement

Generally speaking, existing studies in task-based information seeking can be divided into two broad categories: large-scale log analyses and user studies (i.e., studies that recruit real users). Large scale log analyses utilize real world datasets containing millions of queries and pages (e.g., Lucchese, Orlando, Perego, Silvestri, & Tolomei, 2011). These data reveal users' natural searching and browsing behavior when they work on real life tasks. However, the actual tasks that trigger information seeking activities are not observed, and thus their characteristics and how they can be mapped to search behaviors are not clear (He & Yilmaz, 2017). Search sessions were often inferred by annotators based on time or topics which may not represent users' real task sessions. As a result, researchers may not be able to see how effectively users' tasks have been supported by relying solely on log data.

Rooted in the cognitive viewpoint of information seeking and retrieval, TBIS studies that involve real users (hereafter: user studies) take into account users' work or everyday life tasks as well as their knowledge and experience (Ingwersen, 1996; Ingwersen & Järvelin, 2005). Tasks utilized in those studies should be either assigned by researchers or reported by study participants. Thus, task content and boundary are clear.

User studies have been widely adopted by interactive IR researchers who focus on task-based information seeking.

User studies can be conducted in a controlled laboratory setting where researchers and participants are co-located, or remote/field settings where users work at their own home, office, or other locations and they may be separated spatially and/or temporally from the researcher (Greifeneder, 2012). Researchers may travel to users' work or home places to do field observations or interviews. They may also collect users' search activities remotely using computer programs.

Study participants may be assigned a few simulated task situations in which they are required to look for information to fulfill the task requirements given by the researcher. Simulated tasks may not be relevant to participants' real life and the outcomes of those tasks may not have any real impacts on them. Participants may also be asked to work on their own tasks, which should be something that they need to complete in reality, even without being requested by a study (i.e., authentic tasks). The outcomes of authentic tasks may more or less have a real impact on participants' lives. Few studies employ both authentic tasks and simulated tasks. There are, of course, other components to consider including participant population (e.g., students, professionals), time limit, device (e.g., desktop computer, mobile devices), and information sources (e.g., documentary sources, human sources).

While researchers often chose one set of methods (e.g., running lab studies using simulated tasks) out of convenience or constraints, and discuss the strengths and limitations of their methods of choice, only a few have investigated the way research is conducted and how different methods affect study results. It was often pointed out in the limitation section of a paper that a controlled lab study or simulated tasks may not represent users' real search environment or tasks. However, if researchers have not actually made comparisons between different methods in the same study, it is not clear how different they are in terms of the influences on users' search behaviors or if they affect search behaviors at all. Also, even when participants work on both authentic and simulated tasks in a study, their behaviors may still be constrained by the study design. For example, Borlund and Ingwersen (1999) required their participants to bring an information need that was able to be met by a newspaper full-text collection. This kind of boundary may potentially lessen the authenticity of an authentic task and search behaviors.

The difference between authentic and synthetic study settings and tasks remain unclear. If we are to use real-life behavioral variables to predict task characteristics and personalize search for users, it is important to know how different study settings and tasks affect search behaviors and what the differences mean for us researchers. Therefore, I address the following research questions in this dissertation:

- RQ1: Do study settings (i.e., lab setting versus remote setting) affect users' Web search behavior (e.g., average query segment time and page visits per task)?
- RQ2: Does the authenticity of study tasks (i.e., authentic tasks versus simulated tasks) affect users' Web search behavior?
- RQ3: Do users' perceptions and experiences of tasks (e.g., task difficulty, engagement) differ in different study settings and tasks?

1.3 Summary

To summarize, two major components to consider when designing a TBIS study are study settings and task authenticity. While there have been constant concerns regarding the authenticity of laboratory studies and simulated tasks, limited research was done to systematically compare search behaviors and experiences between lab and field settings and between authentic and simulated tasks. There has not been enough evidence showing if users look for information or perceive tasks differently while performing tasks of different authenticity in different study settings, and more importantly, what the differences are and why those differences occur. This dissertation investigates the impact of study setting and task authenticity on users searching and browsing behaviors, task perceptions, and search experiences.

Chapter 2 : Literature Review

This literature review attempts to understand 1) the different methods used in existing research to study task-based information seeking and 2) what existing research has reported on the impact of study setting and task authenticity on user behavior. To offer a context for this research, I first provide a general overview of task-based information seeking, relevant models, and task classifications. Next, I review TBIS studies that were conducted in lab or field settings (including remote and naturalistic studies), followed by a comparison of the two settings. Comparison studies that examined the differences between authentic tasks and simulated tasks are reviewed next. Lastly, I report on a preliminary work that has inspired this dissertation.

2.1 Task-based Information Seeking

A task is a set of linked concrete or cognitive activities performed by individuals or systems to fulfill a goal (Byström, 2007; Vakkari, 2003). It often has some type of requirement set by the task doer or others. Researchers have defined or bounded tasks differently in various studies. A task can be work-related (i.e., something people do to fulfill their job responsibilities) (Li & Belkin, 2008). It can also be generated by interests, problems, or situations outside of a work context, such as finding a doctor and planning a trip.

Information seeking activities do not occur in a vacuum but are often triggered by an underlying task. Users encounter a problematic situation in which their state of knowledge is inadequate to complete a task, and thus have an information need perception (Belkin, Oddy, & Brooks, 1982). In TBIS, people look for information to finish a task, either related to their work or everyday lives. Researchers in TBIS are mainly concerned with information seeking behavior during task performance. In early studies, tasks were mostly studied in a professional or academic context (e.g., Kuhlthau, 1991; Leckie, Pettigrew, & Sylvain, 1996). Savolainen (1995) argued that everyday life information seeking (ELIS) is of equal importance as job-related information seeking. ELIS refers to the acquisition of information that is used to solve daily life problems (e.g., consumption, health care) not directly connected to the performance of occupational tasks, though issues of work-related or non-work-related information seeking may overlap. There were also researchers who use the term “work task” to cover all kinds of tasks including everyday life tasks (Ingwersen & Järvelin, 2005}.

Researchers have demonstrated that various task characteristics affect how users seek and use information. Task characteristics are believed to have a fairly predictable influence on users’ information behavior (Talja & Nyce, 2015). Some of the most frequently studied task facets or characteristics include *task complexity* (Byström, 2002; Liu, Gwizdka, Liu, Xu, & Belkin, 2010; Kelly, Arguello, Edwards, & Wu, 2015; Saastamoinen & Järvelin, 2017), *task difficulty* (Aula, Khan, & Guan, 2010; Gwizdka & Spence, 2007; Li & Hu, 2013), *task goal* (Jiang, He, & Allan, 2014; Liu, Kim, Creel, 2015), *task knowledge (about the topic or procedure)* (Wirth, Sommer, Pape, & Karnowski, 2016), *task products* (Xie, 2009; Liu et al., 2015), *task stage* (Kuhlthau, 1991; Vakkari, 2001), and *task urgency* (Xie, 2009; Wang & Shah, 2017). Some characteristics, such as task complexity and difficulty, could be defined objectively by the researchers or subjectively as perceived by information seekers. For example, some studies have manipulated objective task complexity in the study design while also eliciting perceived task complexity from the participants (e.g., Liu et al., 2010). Also, users' perceptions of task complexity and difficulty may not be static during task performance so their post-task perceptions may not match pre-task perceptions (Liu, Liu, Yuan, & Belkin, 2011).

The literature has presented the relationships between task characteristics and information seeking behaviors of various levels. Broadly speaking, task characteristics may affect individuals' overall information seeking strategies. An information seeking strategy includes the sources an individual use and the techniques employed to use that information source (Pharo, 2004). For example, individuals are more likely to consult people than documentary sources when they perceive a task to be more complex (Byström, 2002). In the context of Web searching, researchers have examined the relationships between search task characteristics and a number of browsing and searching behavioral variables on task, query, and page level (e.g., query length, number of queries, page dwell time, decision time) (e.g., Gwizdka & Spence, 2006; Liu et al., 2010). Tasks and individuals' information seeking behaviors under the influence of task characteristics subsequently influence information seeking outcomes (i.e., successes) (e.g., Wirth et al., 2016).

2.2 Models in Task-based Information Seeking

Several information seeking and retrieval models have taken into account the characteristics of tasks. Kuhlthau (1991) developed a model of information search process (ISP) based on a series of empirical research conducted in schools and libraries.

This model identifies six stages of information seeking for complex tasks (e.g., writing a research paper). Different task stages (e.g., initiation, exploration) are associated with different types of information sought (e.g., background information, focused information). One major contribution of this model is the affective aspect of information seeking. Individuals feel uncertain at the initial stage of task performance and go through the transition from frustration to clarity. Their uncertainty fades away as they move forward in their search process.

Vakkari (2001) built on the ISP model and developed a theory of the task-based information retrieval process which examines the relationships between task stage and information searching behaviors. It sets individuals' task performance process as the point of departure for studying information searching. Compared to Kuhlthau's model, Vakkari's theory is more specific in the domain of information retrieval. In addition to the common behavioral variables shared by Kuhlthau's and Vakkari's studies (e.g., type of information sought), Vakkari also pointed out the relationships between task stage and other variables such as the degree of relevance, relevance criteria, search queries, and search tactics. Both researchers studied complex long-term tasks (e.g., writing a paper) that may be composed of smaller sub-tasks, while there can also be short-term tasks in which task stage is not that apparent (Wang & Shah, 2017).

Byström and Järvelin (1995) conceptualized work-related tasks by their complexity and used a task's pre-determinability to indicate complexity. For example, the least complex tasks are automatic information processing tasks that are completely determinable, while genuine decision tasks are considered most complex because of their unexpected and unstructured nature. The task as perceived by the information seeker triggers the information seeking. The ways in which information seekers interpret information needs and proceed with information seeking depend on their perceived task complexity and other personal and situational factors (e.g., attitude, time). Byström and Järvelin discovered systematic relationships among task complexity, types of information, and information sources and channels. For instance, the complexity of information needed and the number of sources used increase when task complexity increases. Leckie et al. (1996) also limited the focus to professional context and considered "work roles" the root of information seeking which influence tasks. Tasks in turn influence the characteristics of information needs. Work and everyday life tasks also lie in the center of Ingwersen and Järvelin's (2005) cognitive framework of information seeking & retrieval, which is one of the theoretical frameworks that has guided this dissertation. The cognitive framework of IS&R is further discussed in Chapter 3. Conceptual Framework.

2.3 Task Classifications

The aforementioned models have underscored the importance of understanding the motivating tasks that generate information seeking and have served as the theoretical foundation for research in task-based information seeking. The TBIS literature has presented three levels of tasks: tasks, information seeking tasks, and information searching tasks (Byström & Hansen, 2005). Information seeking tasks are generated by their motivating tasks and a task may involve more than one information seeking tasks. They refer to users' information problem which must be solved for users to move on with their tasks. They could be accomplished through one or more consultations of information sources, such as searching in information systems, consulting other people, and reading physical documents. Information searching tasks are subsets of information seeking tasks in which users search for information from one or more sources.

Different tasks may motivate different information seeking or searching tasks (Li & Belkin, 2010). Information seeking or searching tasks can also be identical to work or everyday life tasks (Pharo & Järvelin, 2004). For instance, when a librarian is searching for information for a patron, his/her work task is a search task. Researchers seldom investigate tasks of all three levels and their relationships in one study. They have mostly examined the relationships between the characteristics of tasks at one level and individuals' behaviors. Li (2009) examined the relationships between work tasks and search tasks and the intra-relationships between the facets of work tasks and search tasks. She suggested that work tasks and search tasks are related in some aspects, but not all. Work tasks and search tasks may have different influences on individuals' information searching behavior and their effects should be considered differently.

Within each level, tasks can be classified into categories along different dimensions. Early studies primarily classified tasks along one dimension (e.g., stage, complexity). For example, tasks were often classified by their difficulty and complexity. Task difficulty has mostly been defined as a subjective construct based on task doers' perceptions such as perceived efforts (Wildemuth, Freund, & Toms, 2014). By contrast, task complexity can both be objectively defined or manipulated and subjectively perceived by task doers. Campbell (1988) identified four basic attributes of complexity: (1) presence of multiple paths to arrive at a desired outcome, (2) presence of multiple desired outcomes, (3) presence of conflicting interdependence among paths to multiple desired outcomes, and (4) presence of uncertain or probabilistic linkage among paths and outcomes. Based on these attributes, he created 16 possible task types that can be characterized into Simple tasks, Decision tasks, Judgment tasks, Problem tasks, and Fuzzy tasks. Simple tasks are the least complex ones that involve none of the task complexity attributes while Fuzzy tasks are the most complex tasks that have both multiple desired outcomes and multiple ways of reaching each outcome. Byström and Järvelin (1995) categorized tasks based on their *a priori* determinability. The least complex tasks are automatic information processing tasks in which the type of the task result, the work process, and types of information used can be determined in detail in advance. On the contrary, in genuine decision tasks, none of these aspects could be described in advance, making this type of task the most complex.

Compared to the higher-level work or everyday life tasks, information seeking and searching tasks have drawn more attention among researchers in information science. This dissertation will also limit the focus to search tasks. Apart from the complexity and difficulty facets, information seeking and searching tasks have been mainly classified by their source (e.g., internally generated, externally generated) (Reid, 2000) or product (e.g., factual tasks, interpretive tasks, exploratory tasks) (Kim, 2008; Kellar, Watters, & Shepherd, 2007; Marchionini, 1989), though task products have been defined somewhat differently by different researchers. Broder (2002)'s Web search taxonomy was one of the first classifications of Web search tasks, which categorized Web search into three classes: navigational search, informational search, and transactional search.

Li and Belkin (2008) filled in the research gap of no generally accepted classification and proposed a faceted classification that can be used for classifying all levels of tasks (i.e., task, information seeking tasks, and information search tasks).

Utilizing a consistent classification of tasks is essential for researchers to define and design tasks in a consistent way across studies, making the results from different studies commensurable. The scheme captures various generic facets including source, task doer, time, outcome, process, and goal; as well as common attributes such as interdependence, and users' perceptions of tasks. It holistically contextualizes users' information needs into their generative tasks. This faceted classification has been widely adopted by researchers in TBIS to define task facets and to capture users' subjective perceptions of task characteristics (Jiang et al., 2014; Liu et al., 2011).

The three sections above provide a general overview of tasks, task-based information seeking models, and task classifications. In the remainder of this chapter, I review different study settings and tasks used by TBIS research that examined task characteristics and user behaviors and existing comparison studies that examined the influence of various methods.

2.4 Methods for Studying Task-based Information Seeking

This subsection provides an overview of the existing TBIS studies that used different methods, categorized by study settings and task authenticity. It starts with a discussion of existing lab studies, followed by field studies (including remote and naturalistic studies). Next, a review of comparison studies of different study settings is presented with a highlight of the findings and limitations. This is followed by a review of studies that compared authentic and simulated tasks. A preliminary research that has inspired this study is briefly reported at the end.

2.4.1 Laboratory User Studies

Empirical research in task-based information seeking has heavily focused on the online environment. Particularly, observable online browsing and searching behaviors such as querying and dwelling behaviors (e.g., query length, page dwell time) have drawn continuous attention. The rationale behind this line of research is: if we know the relationship between task characteristics and search behaviors, we could predict task characteristics using users' search behaviors and personalize search to help users complete their tasks.

Most studies were conducted in lab settings where participants came individually to a computer lab and finished information search tasks. A lab is considered an artificial environment created by a researcher because it is not a place where users usually search for information. Search behaviors captured in search logs, coupled with questionnaires are usually used by researchers to analyze task characteristics and their relationships with user behaviors. Considering task characteristics along with observable behaviors like page dwell time could also predict document usefulness or relevance (Liu & Belkin, 2010). Kim and Allen (2002) suggested that search task type (known-item search vs. subject search) interacted with cognitive variables (e.g., problem-solving style) and together they influenced search behaviors, such as the number of pages viewed and the use of keywords. A series of lab studies as well as limitations of the lab setting are reviewed in the following paragraphs.

Although the results from different studies are somewhat inconsistent, this line of research has revealed the systematic relationships between task characteristics and users' search behaviors that could be used to personalize search for users. Gwizdka and Spence (2007) asked participants to work on factual information tasks (i.e. tasks in which users seeks a specific piece of data such as a name or a date) in a controlled lab setting. They adopted an objective measure of task complexity by considering experts' ratings of the length of navigational path, page complexity, and page information assessment. Meanwhile, they also collected participants' self-assessment of post-task difficulty. They discovered positive correlations between subjective post-task difficulty and various search behaviors such as the number of Web pages visited and page dwell time in factual task performance. They also found a negative correlation between search task difficulty and the degree of navigation path's linearity. Objective task complexity affected the relative importance of those behavioral variables and subjective perception of task difficulty. Kim (2008) also found higher post-task difficulty to be associated with more pages visited, though, unlike Gwizdka and Spence' results, this relationship only existed in exploratory tasks. Participants reformulated more queries, had more interaction with search engines, and spent more time when their exploratory tasks became more difficult.

Li and Belkin (2010) controlled two task facets in their lab study: product and complexity. Unlike previous studies that used only one information system, they discovered that work tasks affected individuals' selection of systems and their level of dependence on interaction with IR systems. Their results also indicated that product and complexity played different roles in influencing search behavior and significant interaction between these two was found in various aspects of search behavior. For example, when working on tasks at the same level of complexity, individuals' behaviors could still be varied significantly by task products. It is repeatedly echoed later by other studies that task should be viewed as a multifaceted variable and each facet taken alone may not provide enough information about the associations between task characteristics and search behaviors.

Liu et al. (2015) designed search tasks along four dimensions (i.e., product, goal, naming, and complexity). In addition to the three facets in Li and Belkin's (2008) classification, they also added a "naming" facet to define a named task that has a specified search target. They invited 48 undergraduate participants from two different universities to perform four tasks in a lab and elicited their self-reported task difficulty and difficulty reasons. They found some common reasons that led the users to perceive the tasks as difficult such as specific requirements and too much unrelated information. Some other difficulty reasons were also varied by tasks. Based on the findings, they suggested various implications for system design that could help users work on different tasks. Jiang et al. (2014) considered task product and goal in their lab study and adopted four tasks from TREC 2014 Session track that can be categorized along these two dimensions. They found that user behaviors are differentiated by different types of tasks in search activeness, browsing style, clicking strategy, and query formulation. More importantly, participants' behaviors changed as their tasks progressed.

Lab user studies give researchers a great extent of control since every participant uses the same computer in the same environment and researchers know exactly what has happened in a study session. As the aforementioned studies indicated, lab studies are useful in examining the effects of one or a few task facets while keeping the other facets constant across tasks. The lab environment helps reduce confounding variables like distraction and multitasking, which are hard to control in field settings. In the meantime, lab studies have a few major limitations. First, they have been criticized for being too artificial as almost all components in a lab setting are predefined by researchers and may not reflect users' real-life information seeking environment (Kelly, 2009; He & Yilmaz, 2017). In lab studies, participants typically finish all tasks straight without taking breaks or multitasking, which may be different from what they usually do in naturalistic settings. Multiple researchers suggested that participants may not perform tasks in the same way as they would do in a familiar environment due to laboratory effects brought by the setting itself or participant-researcher interaction (Andrzejczak & Liu, 2010; Greifeneder, 2012; Toms, Freund, & Li, 2004). They may also feel obliged to interact with the system in a lab setting until the allocated session ends (Zuccon et al., 2013).

Second, since participants and researchers are required to be in a lab at the same time, running those studies include both high temporal and monetary costs (e.g., traveling time and expenses, scheduling issues). Because of these high costs, lab studies often use a small number of participants who complete a limited number of tasks with a single system (Saastamoinen & Järvelin, 2017; Toms et al., 2004). In addition, most lab studies recruited university students and/or a restricted set of professionals (e.g., university faculty or staff) from the university where the lab study takes place because participants must visit a lab to finish the study. An alternative to lab studies is to study TBIS in users' natural search environment such as their home or workplaces, which encourage natural search behavior. A review of field studies is provided in the next section.

2.4.2 Field Studies

The environment and procedure in different lab studies can be very similar, typically including working on search tasks on a desktop computer and completing questionnaires. By contrast, field studies could utilize a wide range of methods. Some field studies are done completely online so a participant's physical appearance is not required. Those studies are also called remote user studies because of the geographical distance between participants and researchers (Greifeneder, 2012). Remote studies can be naturalistic (i.e., using natural setting and authentic tasks) or controlled by researchers (e.g., using simulated tasks, setting a time restriction). They can also be moderated or unmoderated. In moderated remote studies, participants and researchers have some kind of real-time connections, usually via online chatting (e.g., Andrzejczak & Liu, 2010). There are also naturalistic studies in which participants work at their own places, doing authentic tasks, and are interviewed and/or observed by researchers (e.g., Kumpulainen, 2014; Vuong, Saastamoinen, Jacucci, & Ruotsalo, 2019). In naturalistic studies, researchers and participants may be co-located. Their naturalistic information seeking behaviors may be captured by a combination of interview, observation, diary, or screen recording.

A few researchers examined the connections between task characteristics and information seeking behaviors by conducting naturalistic studies. Byström (2002) gathered data from two Finnish local governmental organizations. She used diaries and subsequent interviews to elicit participants' (municipal administrators) tasks, information required by those tasks, and information sources. This is one of a few studies that investigated information seekers' use of non-Web and non-document sources (i.e., human sources). A primary finding was that the increase of task complexity was related to the number of types of information needed, which was also related to the use of people as sources (e.g., experts, meetings). Similar findings turned up in a naturalistic study conducted by Saastamoinen and Järvelin (2012). They used multiple data collection methods and tools including forms, shadowing, and video recording. They found that the differences in searching behaviors between complex and simple tasks were the most obvious while the special features of semi-complex tasks were difficult to discover.

Kelly and Belkin (2004) conducted a fourteen-week naturalistic study in which seven Ph.D. students' online information seeking activities were collected. Participants reported the tasks that they were working on and classified visited document to their corresponding tasks using questionnaires and evaluation software. Kelly and Belkin observed significant differences between mean display times according to tasks. In addition, there was a lack of significant relationship between document display time and usefulness rating when contextual factors (e.g., task) were not included, showing the important role task plays in information seeking. Borlund and Dreier's (2014) recruited participants in the age range of 20–25 years from various professions and conducted naturalistic interactive information retrieval study. They investigated whether the information search behavior associated with Ingwersen's (1986) three types of information needs could be corroborated by empirical data. They confirmed that different behaviors were associated with different types of information need. For example, participants used the least of unique search terms, web pages, and unique websites when checking a specific piece of information.

Saastamoinen & Järvelin (2017) pointed out the scarcity of naturalistic field research of task-based information seeking using authentic tasks. Particularly, the authentic work tasks and search tasks they collected in the field did not resemble the simulated work or search tasks used in past lab studies. Most authentic search tasks they observed were simple with only one query involved and 42% of the work tasks did not include searching. This was mirrored by He and Yilmaz's (2017) field study of individuals' everyday life tasks that 41% of reported tasks included zero-query. This again calls for more comparison between authentic and simulated tasks and study settings.

Similar to lab studies, naturalistic studies that include intensive observation, shadowing, or interviews are expensive in terms of time and resources. Because of the high cost, naturalistic studies often have a small sample size. There may also be other restrictions or issues. For example, in Saastamoinen, Kumpulainen, and Järvelin's (2012) study of city administrators, installation of logging software or video recordings was not allowed by the city. In the meantime, the tasks collected in naturalistic studies often vary on a number of dimensions so a large sample size is needed in order to run statistical tests to examine the effects of each task characteristic.

One way to reduce costs and to allow participants work in their natural environment is conducting remote (online) studies in which participants and researchers are not co-located. For example, Kellar et al. (2007) focused on online interactions and instructed participants to search in their natural environment and to provide task categorizations of their Web usage. Information gathering tasks were shown to be the most complex for which participants spent more time and viewed more pages. Capra, Arguello, & Zhang (2017) recruited participants from Amazon's Mechanical Turk (hereafter MTurk) and tested the effects of task determinability on their search behaviors. Participants were required to bookmark at least three pages for a task. They found out that specifying dimensions by which the items can differ in comparative tasks significantly varied participants' search behaviors from not specifying dimensions. One major advantage of conducting remote studies on crowdsourcing platforms is the low monetary cost. In Capra et al.'s study, participants were only paid 0.3 USD per task. Although these two studies were not conducted in a lab, they were not naturalistic because researchers still controlled certain aspects of search, and thus they could also be called remote studies.

Compared to lab studies and field observations or interviews, remote user studies require less or no travel time on participants' or researchers' side. In the studies that do not require synchronized communication with researchers, scheduling issues can also be avoided. These advantages permit researchers to recruit more heterogeneous groups of participants, to utilize crowdsourcing platforms, and to run multiple remote tests simultaneously. However, they are not without drawbacks. Remote studies require clearer (and possibly longer) written instructions, as the researcher may not be available to promptly answer users' questions (Toms et al., 2004). Remote locations may contain uncontrollable factors that affect their search behavior such as distraction by other people or tasks that researchers do not know enough about (Greifeneder, 2016; Hendahewa & Shah, 2015). Although these confounding variables are real parts of users' natural experiences, they make it challenging to compare behaviors among participants and to examine the relationship between a specific pair of variables. Also, researchers are not able to see other cues that may show users' experience such as facial expression and body movement. For example, it may not be clear that if a longer page dwell time indicates user engagement, frustration, or distraction.

2.4.3 Comparison Between Lab and Field Studies

Study settings' impact on users' behaviors have been extensively reviewed by social science researchers in psychology and human-computer interaction. However, studies comparing search behaviors are very limited and have shown inconsistent results. Findings from existing lab studies and field studies are often not directly comparable due to different behavioral variables measured and different study designs adopted. As discussed in the last section, some field studies are naturalistic studies that aimed to explore participants' natural information seeking behaviors. They provided little constraint in terms of tasks, time, location, and sources (e.g., Saastamoinen & Järvelin, 2017). They have a heavier focus on users' information seeking strategies and tactics such as information source selection but less on the connections to systems. Their goal is not limited to examining the relationships between task characteristics and observable online search behaviors. Thus, participants' activities are often not restricted to searching on the Web. For instance, participants could consult other people, organizational systems or physical documents (e.g., Bystrom, 2002; Saastamoinen & Järvelin, 2012). Therefore, search behavioral measures (e.g., query length, page dwell time) were very limited in those studies. In some studies, detailed search data may not be obtained at workplaces due to privacy issues. By contrast, lab studies mostly only focus on online searching and they measure implicit querying and dwelling behaviors (e.g., query length, page dwell time) that can be extracted from log data.

A few researchers have compared various aspects of online information interactions between lab and field participants. Kelly and Gyllstrom (2011) were among the first who explored the behavioral differences between remote participants and lab participants. They used a TREC test collection that contained over a million newspaper articles and assigned each participant four test topics. They compared search behaviors (e.g., number of queries), participation behaviors (e.g., time taken to complete the study), and evaluation behaviors (e.g., self-reported topical knowledge, engagement) between undergraduate participants in a lab and those who chose to work at locations of their choices. Overall, they did not find significant differences between two settings for most measures. However, greater variances were observed for the number of documents opened and saved by remote participants and the amount of time taken by remote participants to complete the study. Lab participants were also more interested in learning about the topic than remote participants.

Borlund, Dreier, and Byström (2012) compared the time spent on searching as well as participants' explanation of search time between two independent studies, one conducted in a lab and one conducted in naturalistic settings. The two studies were carried out in a similar procedure despite the location. Participants in both studies finished all tasks in one sitting with the presence of a researcher. Overall, lab participants spent more time on tasks than field participants. Borlund et al. suggested that the lab study might create a "pleasing effect", which led the participants to try to "please" the researcher by devoting extra efforts and time. It is worth noting that the two studies were not directly comparable because they involved two different samples of participants (university students vs. professionals from various fields) who searched for different types of tasks.

Zuccon et al. (2013) examined the differences in search behaviors, answer correctness, and system effectiveness assessments between lab participants (university students) and crowdsourcing participants from Amazon's Mechanical Turk who finished the study remotely. Similar to Kelly and Gyllstrom's (2011) results, remote sessions showed a higher variance than lab sessions. They also observed that lab participants submitted more queries than crowdsourcing participants. However, considering that both the study settings and participant populations (i.e., college students vs. MTurk Workers) were different between groups, it was not clear whether the differences could be attributed to the participants or study settings. Denvir (2017) recruited university and high school students for a comparison between lab and remote experiments. She had a similar finding that remote participants executed fewer searches than participants in a school computer lab, though the lab used were not an individual lab typically used by lab studies due to study constraints.

Generally speaking, the aforementioned studies suggested that user behaviors are not significantly different between lab and field studies for the most part. To some extent, lab participants may work harder than field participants, but this divergence manifested in different ways in each study (e.g., issuing more queries, spending more time). However, a newer study of image searching carried out by Wu, Mao, Liu, Zhang, and Ma (2019) provided contrary evidence that users spent more time and issued more queries in the field setting. In their study, field participants' natural image search activities were recorded for a month in naturalistic settings while lab participants completed tasks designed by the researchers. Thus, the difference may be caused by both environments and tasks or one of them.

Study settings' influences have been relatively more researched in related fields such as human-computer interaction. A few studies with mixed results are discussed here. Bruun, Gull, Hofmeister, and Stage (2009) compared three remote asynchronous usability testing methods while also used a conventional lab-based think-aloud testing as a benchmark. Although all three remote methods performed significantly worse than the classical lab test in terms of the number of usability issues discovered, they also took significantly less efforts than a lab study. Andrzejczak and Liu (2010) reported no differences between lab and remote environments except the average task times. However, they instructed remote participants to perform tasks in a public computer lab, not their natural environments. While the researchers were not by participants' side, they communicated through an instant messaging tool during the study session.

Greifeneder (2012) examined if distraction in users' natural environment produces systematic mistakes in digital library usability testing by comparing participants' search behaviors and task outcomes in a lab and locations of their choices. She suggested that distractions and multitasking in naturalistic settings could require a mental shift into another activity and reduce users' attention even if the interruption is minimal. In her research, distraction during a remote session was self-reported by participants. She did not find a significant effect on most of the measures except that participants in the naturalistic setting spent significantly more time to complete the questionnaires than lab participants. She believed that the location of a study does not matter if the researcher has enough knowledge of participants' distractions and activities in a naturalistic setting. Alharbi and Mayhew (2015), however, observed no significant difference in the time spent on tasks and performance between participants in the lab and remote environments, despite the distraction caused by other tasks in participants' natural environments. Takahashi and Nebe (2019) found that remote participants offered richer feedback regarding their system than lab participants. They also noticed that the presence of a moderator did not bring any significant differences between groups. It is important to note that the main focus of these studies was usability testing, not search behaviors, and thus the results are not entirely transferable to TBIS research.

In summary, studies that compared users' online searching and browsing behaviors in the lab and field settings paint a mixed picture. While most measures were not significantly affected, several effects of study settings were discovered in various findings. Due to the limited existing comparison research in TBIS, I also consulted relevant studies in human-computer interaction, which generally showed that field participants took a longer time than lab participants in completing tasks or questionnaires.

2.4.4 Task Design

Another critical component in TBIS study design is the task design, as the nature of the search tasks and how they are presented to users may influence study results (Wildemuth, Freund, & Toms, 2014). The majority of lab studies required participants to work on tasks created and assigned by researchers (i.e., simulated tasks). Simulated tasks are often used in experiments because task characteristics can be controlled (Byström & Hansen, 2005). Some studies provided a simulated work task situation, as suggested by Borlund (2000), to create a context of information seeking and motivate participants to construct their own information need as in real life. Rieh (2002) assigned participants "generic tasks", which outlined a context, but did not specify information problems. For example, participants were asked to find good papers for a research project they were engaged. While they were required to find "good papers", they could choose their own research projects. This type of tasks allows comparison across tasks while also be partly authentic. The authenticity of a task is important in motivating authentic task engagement. In information seeking and retrieval research, simulated tasks usually have no actual consequences for the participants because they are paid regardless of their performance. This makes the authenticity of participants' information seeking behaviors questionable.

Borlund (2000) developed the concept of simulated work task situation to make IR experiments close to actual information seeking environment but are still relatively controlled so different users' information seeking behaviors could be compared. A simulated work task situation is a short 'cover story' describing a situation that requires retrieving information. The simulated task-based approach bridges the system-driven and user-centered approach of IR research by creating realism while also ensuring experimental control since the tasks used for a study are the same for every participant. The concept has been widely adopted in information seeking and retrieval research, particularly IR evaluations. A group of researchers from the University of North Carolina at Chapel Hill, the University of British Columbia, and the University of Sheffield built an online database of assigned search tasks that provides access to a large collection of search tasks assigned in interactive information retrieval studies (ils.unc.edu/searchtasks/search.php) (Wildemuth & Freund, 2012). However, only several studies have attempted to examine the validity of simulated work task situations.

In Borlund and Ingwersen's (1999) meta-evaluation, they compared seven behavioral measures—such as number of searches and search time spent—between simulated tasks and participants' real tasks. The only significant difference was that participants made more relevance assessments based on full-text when working on their own tasks. Behavioral measures such as task time and number of unique search terms per person were not significantly different between groups. The researchers concluded that simulated work task situations were not significantly different from real information needs and that most users found the simulated situations realistic.

Li and Hu (2013) investigated the differences between simulated and authentic tasks in a digital library evaluation and discovered no significant differences in participants' interactive behaviors or performance. By contrast, Blomgren, Vallo, and Byström (2004) discovered that their system (i.e., a full-text database containing news articles) performed better for real tasks than for simulated tasks. Their focus was the effectiveness of the system (e.g., relevance) rather than users' behaviors. In another comparison between simulated and authentic tasks, Borlund, Dreier, and Byström (2012) compared the time spent on search between two independent studies, one lab study and one naturalistic study, that recruited different participants (college students and people employed in different jobs). They observed that participants spent more time on simulated work task situations than their self-formulated information needs, though this finding was drawn from descriptive statistics because their datasets did not permit statistical testing. All of these studies had specific requirements for the types of tasks that participants could bring due to study objective or system restrictions (e.g., specialized database).

Other than observable impact on search behaviors, authenticity of tasks can also influence users' emotions and task perceptions. For example, Poddar and Ruthven (2010) compared the emotions experienced by participants between authentic tasks and simulated tasks. Participants felt more confident and experienced less uncertainty when working on their own tasks. Their confidence was related to ownership of a task as they could change the direction of their personal task. They also had a higher level of topical knowledge in their own tasks than simulated tasks. However, the difference in knowledge may not be due to the task being authentic or simulated, but participants' tendency of bring familiar tasks. Poddar and Ruthven reported that all participants prepared a task similar to the one they had already completed as the authentic task for the study. Similarly, in Li and Hu' (2013) comparison study, participants reported significantly higher pre-task topical familiarity, confidence, and search experience in authentic tasks than simulated tasks. Interestingly, the goal of the simulated task was perceived to be clearer than the authentic task in their study.

2.4.5 Limitations in Existing Research

Existing comparison studies have provided encouraging evidence that the behaviors measured in synthetic settings and tasks may represent users' real search behaviors for the most part. However, previous research has several major limitations that call for further research. First, the field settings used may be too constrained. For instance, Kelly and Gyllstrom (2011) instructed remote participants to finish the study in one uninterrupted session (around an hour) with all other applications on their computer closed. In real-life, users are unlikely to concentrate on searching for an hour. Borlund et al.'s (2012) comparison study also instructed participants to finish tasks in one session with the presence of a researcher. The field setting is considered more natural than the lab setting not only because of the location itself, but the freedom that comes with it. It would be meaningful to see if users' behaviors are different between settings when they are given the flexibility to decide when to work on tasks.

Second, existing studies comparing the lab and field settings often have involved confounding variables. Borlund et al. (2012) compared the time spent on searching between two independent studies, one lab study and one naturalistic study. The two studies might not be directly comparable because they recruited two different samples of participants (university students vs. professionals from various fields) who searched for different types of tasks. Apart from study settings, participants and task characteristics might also affect search behaviors. Similarly, Zuccon et al. (2013) compared university students and crowdsourcing participants. In both studies, participant population may have been a confounding variable since we are not sure if it could affect search behaviors. For example, Hauser and Schwarz (2016) suggest that MTurk workers are more attentive to instructions than university students. Therefore, it is not clear that if the differences between the lab and remote settings could be attributed to the settings or differences in participants.

Third, only a limited number of behavioral variables were measured in existing comparison studies including the number of queries issued, number of pages visited, and task/session time. By contrast, TBIS studies often use a variety of behavioral measurements related to time, queries, and page visits such as different types of dwell time. A finer-grained comparison of user behavior is needed to provide a more complete picture of the influences of study setting and task authenticity on TBIS. Also, existing studies only measured behaviors using whole-session data (behaviors measured using the data from an entire task session). In recent years, within session variables, particularly first-query measures (e.g., first query segment time in each task) that are collected before a search session ends have been proven to be useful for real-time prediction and personalization (e.g., Arguello, 2014; Mitsui, Liu, & Shah, 2018). Using both user study data and TREC data, Mitsui et al. (2018) confirmed that first query measures could be at least as good as whole session measures for the prediction of some task features. For task product prediction, first query measures even outperformed whole session measures. Liu, Mitsui, Belkin, & Shah (2019) confirmed Mitsui et al.'s finding that the effects of task characteristics were stronger at early stages in a task session. Testing the influences of study setting and task authenticity on within-session variables would be a meaningful addition to the existing body of knowledge.

Fourth, most studies comparing authentic and simulated tasks used specialized search systems such as news collections or digital libraries. The type of tasks studied were also restricted to those that could be completed in those specialized systems. Nowadays, many TBIS studies focus on Web searching. This suggests a need for conducting comparison studies on Web search engines as search steps vary significantly with the type of system in use. Borlund (2016) also raised the importance of verifying existing results with newer studies because information searchers are likely to be more experienced today than when the simulated task concept was proposed. Lastly, existing studies only examined the influences of study setting or task authenticity, but it would be useful to explore the interaction effects between these two variables. For example, the impact of study setting may be different in authentic and simulated tasks.

2.5 Preliminary Research

The study conducted for this dissertation was originally inspired by a previous research carried out by my colleagues and me that examined the effects of cognitive authority and peer advice on users' search behaviors in both lab and naturalistic environments (Wang, Liu, Mandal, & Shah, 2018; Liu, Wang, Mandal, & Shah, 2019). This research had two treatment groups in which participants received the study treatment and finished two search tasks and one control group in which nothing was given to the participants and they searched as they would usually do. Although we did not design the study to compare different study settings, we incorporated two field sessions and one lab session to measure users' baseline search behaviors (first field session), behaviors immediately after receiving a treatment (lab session), and short-term lasting effects of the treatment (second field session). As a result, we collected behavioral data both in lab and remote settings, and thus could see some differences in user behavior between settings.

Further details of that study are not reported here since it was done for a very different purpose. What has inspired this study is that some aspects of our users' searching and browsing behaviors significantly varied between study settings. Specifically, in factual specific tasks (i.e., tasks with a specific goal and require users to locate facts), participants visited significantly more content pages and unique content pages per query but spent less time on content pages and SERPs in lab session. These behavioral differences were observed in the control group (no treatment), meaning that these differences represent users' real behavioral differences across different study settings, not the differences caused by study treatments. A possible explanation is that users were more focused in the lab session, so they visited more pages but needed less time on each page. This study has motivated me to have an in-depth inquiry of the effects of different study methods on users' search and participation behaviors.

2.6 Summary

This chapter reviewed the theoretical and empirical studies of TBIS and comparison studies that examined the differences between lab and field settings as well as between authentic and simulated tasks. While different methods have been used to examine task characteristics and their relationships to user behavior, only a limited amount of research was done to examine and compare the methods themselves. Existing results from studies utilizing different methods, particularly different study settings, are not directly comparable due to the different behavioral variables collected and different levels of control.

A few researchers have investigated the differences between study settings and between simulated and authentic tasks. They did not observe significant difference for most of the measures. Main limitations of those studies include limited behaviors measured, restricted field settings, collections adopted, and only examining one aspect of study methods (setting or task). Inspired by existing studies and a preliminary research, this dissertation incorporates both study setting and task authenticity as independent variables in the same study to provide a more comprehensive picture of user behavior collected using different methods. It addresses limitations in existing research with a more rigorous study design and offers a more comprehensive understanding of the effects of synthetic study settings and tasks than what existing works have provided, including a more exhaustive list of behavioral measures examined, interaction effects between study setting and task authenticity, and interview data that explains why users are influenced by the study variables. The detailed study design is reported in Chapter 4. Data Collection.

Chapter 3 : Conceptual Framework

In this chapter, I describe the conceptual framework that has guided this dissertation, which includes a set of theoretical and empirical works. These are: Ingwersen and Järvelin' (2005) cognitive framework for information seeking and retrieval, research that explored laboratory effects in behavioral sciences, Borlund (2003)'s framework for evaluation of interactive information retrieval systems, Li and Belkin' (2008) faceted classification of tasks, and O'Brien and Lebow (2013)'s User Engagement Scale (UES).

3.1 Cognitive Framework of Information Seeking & Retrieval

I use Ingwersen and Järvelin's (2005) cognitive framework of information seeking & retrieval as a starting point for this dissertation. Ingwersen (1996) built on previous works based on the cognitive viewpoint and proposed a cognitive approach to IR theory out of a recognition of the limitations of the Laboratory Model of IR evaluation. The cognitive approach views information seeking and retrieval processes as processes of cognition and place attention on users' cognitive space (e.g., tasks, cognitive & emotional states), their interactions with the system, and the influences of their past experiences as well as social, cultural and organizational context. Ingwersen and Järvelin later extended this framework to integrate information seeking and information retrieval (IS&R). They suggested that IS&R interactions take place in a context and the work or everyday life task situation is the central element of that context. Tasks trigger a problematic situation in which users realize a need of information in order to continue their task at hand. Ingwersen and Järvelin also stressed the concept of perceived work task because the same task may be perceived differently in different situations. The perception of work or everyday life tasks is the central factor that affects IS&R. This framework has guided the design of this study, particularly the pre-task and post-task questionnaires, suggesting a set of variables that can be examined in the study. The questionnaires included questions regarding participants' perceptions of search tasks, their search experiences, as well as the affective aspects such as enjoyment and frustration.

3.2 Study Settings and Laboratory Effects

Although laboratory studies are designed to study individuals' real-world behaviors, they have been criticized for being too artificial to elicit natural behaviors. Researchers in behavioral sciences have examined the potential effects brought by the laboratory setting itself. Orne (1962) questioned the validity of lab experiments, suggesting that employing the experimental model used in physics assumes that the study participant is a passive responder to stimuli. This assumption is hard to justify in behavioral sciences as participants may try to ascertain the real purpose of an experiment and work in a manner that will support the hypotheses being tested. He called the cues that convey an experimental hypothesis the "demand characteristics of the experimental situation", which can significantly influence participants' behaviors in psychological experiments.

Levitt and List (2007) concluded that a fundamental obstacle to extending findings from lab economic experiments to the real world is that the participants know that they are being monitored in a lab experiment. This knowledge can induce them to please the experimenter. Landsberger (1958) and Roethlisberger, Dickson, and Wright (1975) used the term “the Hawthorne effect” to describe the phenomenon in which participants’ performance is temporarily improved when they are observed.

Task-based information seeking studies conducted in different study settings were reviewed in the last chapter. Briefly, user studies examining TBIS can take place in a lab setting where users participate in a computer lab provided by the researcher. User studies can also be conducted in users’ natural environment which can be any places where the users are, such as workplaces, libraries, coffee ships, or their homes. Although lab studies are useful for examining the effects of one or a few task characteristics, they are expensive to run, and their artificial environment may not elicit users’ natural search behavior. Working on tasks in an unfamiliar environment and/or under researchers’ observation may bring laboratory effects that lead participants to deviate from their normal search behavior (Andrzejczak & Liu, 2010; Zuccon et al., 2013). Researchers have suggested conducting more studies in users’ natural environment (Saastamoinen & Järvelin, 2017; He & Yilmaz, 2017). Particularly, remote user studies that can be completed online not only allow participants to work at a place of their choice, but also save expenses in travelling and scheduling. How exactly users’ search behaviors are different between lab and field settings is not clear, which calls for a systematic comparison between lab and field user study settings.

3.3 Simulated Work Task Situations

Borlund (2000, 2003) proposed a framework for the evaluation of IIR systems to facilitate realistic IIR evaluations in a relatively controlled environment. While the framework was originally proposed to guide the design of IIR evaluation, it is also instrumental for the design of studies that examine tasks and user behavior. The framework was proposed to provide a set of components that can ensure a functional, valid, and realistic study setting. Among which, the application of the concept of simulated work task situation is an essential part of the framework and has been utilized by many researchers in IIR evaluation and information seeking studies.

A simulated task situation is a short "cover story" that places users in a scenario that leads them to search for information. The simulated task situation provides study participants with the source of their information need and the environment of the situation, similar to what they would usually have when searching for information in reality. Users could form their own information needs, given the simulated situation. Although the concept of simulated task situation has been applied in many studies, it has rarely been used as Borlund (2000) recommended. Specifically, Borlund suggested employing both simulated task situation and users' real information needs within the same study so that real information needs can serve as the baseline. This requirement has been largely neglected. This has also inspired this dissertation to further compare the differences between simulated and authentic tasks in the web context.

3.4 Task Classification

Li and Belkin (2008) synthesized previous studies and developed a faceted classification to conceptualize tasks of various levels. The classification scheme includes several generic facets (i.e., task source, task doer, time, outcome, process, and goal), common attributes of tasks (i.e., interdependence and objective complexity), and users' perceptions of tasks (i.e., salience, urgency, difficulty, subjective complexity, and knowledge). It has been adopted by a number of IIR studies to define task characteristics that are used to study user behaviors and to perform task predictions (e.g., Jiang et al., 2014; He & Yilmaz, 2017). In this dissertation, this classification was employed to develop questions that guide participants to rate their perceptions of tasks and experiences of working on those tasks.

3.5 Subjective Measurements of User Experience

Cognitive and affective aspects of search such as emotions, interest, and motivation are important parts of users' search experience (Gwizdka & Lopatovska, 2009; Poddar & Ruthven, 2010). Study settings and tasks may have impact on users' perceptions and emotions, which may subsequently influence search behaviors. Examining the impact of study setting and task on subjective factors may help explain the behavioral differences between conditions.

To obtain a comprehensive understanding of the influence of study setting and authenticity of tasks, parts of O'Brien and Lebow's (2013) User Engagement Scale (UES) were included in the post-search questionnaire. This multidimensional scale captures users' perceptions of Perceived Usability (PUs), Aesthetics (AE), Novelty (NO), Felt Involvement (FI), Focused Attention (FA), and Endurability (EN). It has been empirically tested in multiple contexts including searching. While the scale proposed by O'Brien and Lebow has 28 items, some are only relevant to system usability and aesthetics, and thus were excluded from the post-task questionnaire for this dissertation. In addition, a few other subjective factors (i.e., motivation, confidence, and pre-search topical interest) that have been used by existing studies were also measured in the questionnaires.

3.6 Summary

This chapter presents the conceptual framework of this dissertation. Ingwersen and Järvelin’s (2005) cognitive framework of information seeking and retrieval embraces both information objects and the cognitive space of a user and has laid a foundation for this dissertation. Existing behavioral studies conducted in different study settings have suggested potential laboratory effects and how they may affect participation behaviors. Borlund’s (2000) framework for interactive information retrieval (IIR) evaluation proposes key study components for evaluating IIR systems and information seeking behaviors. Li and Belkin’s (2008) task classification and O’Brien and Lebow’s (2013) User Engagement Scale provide a set of task and engagement attributes that may influence search behaviors. The detailed data collection methods are described in the next chapter.

Chapter 4 : Data Collection

This chapter describes the research questions and data collection methods. This includes recruitment procedure, study procedure, task design, data collection instruments, and behavioral measurements.

4.1 Research Questions

The purpose of this dissertation is to examine how study settings and task authenticity impact users' search behaviors, task perceptions, and search experiences in task-based information seeking studies. Specifically, I address the following research questions:

- RQ1: Do study settings (i.e., lab setting versus remote setting) affect users' Web search behavior (e.g., average query segment time and page visits per task)?
- RQ2: Does the authenticity of study tasks (i.e., authentic tasks versus simulated tasks) affect users' Web search behavior?
- RQ3: Do users' perceptions and experiences of tasks (e.g., task difficulty, engagement) differ in different study settings and tasks?

4.2 Research Design

This section reports on the data collection procedure and instruments.

4.2.1 Overview

For this dissertation, I conducted a 2x2 repeated design experimental study with two independent variables: study setting and task authenticity. Study setting had two conditions: laboratory setting and remote setting. All participants finished two sessions:

one lab session in which they came to an individual computer lab and one remote session in which they finished the tasks at any places of their choices. Task authenticity also had two conditions: authentic task and simulated task. In each session, participants finished one simulated task assigned by me and one authentic task brought by themselves.

As I have discussed in the literature review, methods used in field studies could vary greatly from study to study. Remote studies as a type of field study are conducted completely remotely online. Participants receive all study materials and finish the study online without meeting the researcher in person. I chose to set up a remote setting against other types of field studies because the remote setting closely resembles users' natural search environment (e.g., no appearance of a researcher). Also, remote setting has been increasingly adopted by TBIS researchers due to its flexibility in time and location. Data was collected from September 2018 to February 2019. The study recruitment and procedure are detailed as follows.

4.2.2 Participants and Recruitment

Recruitment started in September 2018. I posted recruitment messages on various Facebook groups organized by Rutgers University students (e.g., Rutgers Free & For Sale, Rutgers Housing Sublets & Roommates). I also posted flyers on campus at various academic buildings, dining halls, residential halls, and student centers. To be eligible, participants must be 1) at least 18 years old, 2) Rutgers' undergraduate students, and 3) regular Chrome browser users (as this study required participants to use a Chrome browser and switching browsers may affect search behavior).

Participants followed a link in the recruitment message to sign up for the study. On the registration form, they provided basic demographic information (e.g., age, gender, major, year in college) and signed an informed consent electronically. To verify that they were Rutgers University students, I asked them to provide a Rutgers email address. I reviewed their information, verified their undergraduate student status using their email address, and contacted qualified participants via email to provide further instructions. The recruitment message and registration form are provided in Appendix A & B.

A power analysis for linear models was conducted prior to recruitment using the *pwr* package in R. It suggested that 132 observations (in this research, tasks) would be a sufficient sample size to detect any effects of the study treatments. Given a power of 0.90, an alpha of 0.5, and a medium effect of 0.15 (linear models) (Cohen, 1988). Since each participant performed four tasks, at least 33 participants should be recruited. Thirty-nine undergraduate students from Rutgers University participated in the study and the data collected from 36 participants were complete. Two participants withdrew before finishing the entire study. One participant forgot to turn on the extension while working on the remote session tasks. These three participants were excluded from data analysis. The final dataset contained 144 tasks completed by 36 participants.

To reduce leading effects, participants were not informed about the real purpose of this study. Instead, they were told that I was interested in their search behaviors and their participation would assist in the design of better search systems. Upon request, the real objective of the study was communicated to them at the end of their participation. They were also likely to figure out the real study purpose after the follow-up interview as I asked about the influence of study setting and task authenticity. Regardless, they were asked not to share study details with other students in case that their friends or classmates were also enrolled.

4.2.3 Study Procedure

4.2.3.1 Eliciting Authentic Tasks

After signing up for the study, participants received an email instruction detailed the study procedure as well as a unique username and password combination to log into the study system. Prior to starting the main study, each participant was asked to describe two search tasks that they planned to conduct in the near future but had not done so. These two tasks were later used as the authentic tasks during the main study.

Although several existing studies required participants to bring their own tasks/information needs, some did not document this process in detail or if there were any restrictions on task type or topic. Li and Hu (2013) included the full description of eliciting the real work task situation, which instructed participants to select a task they needed to finish recently, such as a paper, an assignment, or a research project. This description placed the task into a schoolwork context. Borlund and Dreier (2014) instructed participants to formulated three information needs that corresponded to Ingwersen's (2000) three types of information needs: verificative information need, conscious topical information need, and muddled topical information need. The three information needs can be related to any everyday topics such as job, school, or leisure. Overall, to make the tasks comparable among each other, there should be some restrictions on the types of tasks that participants could bring. In the meantime, the restrictions should not be too strict to make authentic tasks artificial. Although the concept of search task could be relatively easy for undergraduate participants to understand in a school or work context, I did not limit the topics of their tasks to school or work in order to keep their task choices flexible. Including everyday context widened the scope of the topics and set less restrictions, and thus could increase the chance of eliciting real tasks.

In order to make the authentic tasks comparable among each other and to the simulated tasks, I required the participants to bring evaluative tasks, which are tasks that require people to find information to compare and evaluate several options that belong to the same category and make a recommendation from those options (Kelly et al., 2015). For example, if a person wants to take an online programming course, this person may need to consider multiple aspects such as course content, price, instructor, and course duration. This person evaluates several courses along multiple dimensions and makes a selection. This type of task also falls into Li and Belkin's (2008) decision or solution task.

I chose to use evaluative tasks mainly for three reasons. First, this restriction can to some extent ensure that the tasks did not vary significantly from one another. Kelly et al., (2015) developed a cognitive complexity framework for task creation and evaluation and divided tasks into five categories with increasing levels of cognitive complexity: remember, understand, analyze, evaluate, and create. They discovered a general trend that values of behavioral variables tended to increase as cognitive complexity increased. Using only evaluative tasks allows me to keep all the tasks within the same cognitive complexity level since they are more commensurable with each other than with tasks from the other levels. Second, since participants were asked to bring their own tasks, the type of task needed to be common in their everyday lives so they could think of two tasks that they would do even without being requested by this study. It is common in everyday life that people make a selection among a few options (Wang & Shah, 2017). Third, evaluative tasks were found to be interesting to participants in Kelly et al.' research, and thus may motivate them to invest more efforts.

To avoid influencing participants' choices of authentic tasks, I did not provide any examples of tasks that they could bring. For instance, if I had included a shopping task as an example, participants may have only considered shopping-related tasks or misinterpreted that as my expectation. I also wanted to make sure that the tasks were not too complicated or simple compared to each other. For example, picking a fast food restaurant for a quick lunch may be too simple compared to picking a doctor for a health condition. Thus, I provided one example of a task not suitable at each extreme (i.e., too complex or too simple).

To ensure that participants had a correct understanding of what an evaluative task was, I made the task preparation process semi-interactive. I reviewed each participant's authentic tasks prior to the main study. Apart from describing the task itself, participants were also asked to provide more details about their tasks in the task submission form, such as why they needed to do those tasks. Even though some task descriptions were quite simple and looked like search queries, I was able to learn more about the tasks from their narratives of the context. I determined if their tasks fulfilled the study requirement (e.g., selecting an option from multiple options, not too easy or difficult, the task goal can be fulfilled by searching online). Participants would not be given access to the next stage in the study system until their tasks were approved. If a task did not fulfill the requirement (e.g., not an evaluative task, too simple or complex), I emailed the participant to request anther task. The tasks submitted by most participants were approved at the first attempt. A few participants submitted wrong types of tasks and needed additional guidance, mostly because they did not read through the instruction or did not fully understand it. Tasks that were too simple were also rejected even though they belonged to the category of evaluative task. For example, one participant wanted to find a session of a class that worked with her schedule. This task only involved comparing the time and location of several available sessions of the same class. It was not only very simple but would also only involve activities on one website, a course registration site. Tasks like this were rejected and the participants were asked to resubmit. The instruction for preparing authentic tasks is provided in Appendix C.

4.2.3.2 Lab Session

In the Lab Session, each participant individually came to a computer lab to finish two search tasks on a desktop computer. In the study system, a page containing a task description and a text box was shown to them. They were asked to search online and to type or copy and paste any information or links they found in a text box. There was no requirement regarding how much information they needed to find. The study system itself did not provide a search interface, and participants could use any search engine of their choice. They could revisit the task description at any time, which could be especially useful in simulated tasks since those were not their own tasks.

Participants were advised to spend as much time as they needed for each task. However, they were informed that each lab session was scheduled for about an hour, which might be interpreted as a soft time constraint on them. They were encouraged to search as they would normally do. They could decide when to stop searching if they had found enough information or decided to give up. According to existing studies, users rarely spent more than 20 minutes on a search task in a lab study unless other assignments were given (e.g., writing a short essay) (e.g., Zhang & Gwizdka, 2014; Liu & Belkin, 2010). It was also true in this study as the large majority of the participants finished within an hour. There was always a half-hour gap between two lab sessions. Therefore, even when participants did not finish within an hour or they arrived late, they were not interrupted and given extra time to finish. All participants completed the lab session without being cut early.

Participants' querying and dwelling behaviors were collected by a Chrome browser extension. Their screen was monitored by me using Morae on another computer during the lab session, just in case they run into any issues. Participants were informed that their screen was monitored in the Lab Session, but not in the Remote Session. The lab setting may have stronger influences on their behaviors when they are being monitored. However, this design of the Lab Session resembles a typical lab study in which users' screens are often observed in some way. More details about the extension are reported in 4.2.5.1 Chrome Browser Extension.

4.2.3.3 Remote Session

Participants went through the same procedure in the Remote Session as in the Lab Session except that they participated remotely at any locations of their choices. Another difference between the lab and the remote sessions was the time restriction. Unlike the lab session, where a time window had to be reserved in advance (even though no time limit was given), the remote session gave participants more freedom to choose when to work on the study. They were given up to three days to finish the questionnaires and tasks remotely. They could approach those at any time within those three days. If they temporarily left the tasks and logged out, they would be taken back to their previous stopping point after they logged in again. Their search data was only collected when they were logged in. This was to provide them a natural and flexible search environment so they would not feel a strict time pressure. I gave the 3-day time limit not because participants would need three days, but because that they were more likely to delay or forget about the study if no time limit was given at all. Unlike researchers who explicitly manipulated time constraints, such as Liu et al. (2019), who limited the time available to 5 minutes per task, I gave a generous 3-day limit which served more as a task reminder than a time constraint.

It is worth noting that although participants may not be under strict time pressure in the remote session, their own tasks could have other deadlines closer than the one I gave. For instance, P13 was comparing different scholarships to apply for and she needed to decide before the deadline for a scholarship application which was earlier than three days.

Participants received an email reminder one day prior to the end of their remote session if they had not finished all tasks and questionnaires. I checked their log and questionnaire data after they finished the remote session to ensure that their data was successfully recorded. The order of their sessions was randomly determined, and they did not know if they would start in the lab or in the field until they were approved to start the main study.

4.2.3.4 Questionnaires

Participants filled out two questionnaires for each task (i.e., pre-task questionnaire and post-task questionnaire) to report their perceptions of tasks and search experiences. These questionnaires were used to reveal if study setting or task authenticity had influences on their perceived task characteristics and experiences that would not be exposed by implicit behavioral data.

Before each task, participants gauged their expectations of task complexity, task difficulty, topic knowledge, and rated how important the task was, how interested they were in the task, how motivated they were to complete the task, and how confident they were in finishing the task on a five-point scale. The first four task characteristics were selected from Li and Belkin's (2008) faceted task classification. The literature has also suggested that these factors affect users' search behaviors and intentions (e.g., Liu et al., 2019; Saastamoinen & Järvelin, 2017; Zhang & Gwizdka, 2014). In addition, other factors that may affect user engagement were selected from the literature (O'Brien and Toms, 2008; Borlund and Schneider, 2010). Task difficulty and complexity were asked again in the post-task questionnaire except that participants responded to those questions based on their actual experience. Not all task characteristics that have been reported in the literature were included in the questionnaire because some of them should have constant value across all tasks (e.g., task product, task doer).

The post-task questionnaire included more items about participants' perceptions and emotions, adopted from O'Brien and Lebow's (2013) User Engagement Scale. Some items in the scale were removed because they were not relevant to this study (e.g., screen layout). In addition, the post-task questionnaire provided a list of search problems selected from Chowdhury, Gibb, and Landoni's (2014) list of information seeking problems to allow participants to select any problems they encountered during the search session. This is to check if users encounter different problems in different settings or tasks. Among the items collected by pre-task questionnaires, task complexity and task difficulty were asked again in the post-task questionnaires because participants may have a more accurate estimation of these two factors. Other items such as topic knowledge and task motivation were not asked again because the pre-task responses should be the ones that influenced participants' behaviors and post-task perceptions may only influence future tasks. The pre-task and post-task questionnaires are provided in Appendix D & E.

4.2.3.5 Exit Interview

While behavioral measures address the "what" questions, they do not explain why things occur. It is equally important to investigate how users thought about the influences of the study variables, which may help explain why their behaviors were (or were not) different between conditions. Each participant was interviewed after they finished and submitted all tasks and questionnaires in the study system. Those participants who finished the study with the Lab Session were interviewed at the end of the Lab Session. Those who finished the study with the Remote Session scheduled another time to come to the interaction lab after they finished the Remote Session. This was not an extra trip for them since they also needed to pick up their compensation in cash. Exit interview was scheduled as soon as their schedule permitted after finishing the second session (mostly within the same week) so they still had fresh memory of their experiences.

The interviews were semi-structured. I had a general interview guide including a set of prepared questions/themes, but I also tailored the questions based on my observation and participants' questionnaire responses. I reviewed their log data and questionnaire responses before each interview and prepared questions in advance. During the Lab Session, I monitored their activities on another computer using Morae and wrote down notes and questions. Questions asked in the interview were primarily about the differences between two types of tasks and settings. For example, if a participant rated the authentic tasks much more difficult than the simulated tasks, he/she was asked to elaborate on that. I also asked them to compare two study settings and tasks as they had experienced both.

The interview was also an opportunity for me to ask about anything interesting, unclear, or irregular in their search logs. For example, P8 was comparing different dairy-free recipes in the remote session but did not use the term "dairy-free" in her search queries at all. Since she was working on this task remotely and also searched for other things in between, it was not clear exactly which queries and pages were relevant to this task. I was able to sort out relevant data from the log with her help in the interview. It turned out that a few pages looked irrelevant to me were actually relevant to her task. The interview guide is in Appendix F.

4.2.4 Simulated Tasks

Each participant searched for information to finish four tasks, two in the laboratory, and two in the remote setting. All tasks were evaluative tasks as defined by Kelly et al. (2015) (i.e., tasks that require people to find information to compare and evaluate several options that belong to the same category and make a selection based on this information). Two of the four tasks were authentic tasks submitted by them in advance. The detailed procedure of eliciting authentic tasks is provided in 4.2.3.1 Eliciting Authentic Tasks and Appendix C. The other two tasks were simulated tasks selected and modified from Kelly et al., (2015); both were evaluative tasks placed in a scenario. Rather than designing new tasks, I chose two tasks that were designed and had been tested by other researchers in multiple previous studies including a study with the same type of users as the current work (i.e., undergraduate students). The two tasks were presented as follows:

1. Simulated Task One: One of your siblings got a spur of the moment tattoo. However, after some careful consideration, they now regret it. You decide to investigate methods for tattoo removal, so you can make some suggestions to your sibling about what he might do to get rid of the tattoo. What are the current available methods for tattoo removal, and how effective are they? Which method do you think is the best? Why?
2. Simulated Task Two: For several years, your friend has complained of periods of extreme fatigue, headaches, and joint pain. After seeing several doctors, a specialist diagnosed her with lupus. What are different ways to treat lupus, and

how effective are they? Which treatment would you recommend to your friend? Why?

4.2.4.1 Task and Session Rotation

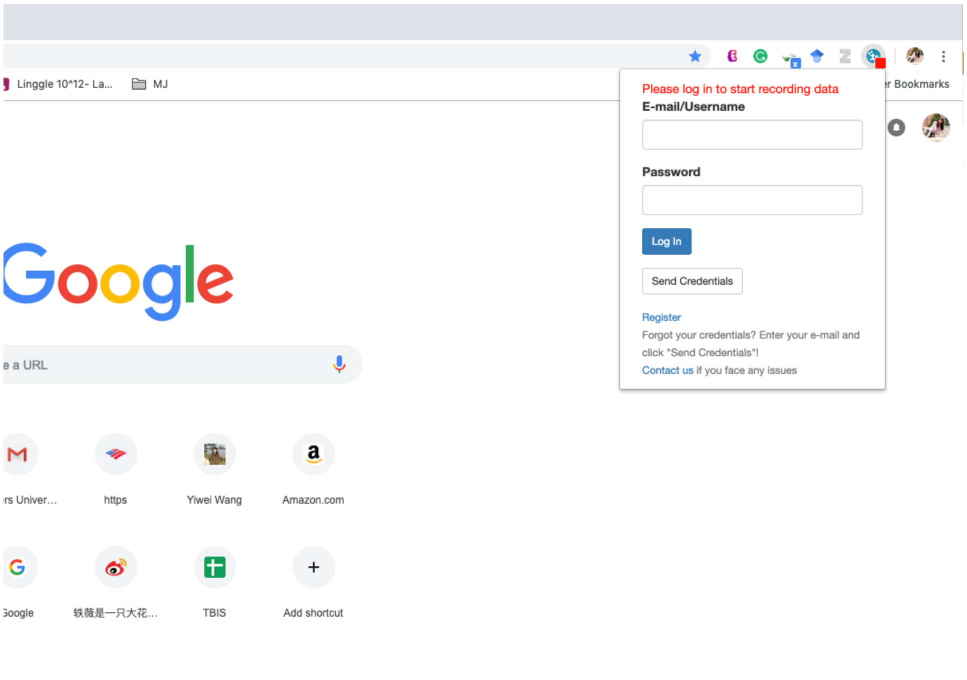
Existing studies have demonstrated that the order of tasks or study sessions may affect users' participation behavior. For instance, their engagement may gradually drop during a study so they may not invest as much effort later in the study as when they start (Liu et al., 2019). To avoid the ordering effect, tasks and study settings were rotated. Half of the participants started the study with the Lab Session, followed by the Remote Session. The other half finished the Remote Session first. The order of their sessions was randomly determined after registration and communicated to participants through email before the main study. Initially, they only had access to the first session tasks. This was to prevent those participants who started remotely from accidentally finishing all tasks. Task order was also counterbalanced. Half of the participants started with the authentic task in each session while the other half started with the simulated task. Table 4.1 lays out the task and session rotation.

	Lab – Remote	Remote – Lab
Authentic – Simulated	9 participants	9 participants
Simulated – Authentic	9 participants	9 participants

Table 4.1: Session & Task Rotation

4.2.5 Chrome Browser Extension

A Coagmento Chrome browser extension (Soltani, Mitsui, & Shah, 2019) was developed to display study instructions, tasks, and questionnaires, and to collect participants' online activities. Participants were provided with a link to the Chrome App Store to download the browser extension and a unique username and password combination to log into the extension. Once they installed the extension, a Coagmento icon appeared on the top right corner of their Chrome browser. The extension was turned off by default and a red square showed on the icon to indicate the off mode. After a user clicked on the icon and enter his/her username and password, the red square turned to green and started to collect log data (see Figure 4.2 & 4.3). They needed my permission to continue the study 1) after they submitted their authentic tasks and 2) after they finished the two tasks for the first session. This was to ensure that they submitted two tasks suitable for this study and they did not accidentally finish all four tasks in one sitting.



Please log in to start recording data

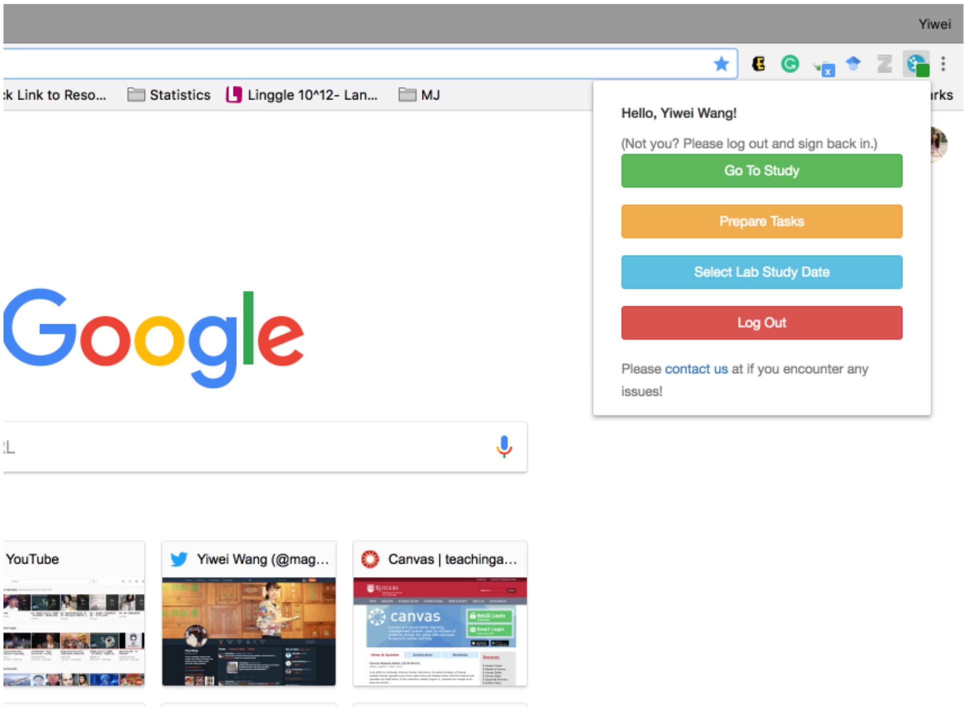
E-mail/Username

Password

Forgot your credentials? Enter your e-mail and click "Send Credentials"!

Contact us if you face any issues

Figure 4.1: Chrome Browser Extension Before Login



Hello, Yiwei Wang!

(Not you? Please log out and sign back in.)

Please contact us at if you encounter any issues!

Figure 4.2: Chrome Browser Extension After Login

The extension itself did not have a search interface and users' search data were automatically captured on any search engines, such as Google and Bing once they logged in. The extension collected their online activities including the queries issued, search engine result pages (SERPs), and content pages viewed with timestamps. All data was automatically stored in a secure server for analysis. Timestamps on the client side were used to extract various behavioral variables, such as different types of dwell time. To protect participants' privacy, the extension allowed them to turn it off at any time by one click if they did not want their search data to be collected. The extension collected data about searching activities as follows:

- user ID
- timestamp (server side)
- local timestamp (client side)

- task ID
- query
- visited URL (Uniform Resource Locator)
- title of Web page
- host of Web page
- search engine result page (SERP)
- date and time

4.2.6 Behavioral Measurements

To examine which aspects of users' searching and browsing behaviors were affected by the study variables, I extracted the following measures of users' behavior (abbreviations in parentheses). These behavioral measures were selected after a comprehensive review of existing TBIS studies that collected online user behaviors. Researchers have used these measures to examine the relationships between task characteristics and user behavior as well as to run predictions. It is worth noting that the page dwell time extracted from search log was the time that a page was opened and shown to a participant. I was not able to determine whether a participant actually focused on a page since I did not record their physical activities. For example, it was possible that a person opened a page but left the computer to do something offline (e.g., chatting with a friend). In that case, the time spent on the offline activity was still counted as part of the page dwell time.

Whole-session measures:

- *Task completion time (task time)*: Total time spent on the entire task session. This was calculated by adding up the time spent on all online activities related to a task;
- *Task searching and browsing time (searching & browsing time)*: Task completion time minus the time spent on questionnaires and the task description page;
- *Total page time*: Total time spent on content pages in a task;
- *Average query segment time (avg. query time)*: Average time spent on query segments in a task. A query segment is everything that occurs from one query to the next;
- *Average unique query segment time (avg. uniq. query time)*: A unique query segment time was calculated by adding up the time spent on identical queries in a task session;
- *Number of queries per task (no. queries);*
- *Number of unique queries per task (no. uniq. queries);*
- *Average query length (avg. query len.)*: Average query length (calculated by words) in a task;
- *Number of queries without page views (no. queries w/o page):* The number of query segments in which no content page was clicked;
- *Number of unique SERPs viewed per task (no. uniq. SERPs)*: The number of unique search engine result pages visited in a task (a user may visit more than one SERPs in each query segment and identical SERPs were only counted once);
- *Average first decision time per task (avg. 1st. deci. time)*: The average time taken from issuing each query to clicking on the first page in this query segment in a task;
- *Number of content pages per task (no. pages)*: The number of content pages visited each task. If a content page was visited multiple times during a task session, it was counted as multiple pages;
- *Number of unique content pages per task (no. uniq. pages)*: Identical content pages were only counted once in a task session;
- *The proportion of time on content pages in a task (prop. page task);*

- *The proportion of time on content pages in searching and browsing time of a task (prop. page task w/o task page)*;
- *Average number of times a page is visited during a task session (avg. page visits);*
- *Average content page dwell time per task (avg. page time):* The average time spent on a content page;
- *Average unique content page dwell time (avg. uniq. page time).*

First-query measures:

- *First query segment time (1st query time);*
- *Number of pages visited in the first query segment (1st no. pages);*
- *Number of unique pages visited in the first query segment (1st num. uniq. pages);*
- *Total page time in the first query segment (1st total page time)*: Total time spent on content pages in the first query segment;
- *Average page time in the first query segment (1st avg. page time);*
- *First decision time (1st deci. time)*: The time from issuing the first query to clicking on the first page.

4.2.7 Research Ethics

The browser extension only captured queries, URLs, and timestamps, and it did not collect private information such as username and password that participants used to log into their personal accounts. They could turn off the browser extension at any time by one click if they did not want to reveal their search activities. This was to protect participants' privacy since many of them worked on their own tasks irrelevant to the study during the remote session. They could also request to remove sensitive data from the search log if they forgot to turn off the extension while working on private topics. This study has been approved by Rutgers Institutional Review Board (IRB) (#Pro2018000197). All data are stored in a password-protected server.

4.3 Pilot Tests

I conducted two rounds of pilot tests in July and August 2018 to test the study instruments and procedure. The first round of pilot tests was conducted before developing the Chrome extension used to display the study tasks and questionnaires. Instead, I used a Qualtrics form to display the study tasks and questionnaires. The purpose of this round was to test if the task descriptions, study instructions, and questionnaires were understandable to the participants. I recruited six participants (undergraduate and master's students from Rutgers University) using convenience sampling. Tasks and questionnaires were clear to the participants for the most part. They pointed out issues that could be fixed by providing clearer instruction. For example, one pilot test participant was not sure if the two authentic tasks could be related to each other. I made modifications to the task descriptions after theses pilot tests.

After revising the study instruments, I worked with a developer to develop the Chrome browser extension. The second round of pilot tests was conducted after the extension was fully developed. The primary purpose of this round was to test the entire study procedure from registration to interviewing. I also tested if the browser extension was functioning well and collecting the right types of data. I followed the exact procedure (including recruitment and registration) as what I planned to do for the actual study. A recruitment message was posted on Free & For Sale, a Facebook group organized by Rutgers University students. Subscribers were from various majors and grades from Rutgers. Participants were required to use their Rutgers email for registration, and they were not officially enrolled in the study until I reviewed their registration form and verified that they were undergraduate students in Rutgers University. Posting on one Facebook group page was sufficient for the pilot tests.

Six participants signed up for the study. Four participants from four different majors logged into the extension and started the study, and three of them finished the entire study. The other two participants did not start the study after signing up. One participant submitted two search tasks but did not continue the study.

After each participant completed the study, they were asked to provide feedback regarding the study instruments, procedure, and compensation. Overall, participants found no difficulties in following the study instruction to finish the entire study. One participant was uncertain if she should start searching immediately after reporting her authentic tasks. I clarified on the task preparation page that they would be given time to work on the two tasks in the main study. Two participants finished their lab session within an hour while the other used 70 minutes. It took participants seven days in average to complete the entire study. All participants felt that the registration form accurately and clearly described the study. Two mentioned that the study took a little less time than they expected. All participants felt that the compensation was fair in respect to the amount of time and efforts required.

I also verified that the extension was functioning well and was able to produce the right types of data to answer the research questions (e.g., local timestamps, actions). Although the data collected from pilot tests did not permit any in-depth analysis, a few interesting observations emerged. Two out of the three participants who finished the study brought up in their exit interviews that the tasks took longer when they worked at their own places compared to when working in our lab because they were distracted by other things. For example, one participant mentioned that "*I was a little bit more distracted at home, definitely. I don't exactly remember, but I wouldn't be surprised if I check Facebook real quick or something of that sort.*" This was reflected in her search log that she checked social networking sites while working on the tasks. Participants pointed out that they had more control over their own tasks. They had a better idea of when to stop searching for their own tasks because they had specific goals. By contrast, the simulated tasks were not something they would do in real life so that they would not spend as much time as if they needed to do them in reality. Notably, they did not necessarily feel more interested in their own tasks than the simulated tasks. Two participants showed great interest in the simulated tasks as they have never searched for those topics before.

These two rounds of pilot tests confirmed that the study instruments and procedure were ready for the main study.

4.4 Summary

This chapter detailed the data collection methods employed for this dissertation. I took a mixed-methods approach incorporating lab and remote experiments, questionnaires, and semi-structured interviews. I conducted a 2x2 repeated measure experiment with two independent variables (i.e., study setting and task authenticity) that had two conditions. A Chrome browser extension was developed to display the study tasks and questionnaires. Thirty-six undergraduate students in Rutgers University completed the study. Each participant completed a lab session and a remote session in which they completed one authentic task and one simulated task, as well as an exit interview. The order of study settings and tasks were rotated to avoid an ordering effect. Log data, questionnaire data, and interview data were collected for analysis. Next chapter will report on the data processing and analysis procedure.

Chapter 5 : Data Analysis

This chapter describes the data analysis methods including methods for data pre-processing, quantitative analysis, and analysis of the interview data.

5.1 Data Pre-processing

To prepare the log data for analysis, I first extracted different types of page dwell time (e.g., content page dwell time, SERP dwell time) by subtracting the local timestamp of each page from the local timestamp of the next page within the same task. This was to calculate the time participants spent on each page regardless of whether the page was relevant to their tasks. Next, I reviewed and cleaned the log data by removing irrelevant activities. A separate dataset was created with any queries and pages irrelevant to the study removed. Irrelevant activities were generated because 1) participants were distracted and searched for their own topics while working on the study, or 2) they forgot to turn off the extension after finishing a task. I only included activities relevant to the study tasks when investigating the influence of study treatments on search behaviors such as query time and number of content page visited. For example, when comparing the time spent on each task in two settings, only the time that participants actually spent on the task was included. Time spent on other pages (e.g., SNS, watching TV shows irrelevant to the task) were not counted. Although I could use keywords to automatically identify the majority of those actions (e.g., listening to music on YouTube, checking Facebook), I cleaned the data manually because the same source might be relevant or irrelevant to a task. For example, a participant might watch videos about methods for tattoo removal on YouTube for the study task or might watch American Idol on YouTube which was irrelevant to the study.

Occasionally, it was not clear that whether an activity was relevant to the study. I reviewed the remote session data before each exit interview, so I was able to check with the participants if I was uncertain. For instance, P8 was searching for dairy-free recipes to suit new diet while most of her queries did not have the term “dairy-free”. In cases like this, I usually needed to ask the participant to go through their search log and sort out relevant activities. Also, the browser extension only captured participants’ search queries on general-purpose search engines like Google and Bing. They occasionally searched on websites where queries were not automatically captured such as shopping websites or travel airfare search sites, particularly for their own tasks. Queries issued on those sites were manually extracted when possible.

I extracted other behavioral measures such as query segment time and the number of pages and queries using the cleaned dataset. The definition of each behavioral measure and how it was calculated were reported in 4.2.6 Behavioral Measurements.

5.2 Data Analysis

This section describes the data analysis methods used for this dissertation. All quantitative analyses were carried out in R. Alpha was set to 0.05 for all analyses.

5.2.1 Two-way Repeated-measures ANOVA

To examine the effects of study settings, task authenticity, and their interactions on users’ subjective ratings of task characteristics and engagement, I conducted two-way repeated-measures ANOVA. The data collected for this dissertation are multilevel data with repeated measurements nested within individual participants (measurements were made on each participant at multiple time points). Regular ANOVA assumes that observations are independent. This assumption was violated by the repeated-measures design of this study as the measures taken on the same participant may be more correlated with one another than they are with the measures taken on other participants. Therefore, repeated-measures ANOVA is suitable for data analysis. One model was estimated for each subjective task characteristic collected in the pre-task questionnaire and each item related to engagement collected in the post-task questionnaire. Study setting, task authenticity, and their interaction were added as independent variables.

5.2.2 Multilevel Modeling (MLM)

I tested the effects of study settings, tasks, and their interaction on searching and browsing behaviors using multilevel modeling (MLM) which is a general linear model for repeated measures. While MLM is usually used for nested data structures, it is also suitable for data that are measured at multiple locations or time points on the same participants (repeated measures) (Finch, Bolin, & Kelley, 2014). The traditional regression assumes that values of outcome variables are independent, which was violated by the repeated-measures design. When measures are clustered under individuals, there may be a separate intercept and slope for each cluster. In the context of this study, the effects of study setting and task authenticity on outcome variables may vary based on individual participants. MLM handles dependent data by accounting for the variability of individual participants (Field et al., 2012).

I chose MLM over other statistical methods also because that MLM could include multiple predictors including treatment conditions and other task-related factors-such as subjective task interest and topic knowledge-that may affect search behavior. Although all tasks used in this study were evaluative tasks, other task facets could not be strictly controlled. Task characteristics are quite subjective, and participants reported their perceptions of task characteristics in questionnaires. It was almost impossible to ensure that all tasks were at the same level of difficulty and complexity to the participants or they had the same amount of interest or topic knowledge for each task. For example, participants might make their authentic tasks those that they were most interested in, and subsequently demonstrated different behaviors because of their interest, not because the tasks were authentic or simulated. Therefore, including those potential confounding variables when estimating models can isolate the effects of the experimental manipulations.

One random intercept model and one random slope model were estimated for each behavioral measure. The one with a lower Akaike information criterion (AIC) value (indicating a lower information loss) is reported in this dissertation. Study setting, task authenticity, and their interaction were added as independent variables (IVs) in each model. Task related factors suggested by existing research were added as IVs to control their effects including task difficulty, task complexity, topic knowledge, topic interest, task importance, task motivation, and task confidence. Each model also accounted for the random effects of study setting and task authenticity. A correlation matrix was generated and none of the correlations between any two IVs were more than 0.8. When modeling the whole-session behavioral variables, experienced task difficulty and complexity collected in the post-task questionnaire were used because they were more accurate representations of the actual task difficulty and complexity that participants experienced in task sessions. When modeling the first-query variables, both pre-task ratings and post-task ratings were tried since pre-task responses reflected participants' perceptions of the difficulty and complexity early in each task session. The results were similar in terms of the influences of study setting and task. Only models that included post-task measures were reported for simplicity. All models were computed in R using maximum likelihood estimation.

MLM has all of the assumptions for a regular regression except the assumptions of independence and independent errors. Diagnostic tests were performed to test if each model meets the assumptions. To assess the assumption of no multicollinearity, the variance inflation factor (VIF) for each independent variable was calculated. If the average VIF is substantially greater than 1 or the largest VIF is greater than 10, the model may be biased (Bowerman & O'Connell, 1990). To test the assumption of homoscedasticity, a variation of Levene's Test was used (Faraway, 2005). The assumption is met if the p-value is greater than 0.5. The assumption of normally distributed residuals was tested by checking the histogram of the residuals. If the histogram resembles a normal distribution, the assumption of normality was fulfilled, though a perfect normal distribution is very rare. The assumption of linearity was tested by plotting the residuals against the predicted values. The assumption of linearity is met if the residuals randomly spread out around x-axis (do not show a pattern) (Hox, 2010). All assumptions were met except the assumption of normality. In the case when the residuals were not normally distributed, all values of that particular variable were transformed by natural logarithm which improves the normality of the residuals. The log transformation does not change the relationships between variables and has been commonly adopted by existing studies (e.g., Jiang, He, Kelly, & Allan, 2017; Liu, Gwizdka, Liu, Xu, & Belkin, 2010).

5.3 Analysis of the Interview Data

I conducted an analysis of the interview data to understand participants' experience of the influences of study setting and task authenticity from their own perspectives. The interview data was used as evidence for explanation and interpretation of the statistical results. Participants' recount of their experiences helped me understand why their task perceptions or behaviors were influenced by the study manipulations. Analysis of the interview data was primarily guided by an overarching theme: why and how do study settings or authenticity of tasks affect users' search experiences?

In exit interviews, participants responded to questions regarding their perceptions of the differences between different conditions and problems encountered during each search session. All interviews were transcribed verbatim. When reviewing the transcripts, I sorted out lines of transcripts that were relevant to the overarching theme. For example, when discussing problems encountered during a task session, participants brought up that price was a crucial factor when working on their own tasks and information related to prices was difficult to locate. However, it was not an issue in simulated tasks since they did not need the information for real. This type of discussion was relevant to the comparison between authentic and simulated tasks and thus was included in the analysis. Even though the large majority of the interview questions were designed to address the overarching theme, they did not necessarily solicit relevant responses. For example, if a participant reported that she was unconfident about finding information for a task because of a particular topic, and she would have had the same issue regardless of study settings or task authenticity, her discussion related to that issue would be counted as irrelevant. Lines of relevant data were categorized based on their nature relations. For example, multiple participants were commenting on feeling a time pressure in the lab, and lines related to "felt a time pressure in the lab" were grouped together. Categories were further grouped into subthemes and themes. The result of this analysis is a set of categorized quotations that assisted me in understanding the influences of study settings and task authenticity revealed by the quantitative analysis.

5.4 Summary

This chapter demonstrates the data analysis methods. For quantitative analysis, I performed two-way repeated-measure ANOVA to examine the impact of study settings and task authenticity on participants' subjective ratings of task characteristics and engagement. To demonstrate the effects of study setting, task authenticity, and their interaction on behavioral measures, I estimated multilevel models. In addition, I analyzed the interview data to understand the influence of the study variables from participants' own perspectives.

Chapter 6 : Results

This chapter presents data analysis results which include basic demographic information about the participants, an overview of the authentic tasks submitted by participants, and the quantitative analysis results. Analysis of the interview data is reported in the next chapter to further explain the quantitative results.

6.1 Participants

Thirty-six undergraduate participants represented all four grade levels and 21 different majors such as engineering, neuroscience, business, and information technology (See Figure 6.1 for the distribution of their field of study). Majors specified in the chart had two or more participants. Their average age was 20.6 (SD=2.05), ranging from 18 to 23. 69% were females and 69% were juniors or seniors. 17% were non-native English speakers who were fluent in English.

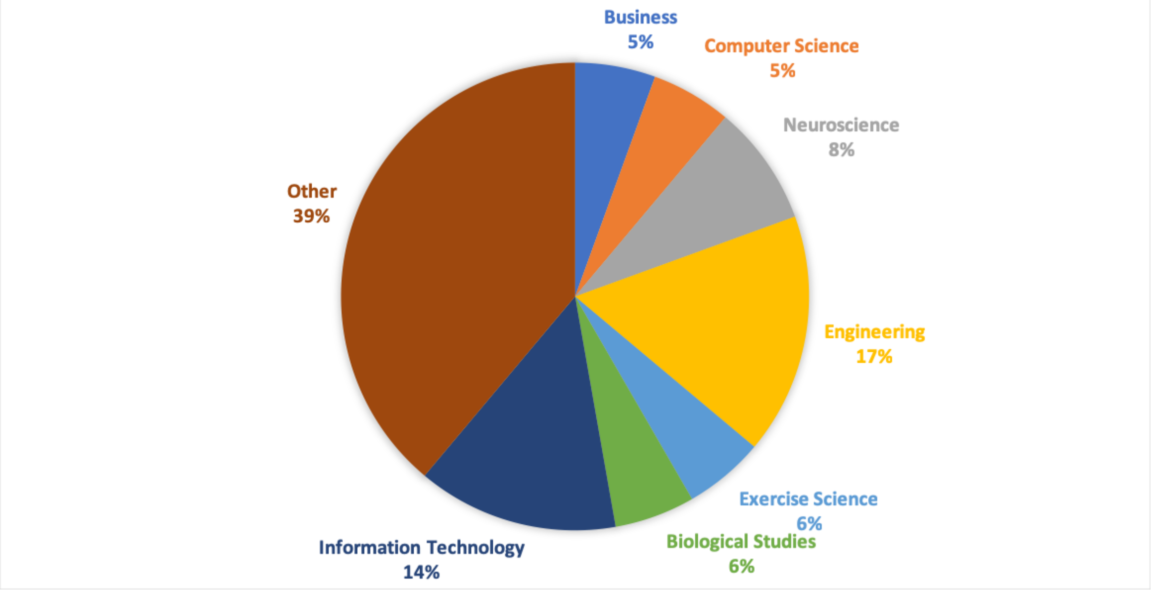


Figure 6.1: Participants' Fields of Study

Participation of the entire study took each person about 1.5-2.5 hours spanning one to two weeks depending on their schedule and the availability of the interaction lab. On average, each participant took 10 days to finish the study. Each of them was paid $40 in cash after their follow-up interviews. The compensation was higher than the minimum wage for student workers in Rutgers University ($11/hour), and thus was considered reasonable to compensate their efforts and time.

6.2 Authentic Tasks

Participants were required to submit two evaluative tasks that they were going to do even if they were not enrolled in the study. I did not limit topics as participants were more likely to have real tasks if less constraint was given. The authentic tasks submitted varied greatly in topics. Shopping for products-such as electronic devices or skin products-were a little more common than other topics. Participants were also evaluating things other than physical products such as methods, services, and courses (see Appendix G for a full list of the tasks solicited). Other than serving as the authentic tasks for this research, this collection of real-life search tasks could also be used to inform task design in future studies.

I inquired about participants' experiences of formulating authentic tasks in the follow-up interviews. Most participants recounted that they had no problem thinking of two evaluative tasks since they were quite common in everyday lives. However, their tasks were not necessarily the ones that they needed to do right away. Some participants used this study as a chance to do non-urgent tasks that had been postponed before. For instance, P5 remarked that "*It kind of gave me an excuse to kind of get to doing it because*

I was procrastinating on it. The speakers have been bad for a couple of months now. Like, I'll look one up later, but this got me like going as you work on it." Although those tasks were not urgent or important enough for the participants to finish soon, they were chosen because participants were given the opportunity.

Participants were not always interested in or motivated to do their own tasks despite that they had to finish them, as commented by P11: "*The one I picked for the banking, to open a savings account. I don't have much knowledge about banking so I don't know a lot of the terms like APR and all that stuff. So I kind of found that a little boring. I mean if someone maybe were to explain it a little bit better*." Some participants even considered the simulated tasks more interesting. There were also a few participants who picked the tasks that they had been working on before the study. For instance, P21 was selecting food that was good for thyroid health. She recently got diagnosed with Hashimoto's disease and had been dealing with that. She said that the task has been "*kind of something that's been on my mind*" and "*I'm always trying to research more about that.*"

Notably, some participants pointed out that the authentic tasks they worked on for this study were not exactly the same as real-life tasks. They thought that those authentic tasks did not have to be completely realistic when they worked for a study. For example, P16 mentioned that he did not consider the financial aspect when searching for over-ear headphones as he would have done in real life since he did not need to buy it for real within the study. In other words, participants did not necessarily approach their own tasks exactly in the same way as they would have done in reality. As a result, the difference between authentic and simulated tasks may be bigger in reality.

6.3 Task Perceptions & Experiences

This section reports on the effects of study settings and task authenticity on participants' subjective ratings of task characteristics and engagement. Table 6.1 presents the means and standard deviations (SDs) of pre-task questionnaire responses. Table 6.2 presents the two-way repeated-measures ANOVA results for the pre-task questionnaire responses. For effect sizes, generalized eta squared (η^2_G) is reported as recommended by Bakeman (2005).

The two-way repeated-measures ANOVAs showed that task authenticity had significant effects on all of the perceived task characteristics. Coupled with descriptive statistics, they revealed that participants expected simulated tasks to be more difficult and complex than authentic tasks (difficulty: $F(1, 35) = 7.68$; $p = .009$; complexity: $F(1, 35) = 16.03$; $p < .001$). Participants were less topically knowledgeable, confident, motivated, and interested in simulated tasks than in authentic tasks (knowledge: $F(1, 35) = 61.47$; $p < .001$; confidence: $F(1, 35) = 14.59$; $p < .001$; motivation: $F(1, 35) = 18.49$; $p < .001$; interest: $F(1, 35) = 37.82$; $p < .001$). They also felt that their own tasks were more important than simulated tasks, $F(1, 35) = 55.04$; $p < .001$. Study settings had no significant effect on these perceived task characteristics. There was also no interaction effect detected.

Table 6.1: Means and SDs of pre-task questionnaire responses

	Lab Setting Mean (SD)	**Remote Setting Mean (SD)**	**Simulated Task Mean (SD)**	**Authentic Task Mean (SD)**
Task difficulty	2.94 (1.06)	2.86 (1.12)	3.13 (1.11)	2.68 (1.02)
Task complexity	3.22 (1.08)	3.25 (1.16)	3.53 (1.06)	2.94 (1.10)
Topic knowledge	2.56 (1.30)	2.71 (1.28)	2.00 (1.17)	3.26 (1.07)
Task importance	3.76 (1.18)	3.72 (1.01)	3.10 (1.13)	4.39 (0.55)
Topic interest	3.85 (1.03)	3.94 (0.93)	3.39 (1.03)	4.40 (0.60)
Task motivation	4.07 (0.88)	3.86 (0.89)	3.61 (0.91)	4.32 (0.71)
Task confidence	4.07 (0.76)	4.08 (0.71)	3.90 (0.72)	4.25 (0.71)

Table 6.2: The effects of study setting, task authenticity, and their interactions on pre-task questionnaire responses, *$p < .05$; **$p < .01$; ***$p < .001$

	Study Setting	**Task Authenticity**	**Setting x Task**
Task difficulty	F (1, 35) = 0.23, p = .631, η^2_G = .00	**F (1, 35) = 7.68**, p = .009**, η^2_G = .04	F (1, 35) = 0.11, p = .739, η^2_G = .00
Task complexity	F (1, 35) = 0.03, p = .860, η^2_G = .00	**F (1, 35) = 16.03***, p < .001**, η^2_G = .07	F (1, 35) = 0.02, p = .889, η^2_G = .00
Topic knowledge	F (1, 35) = 0.88, p = .355, η^2_G = .00	**F (1, 35) = 61.47***, p < .001**, η^2_G = .25	F (1, 35) = 2.68, p = .111, η^2_G = .01
Task importance	F (1, 35) = 0.13, p = .723, η^2_G = .00	**F (1, 35) = 55.04***, p < .001**, η^2_G = .35	F (1, 35) = 3.16, p = .084, η^2_G = .02
Topic interest	F (1, 35) = 0.70, p = .407, η^2_G = .00	**F (1, 35) = 37.82***, p < .001,** η^2_G = .27	F (1, 35) = 2.56, p = .119, η^2_G = .01
Task motivation	F (1, 35) = 2.78, p = .105, η^2_G = .02	**F (1, 35) = 18.49***, p < .001**, η^2_G = .16	F (1, 35) = 3.50, p = .069, η^2_G = .01

Task confidence	*F* (1, 35) = 2.07, $p = .886$, $\eta^2_G = .00$	***F* (1, 35) = 14.59***, *p* < .001**, $\eta^2_G = .06$	*F* (1, 35) = 3.67, $p = .549$, $\eta^2_G = .00$

Table 6.3 presents the means and standard deviations (SDs) of participants' post-task ratings of task difficulty, complexity, and engagement. Table 6.4 presents the two-way repeated-measures ANOVA results for these post-task questionnaire responses. Study settings had no significant impact on participants' engagement except that the lab setting was significantly more demanding than the remote setting, *F*(1, 35) = 10.72, *p* = .002. Participants were more interested, involved, drawn into searching, and had more fun in authentic tasks than in simulated tasks (interested: *F*(1, 35) = 24.23, *p* < .001; involved: *F*(1.35) = 4.48; *p* = .041; drawn into searching: *F*(1, 35) = 29.05; *p* < .001; fun: *F*(1, 35) = 5.85; *p* = .021). They were also more likely to lose themselves and lose track of time in their own tasks than in simulated tasks (lost self: *F*(1, 35) = 27.25; *p* < .001; lost track of time: *F*(1, 35) = 8.04; *p* < .008).

Table 6.3: Means and SDs of post-task questionnaire responses

	Lab Setting Mean (SD)	**Remote Setting Mean (SD)**	**Simulated Task Mean (SD)**	**Authentic Task Mean (SD)**
Task difficulty	2.68 (1.12)	2.43 (1.11)	2.60 (1.16)	2.51 (1.09)
Task complexity	3.01 (1.20)	2.79 (1.23)	3.04 (1.25)	2.76 (1.18)
Discouraged	2.08 (0.95)	2.33 (1.13)	2.32 (1.11)	2.10 (0.97)
Frustrated	1.99 (0.91)	2.10 (0.98)	1.97 (0.86)	2.11 (1.03)
Not as expected	2.18 (0.98)	2.17 (0.96)	2.19 (1.04)	2.15 (0.90)
Demanding	2.83 (1.19)	2.38 (0.96)	2.61 (1.18)	2.60 (1.02)
In control	3.92 (0.95)	4.13 (0.79)	3.96 (1.00)	4.08 (0.73)

Interested	4.06 (0.93)	4.06 (0.89)	3.68 (0.98)	4.43 (0.65)
Had fun in searching	3.51 (.14)	3.56 (1.10)	3.26 (1.14)	3.81 (1.03)
Involved	4.13 (0.84)	4.03 (0.87)	3.90 (0.87)	4.25 (0.80)
Rewarding	3.92 (0.95)	3.86 (0.97)	3.50 (1.03)	4.28 (0.68)
Drawn into searching	3.82 (0.91)	3.67 (1.01)	3.35 (1.05)	4.14 (0.66)
Lost self	2.69 (1.10)	2.75 (0.88)	2.40 (0.88)	3.04 (1.00)
Lost track of time	2.68 (1.10)	2.61 (0.94)	2.44 (0.99)	2.85 (1.02)
Success	4.17 (0.73)	4.04 (0.76)	4.04 (0.80)	4.17 (0.69)

Table 6.4: The effects of study setting, task authenticity, and their interactions on post-task questionnaire responses, $*p < .05$; $**p < .01$; $***p < .001$

	Study Setting	**Task Authenticity**	**Setting x Task**
Task difficulty	$F (1, 35) = 0.91$, $p = .176$, $\eta^2_G = .02$	$F (1, 35) = 0.16$, $p = .690$, $\eta^2_G = .00$	$F (1, 35) = 0.55$, $p = .461$, $\eta^2_G = .00$
Task complexity	$F (1, 35) = 1.32$, $p = .259$, $\eta^2_G = .01$	$F (1, 35) = 2.25$, $p = .143$, $\eta^2_G = .02$	$F (1, 35) = 0.56$, $p = .459$, $\eta^2_G = .01$
Discouraged	$F (1, 35) = 2.56$, $p = .119$, $\eta^2_G = .02$	$F (1, 35) = 1.93$, $p = .173$, $\eta^2_G = .02$	**$F (1, 35) = 8.62**$, $p = .006$**, $\eta^2_G = .07$
Frustrated	$F (1, 35) = 0.46$, $p = .504$, $\eta^2_G = .01$	$F (1, 35) = 0.94$, $p = .338$, $\eta^2_G = .01$	$F (1, 35) = 2.74$, $p = .107$, $\eta^2_G = .02$
Not as expected	$F (1, 35) = 0.01$, $p = .930$, $\eta^2_G = .00$	$F (1, 35) = 0.06$, $p = .815$, $\eta^2_G = .00$	$F (1, 35) = 0.14$, $p = .707$, $\eta^2_G = .00$
Demanding	**$F (1, 35) = 10.72**$, $p = .002$**, $\eta^2_G = .06$	$F (1, 35) = 0.00$, $p = .945$, $\eta^2_G = .00$	$F (1, 35) = 0.46$, $p = .502$, $\eta^2_G = .00$
In control	$F (1, 35) = 1.47$, $p = .233$, $\eta^2_G = .02$	$F (1, 35) = 1.15$, $p = .292$, $\eta^2_G = .01$	$F (1, 35) = 1.13$, $p = .719$, $\eta^2_G = .00$
Interested	$F (1, 35) = 0.00$, $p = 1.000$, $\eta^2_G = .00$	**$F (1, 35) = 24.23***$, $p < .001$,** $\eta^2_G = .23$	$F (1, 35) = 1.45$, $p = .237$, $\eta^2_G = .01$

Had fun in searching	F (1, 35) = 0.10, p = .758, $\eta^2_G = .00$	**F (1, 35) = 5.85*,** **p = .021,** $\eta^2_G = .09$	F (1, 35) = 0.09, p = .768, $\eta^2_G = .00$
Involved	F (1, 35) = 0.86, p = .361, $\eta^2_G = .01$	**F (1, 35) = 4.48*,** **p = .041**, $\eta^2_G = .07$	F (1, 35) = 2.19, p =.148, $\eta^2_G = .01$
Rewarding	F (1, 35) = 0.19, p = .669, $\eta^2_G = .00$	**F (1, 35) = 24.41***,** **p < .001**, $\eta^2_G = .24$	F (1, 35) = 0.23, p = .634, $\eta^2_G = .00$
Drawn into searching	F (1, 35) = 1.64, p = .208, $\eta^2_G = .01$	**F (1, 35) = 29.05***,** **p < .001**, $\eta^2_G = .25$	F (1, 35) = 0.27, p = .607, $\eta^2_G = .00$
Lost self	F (1, 35) = 0.25, p = .618, $\eta^2_G = .00$	**F (1, 35) = 27.25***,** **p < .001**, $\eta^2_G = .22$	**F (1, 35) = 6.15*,** **p = .018**, $\eta^2_G = .06$
Lost track of time	F (1, 35) = 0.26, p = .611, $\eta^2_G = .00$	**F (1, 35) = 8.04**,** **p = .008**, $\eta^2_G = .07$	F (1, 35) = 1.37, p = .250, $\eta^2_G = .01$
Success	F (1, 35) = 1.62, p = .212, $\eta^2_G = .01$	F (1, 35) = 0.93, p = .342, $\eta^2_G = .01$	F (1, 35) = 0.89, p = .352, $\eta^2_G = .01$

Study settings and task authenticity had interaction effects on participants' ratings of feeling discouraged and losing themselves in the search tasks (discouraged: F (1, 35) = 8.62; p = .006; losing self: F (1, 35) = 6.15, p = .018). This indicates that study settings had a different impact on these two variables depending on which task participants were working on. I use bar charts to determine the nature of these interactions. As Figure 6.2 shows, participants felt more discouraged in the remote setting than in the lab setting when they were working on authentic tasks. However, they felt more discouraged in the lab than in the remote setting when they worked on simulated tasks. Similarly, Figure 6.3 shows that participants were more likely to lose themselves in the lab than in the remote setting when they worked on authentic tasks. In simulated tasks, however, they were more likely to lose themselves in the remote setting than in the lab.

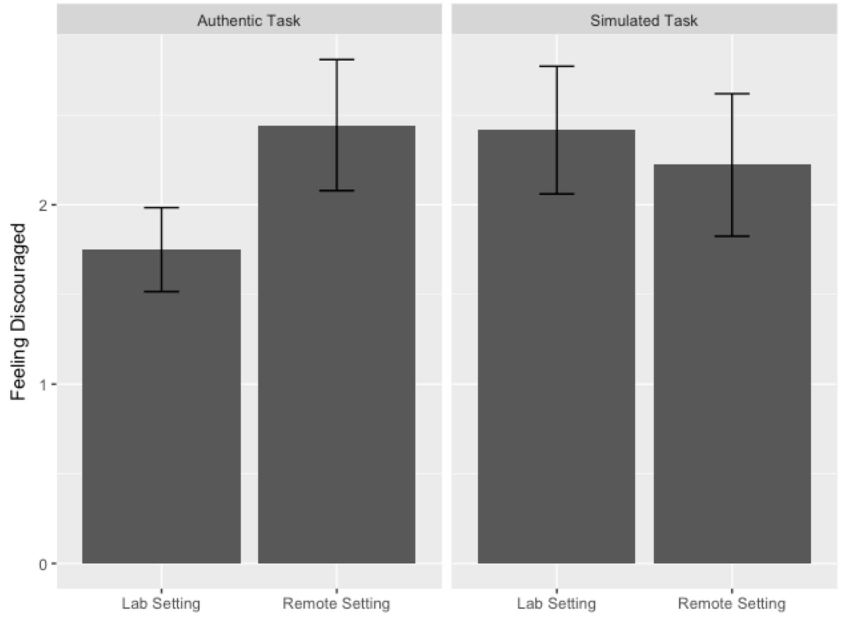


Figure 6.2: Participants’ post-task ratings of feeling discouraged (means, 95% confidence intervals)

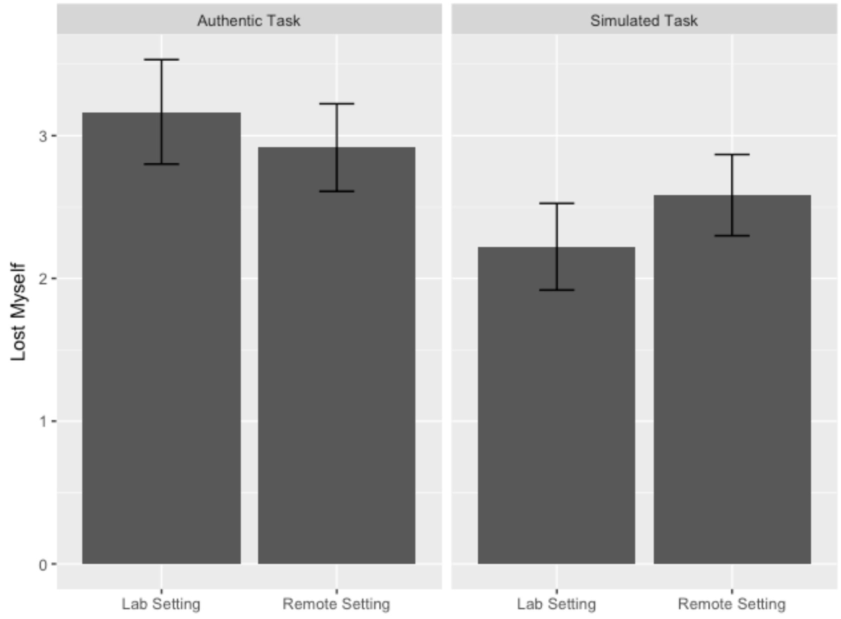


Figure 6.3: Participants’ post-task ratings of losing themselves in the search experience (means, 95% confidence intervals)

Table 6.5 reports the frequency of information seeking problems selected by the participants in the post-task questionnaire. I did not run statistical tests to examine the relationships between these problems and study variables because most of them were only selected a few times. Frequency analysis shows that participants were more likely to encounter problems in articulating information needs (n=10, 14% of observations) and obtaining relevant results (n=12, 17%) in the lab setting than in the remote setting (n=4, 9, 6% and 13% respectively). They were more likely to be impatient (n=10, 14%), unaware of relevant information sources (n=16, 22%) and feeling unconfident about finding information (n=14, 19%) when they worked remotely compared to when they were in the lab (n=7, 11, 5, 10%, 15%, and 7%, respectively). Participants were more likely to feel unconfident about finding information in simulated task (n=12, 17%) than in their own tasks (n=7, 10%). However, more of them felt that the information was too scattered (n=20, 28%) and there were too many irrelevant results (n=13, 18%) in authentic tasks than in simulated tasks (n=10, 8, 14% and 11%, respectively).

Table 6.5: Frequency of information seeking problems selected by the participants in post-task questionnaires

	Lab Simulated	**Lab Authentic**	**Remote Simulated**	**Remote Authentic**
Lack of sufficient patience	8%	11%	14%	14%
Unable to articulate information needs	14%	14%	6%	6%
Unaware of relevant information sources	19%	11%	22%	22%
Too much information	36%	25%	31%	33%
Information was too scattered	14%	28%	14%	28%
Information was not up-to-date	3%	8%	6%	3%

Too many irrelevant results	14%	19%	8%	17%
Look at wrong sources	6%	8%	8%	0
Unconfident about finding information	11%	3%	22%	17%
Financial constraints	3%	0	0	0
Poor search skills	0	0	6%	6%
Unable to understand the information found	11%	8%	8%	3%

6.4 Behavioral Measures

This section reports the descriptive statistics for the behavioral measures and the results of multilevel models that examined the influence of study setting and task authenticity on behavioral measures. The multilevel models controlled the effects of user-perceived task characteristics and demonstrated the effects that could be attributed to the study manipulations. One model was built for each behavioral measure and independent variables were the same in each model which include study setting, task authenticity, the interaction between setting and task, experienced task difficulty, experienced task complexity, topic knowledge, task importance, topic interest, motivation, and confidence. Marginal R^2 (R^2m) associated with fixed effects and conditional R^2 (R^2c) associated with both fixed and random effects (Nakagawa & Schielzeth, 2013) are reported at the end of each table. In addition to searching and browsing behaviors, I also estimated models for pre-task and post-task questionnaire completion time to see if the study variables had an influence on users' participation behavior.

Descriptive statistics in Table 6.6 shows that participants spent less time on tasks and viewing content pages when they worked in the lab than when they worked remotely.

They also visited more pages when working in the lab. Regarding the differences between tasks, participants took more time to view pages in simulated tasks than in authentic tasks but looked at more pages and spent more time in completing tasks in authentic tasks.

Table 6.6: Means and SDs of behavioral measures

	Lab Setting Mean (SD)	Remote Setting Mean (SD)	Simulated Task Mean (SD)	Authentic Task Mean (SD)
Avg. query time	246.13 (305.84)	264.35 (433.48)	257.50 (450.23)	252.97 (280.88)
Avg. uniq. query time	254.33 (302.16)	287.39 (431.98)	267.46 (447.73)	274.26 (279.25)
Avg. query length	4.11 (1.60)	4.08 (2.00)	4.08 (1.86)	4.11 (1.76)
No. queries	4.99 (6.73)	4.04 (3.81)	3.60 (3.66)	5.43 (6.72)
No. uniq. queries	3.85 (3.61)	3.44 (3.03)	3.00 (2.60)	4.29 (3.83)
Avg. page visits	1.75 (0.89)	1.57 (0.82)	1.67 (1.00)	1.64 (0.70)
Avg. page dwell time	30.35 (37.32)	43.43 (54.48)	49.46 (56.64)	24.32 (30.31)
Avg. uniq. page dwell time	52.01 (58.60)	67.42 (71.16)	76.83 (74.86)	42.60 (49.18)
Avg. 1st. deci. time	10.69 (8.17)	16.28 (19.42)	14.67 (16.58)	12.31 (13.49)
No. pages	31.60 (38.29)	22.07 (22.98)	14.75 (16.00)	38.92 (38.57)
No. uniq. pages	16.64 (20.34)	11.43 (10.22)	7.21 (6.12)	20.86 (20.00)
No. query w/o pages	1.01 (2.78)	0.72 (1.05)	0.57 (1.05)	1.17 (2.75)
Task time	940.75 (557.90)	1245.47 (1190.42)	989.82 (798.09)	1196.40 (1056.80)
Searching & browsing time	586.29 (441.37)	799.00 (994.89)	549.00 (553.45)	836.29 (927.03)

No. uniq. SERPs	5.50 (5.61)	4.99 (5.58)	3.74 (2.95)	6.75 (7.03)
Total page time	504.89 (415.72)	692.96 (952.42)	476.18 (506.12)	721.67 (900.64)
Prop. page task	0.52 (0.22)	0.49 (0.25)	0.46 (0.20)	0.55 (0.25)
Prop. page task w/o task page	0.82 (0.16)	0.80 (0.23)	0.82 (0.18)	0.81 (0.21)
1st deci. time	11.46 (10.71)	16.36 (19.10)	16.10 (18.12)	11.72 (12.40)
1st query time	250.42 (338.93)	260.44 (436.28)	268.46 (476.99)	242.40 (278.22)
1st total page time	221.39 (332.11)	223.71 (407.06)	233.36 (449.33)	211.74 (271.77)
1st no. pages	14.15 (24.58)	8.15 (9.85)	7.60 (10.36)	14.71 (24.22)
1st no. uniq. pages	6.39 (11.39)	3.64 (4.26)	2.47 (2.63)	7.56 (11.47)
1st avg. page time	21.69 (29.82)	30.89 (45.83)	28.43 (44.63)	24.15 (32.11)
1st avg. uniq. page time	46.17 (68.29)	58.44 (78.26)	70.92 (95.02)	33.69 (33.66)
Pre. ques. time	59.54 (37.06)	94.14 (137.97)	89.44 (99.22)	64.24 (104.15)
Post. ques. time	72.26 (30.18)	89.89 (95.14)	77.25 (44.97)	84.90 (89.82)

Table 6.7 reports the MLM results for different types of page dwell time. Post-task complexity and difficulty were used in all models. When estimating the model, I coded the remote setting and authentic tasks as 0 and the lab setting and simulated tasks

as 1. Thus, a positive effect of study setting or task on a behavioral measure means that the synthetic configuration of setting or task have a positive effect on this variable. For example, Table 6.7 shows that study setting negatively influenced average page dwell time (-0.34, *p* = .022) per task, meaning that the lab setting negatively influenced the average dwell time that participants spent on content pages in a task. By contrast, simulated tasks positively influenced average page dwell time (0.47, *p* = .042) per task. Simulated tasks also had a positive effect on the average time participants spent on unique content pages (0.69, *p* = .003) while study setting had no significant effect.

Study setting and task authenticity had an interaction effect on the total time that participants spent on content pages (-0.49, *p* = .018). This means that study settings had different effects on total page time, depending on which task that participants were working on. Meanwhile, task complexity positively influenced total page time (0.21, *p* = .007).

Table 6.7: Multilevel model results for page dwell time, **p* < .05; ***p* < .01; ****p* < .001

	Total page time		**Avg. page time**		**Avg. uniq. page time**	
	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value
Study setting	0.25	.226	**-0.34***	.022	-0.11	.524
Task authenticity	-0.05	.821	**0.47***	.042	**0.69****	.003
Setting x task	**-0.49***	.018	0.05	.793	-0.19	.416
Difficulty	-0.03	.654	-0.04	.594	-0.02	.793
Complexity	**0.21****	.007	0.04	.554	0.05	.494
Knowledge	0.06	.421	-0.01	.806	0.00	.963
Importance	-0.06	.545	0.04	.669	0.00	.939
Interest	-0.12	.363	0.01	.953	0.02	.858

Motivation	0.11	.426	-0.09	.379	0.06	.662
Confidence	-0.06	.556	0.01	.944	-0.02	.817
Intercept	**4.63*****	< .001	**3.36*****	< .001	**3.10*****	< .001
R^2m	.09		.13		.10	
R^2c	.81		.69		.55	

Table 6.8 displays the MLM results for content page visits. The lab setting significantly increased the number of content pages visited (0.48, $p = .009$) and unique content pages visited (0.46, $p = .003$) during a task session. Task motivation also had a positive effect on the number of pages visited (0.25, $p = .031$). Simulated tasks had a negative effect on the number of pages (-0.62, $p = .003$) and unique pages (-0.61, $p < .001$) visited per task. Their interaction effect was a significant predictor of the unique pages (-0.40, $p = .047$), meaning that the effect of study settings on the number of unique pages visited were different in authentic and simulated tasks. In addition, the lab setting positively affected the average number of times that a page was visited within a task session (0.11, $p = .030$). In brief, the synthetic setting influenced users to look at more content pages while synthetic tasks did the contrary. Plus, participants also visited each page more times in the lab than in the remote setting.

Table 6.8: Multilevel model result for page visits, *$p < .05$; **$p < .01$; ***$p < .001$

	No. page		**No. uniq. page**		**Avg. page visits**	
	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value
Study setting	**0.48****	.009	**0.46****	.003	**0.11***	.030
Task authenticity	**-0.62****	.003	**-0.61*****	< .001	0.03	.601
Setting x task	-0.46	.068	**-0.40***	.047	-0.11	.127
Difficulty	-0.02	.811	0.01	.879	-0.00	.865

Complexity	0.08	.314	0.07	.229	0.30	.184
Knowledge	-0.00	.987	0.02	.681	0.00	.977
Importance	-0.02	.837	-0.03	.668	-0.04	.220
Interest	-0.10	.364	-0.04	.678	0.00	.897
Motivation	**0.25***	.031	0.15	.149	0.05	.130
Confidence	0.09	.433	0.00	.994	0.00	.969
Intercept	**1.78****	.007	**1.91*****	< .001	**0.70*****	< .001
R^2m	.26		.29		.06	
R^2c	.99		.60		.53	

Table 6.9 reports the MLM results for the number of queries issued and the number of unique SERPs viewed. Neither of the study manipulations had significant effects on these behavioral measures. However, several user-reported task characteristics demonstrated significant impact. The number of queries issued per task was negatively affected by topic knowledge (-0.23, *p* < .001) and positively affected by confidence (0.30, *p* = .003). Similarly, the number of unique queries used was also negatively influenced by topic knowledge (-0.15, *p* = .022) and positively affected by confidence (0.23, *p* = .020). Although these two measures were not significantly affected by task authenticity itself, they were significantly different between authentic and simulated tasks because task knowledge and confidence were different between tasks. Similarly, no significant effects of study setting or task authenticity were detected on the number of unique SERPs visited per task. Participants' topic knowledge negatively affected the number of unique SERPs visited (-0.20, *p* = .002) while their confidence in completing the task positively affected this measure (0.26, *p* = .008).

Table 6.9: Multilevel model results for number of queries and unique SERPs, *$p < .05$; **$p < .01$; ***$p < .001$

	No. queries		**No. uniq. queries**		**No. queries w/o page**		**No. uniq. SERPs**	
	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value
Study setting	0.14	.338	0.06	.669	0.04	.741	0.23	.133
Task authenticity	-0.10	.565	-0.23	.175	-0.04	.777	-0.29	.107
Setting x task	-0.21	.283	-0.01	.944	-0.12	.463	-0.25	.189
Difficulty	0.06	.427	0.03	.673	0.01	.791	0.04	.598
Complexity	0.02	.762	0.06	.354	0.02	.707	0.04	.563
Knowledge	**-0.23*** **	< .001	**-0.15***	.022	-0.08	.089	**-0.20****	.002
Importance	0.10	.200	0.12	.124	0.03	.579	0.11	.152
Interest	0.12	.227	0.02	.856	0.08	.299	0.01	.898
Motivation	-0.05	.636	-0.03	.800	0.03	.689	-0.03	.771
Confidence	**0.30****	.003	**0.23***	.020	0.11	.158	**0.26****	.008
Intercept	-0.06	.907	-0.16	.761	-0.44	.274	0.33	.537
R^2m	.14		.11		.08		.15	
R^2c	.62		.60		.33		.59	

Table 6.10 reports the MLM results for average query dwell time and unique query dwell time per task, average query length per task, and average first decision time per task. Simulated tasks positively affected the average first decision time per task (0.35, $p = .038$), meaning that participants took more time to decide the first page to click on after issuing a query in simulated tasks than in authentic tasks. The study variables did not have a significant impact on other measures in this table. Average query length was positively affected by task interest (0.10, $p = .035$) and negatively affected by task importance (-0.08, $p = .049$).

Table 6.10: Multilevel model results for query segment time, query length, and first decision time, *$p < .05$; ***$p < .001$

	Avg. query time		**Avg. uniq. query time**		**Avg. query len.**		**Avg. 1st. deci. time**	
	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value
Study setting	0.18	.367	0.06	.755	0.09	.307	-0.21	.189
Task authenticity	0.17	.476	0.05	.805	-0.06	.475	**0.35***	.038
Setting x task	-0.47	.090	-0.33	.140	-0.11	.217	0.04	.839
Difficulty	0.05	.565	0.08	.339	0.02	.558	0.09	.165
Complexity	-0.09	.287	0.06	.466	0.02	.564	-0.03	.609
Knowledge	0.08	.324	0.06	.419	-0.05	.059	0.09	.101
Importance	-0.04	.705	-0.04	.722	**-0.08***	.049	-0.01	.899
Interest	-0.10	.439	-0.04	.723	**0.10***	.035	**0.18***	.044
Motivation	0.18	.181	0.13	.281	-0.02	.630	-0.13	.205
Confidence	-0.17	.195	-0.11	.328	-0.06	.184	-0.15	.061
Intercept	**5.4*****	< .001	**5.19*****	< .001	**1.88*****	< .001	**2.39*****	< .001
R^2m	.05		.04		.13		.13	
R^2c	.36		.51		.99		.47	

Table 6.11 reports the MLM results for task completion time and the proportion of time spent on content pages in a task. Since participants also spent time reading task descriptions and taking notes, two types of task completion time were calculated: *task completion time* that includes all activities within a task and *searching and browsing time* that excludes the time spent on reading task descriptions and typing notes. Study setting and task authenticity did not have significant effects on task completion time. Their interaction influenced the proportion of time on content pages in a task (-0.10, $p = .023$).

Table 6.11: Multilevel model results for task time and the proportion of time on content pages in a task, **p* < .05; ***p* < .01; ****p* < .001

	Task time		**Searching & browsing time**		**Prop. page task**		**Prop. page task w/o task page**	
	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value
Study setting	-0.10	.554	0.10	.567	0.07	.066	-0.05	.791
Task authenticity	-0.06	.736	-0.17	.305	-0.09	.093	-0.23	.328
Setting x task	-0.15	.388	-0.31	.083	**-0.10***	.023	0.12	.576
Difficulty	0.07	.288	0.03	.620	0.00	.832	-0.14	.129
Complexity	0.07	.259	0.09	.144	0.02	.274	0.13	.084
Knowledge	0.06	.373	-0.01	.838	0.00	.927	0.10	.198
Importance	-0.05	.514	0.00	.962	-0.02	.349	-0.17	.100
Interest	-0.02	.833	0.09	.389	0.01	.707	-0.19	.123
Motivation	0.19	.079	0.05	.672	-0.02	.587	**0.30***	.028
Confidence	0.01	0.942	-0.01	.882	-0.03	.315	-0.18	.118
Intercept	**5.87*****	< .001	**5.44*****	< .001	**0.67*****	< .001	**2.81*****	< .001
R^2m	.07		.09		.08		.08	
R^2c	.69		.73		.74		.64	

In respect to first-query behaviors, simulated tasks positively influenced first decision time (0.65, *p* = .001) which was the time participants spent to decide which page to click on first after issuing the first query (see Table 6.12). Simulated tasks also negatively influenced the number of unique pages visited in the first query (-0.47, *p* = .034) (see Table 6.13), same as their influence on the total number of unique pages visited per task. Overall, study settings and task authenticity did not have as much influence on first-query measures as they had on whole-session measures. At least on the first-query behaviors measured in this study, study settings did not have a significant impact. Table 6.13 also reports the multilevel modeling results for questionnaire completion time. Simulated tasks positively affected the time taken to complete the pre-task questionnaire (0.49, *p* = .005). The time taken to complete the post-task questionnaire was not affected by study variables or task characteristics.

Table 6.12: Multilevel model results for different types of first query dwell time, **p* < .05; ***p* < .01; ****p* < .001

	1st query time		**1st page time**		**1st deci. time**		**1st avg. page time**		**1st avg. uniq. page time**	
	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value
Study setting	0.13	.608	0.36	.463	-0.02	.878	0.10	.765	0.02	.948
Task authenticity	-0.05	.876	-0.46	.396	**0.65****	.001	-0.13	.708	-0.26	.475
Setting x task	-0.60	.070	-0.56	.314	-0.32	.103	-0.31	.406	-0.31	.402
Difficulty	0.17	.178	0.19	.367	0.11	.136	0.09	.484	0.12	.389
Complexity	-0.17	.135	-0.21	.257	-0.02	.667	-0.07	.541	-0.07	.570
Knowledge	-0.01	.946	0.02	.899	0.06	.338	-0.02	.861	-0.03	.805
Importance	0.06	.674	-0.09	.717	0.04	.624	0.03	.861	0.02	.917
Interest	-0.19	.299	-0.28	.349	**0.22***	.031	-0.15	.429	-0.21	.270
Motivation	0.22	.241	0.44	.169	-0.17	.139	0.03	.873	0.04	.831
Confidence	-0.26	.125	-0.25	.366	-0.12	.229	-0.12	.482	-0.10	.561
Intercept	**5.63*****	< .001	**4.99****	.002	**1.88****	.001	**3.49*****	< .001	**3.72*****	< .001
R^2m	.06		.05		.13		.02		.03	
R^2c	.56		.50		.51		.47		.51	

Table 6.13: Multilevel model results for first query measures and questionnaire completion time, **p* < .05; ***p* < .01; ****p* < .001

	1st no. page		**1st no. uniq. page**		**Pre. ques. Time**		**Post. ques. time**	
	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value	Fixed effects	*p*-value
Study setting	0.35	.168	0.34	.073	-0.12	.404	-0.05	.705

Task authenticity	-0.31	.295	**-0.47***	.034	**0.49****	.005	0.09	.543
Setting x task	-0.29	.372	-0.30	.235	-0.06	.708	0.04	.816
Difficulty	0.15	.196	0.10	.223	-0.04	.568	-0.01	.789
Complexity	-0.14	.186	-0.08	.267	0.01	.862	-0.07	.195
Knowledge	0.07	.499	0.04	.565	-0.01	.874	0.02	.753
Importance	-0.07	.602	-0.01	.925	0.08	.266	0.10	.119
Interest	-0.17	.299	-0.13	.236	0.01	.915	-0.00	.965
Motivation	**0.35***	.042	0.21	.092	-0.08	.390	-0.06	.480
Confidence	-0.16	.316	-0.11	.322	-0.06	.481	-0.09	.240
Intercept	**1.85***	.036	**1.52***	.017	**4.18*****	< .001	**4.61*****	< .001
R^2m	.11		.18		.12		.04	
R^2c	.47		.36		.52		.37	

6.5 Summary

This chapter reports on participants' demographic information, their authentic tasks, and quantitative analysis results. Participants were recruited campus-wise from Rutgers University. Thirty-six undergraduate students from 21 different disciplines finished the study. Each person submitted two evaluative tasks that they were going to do in real life; and those tasks varied in topics and urgency.

Two-way repeated-measures ANOVA and multilevel modeling were conducted to examine the impacts of study settings and task authenticity on subjective task characteristics, engagement, and behavioral measures. To summarize, all of users' pre-task subjective task perceptions were significantly different between authentic and simulated tasks. Overall, they were more knowledgeable, confident, interested in, and motivated to do their own tasks and their own tasks were more engaging than simulated tasks. Study settings did not have a significant impact on pre-task ratings. However, the lab was considered more demanding than the remote setting.

Regarding the behavioral measures, study settings and task authenticity had significant impact mostly on measures related to content pages such as average page dwell time and page visits, but not on query-related measures or overall task completion time. Their effects on first-query measures were not as wide as on whole-session measures. While study settings had no significant influence on first-query measures, task authenticity affected first decision time and the number of unique pages visited in the first query. In the next chapter, I discuss the findings by returning to each research question.

Chapter 7 : Discussion

This chapter addresses each research questions by discussing the statistical analyses results. Where applicable, interview data was used to explain the quantitative findings.

7.1 Subjective Task Characteristics and Experiences

I start this discussion by answering RQ3: *Do users' perceptions and experiences of tasks (e.g., task difficulty, engagement) differ in different study settings and tasks?*

Task authenticity had significant impacts on all of the pre-task ratings of subjective task characteristics. Participants were more topically knowledgeable and confident in their own tasks mostly because they had searched for or had been working on similar topics before, as commented by P21: "*I have researched that in the past, you know I was looking up skin care stuff and everything, so I did have some background on it and have heard of the things that I was searching already, so I didn't feel like I needed to take too much time with that.*" Some of their tasks were routine tasks that they likely needed to do once in a while, such as looking for cloths, electronic devices, or funding opportunities. Some participants mentioned that they "*knew exactly where to look*" when working on authentic tasks because they have researched those topics before. This is consistent with Poddar and Ruthven's (2010) and Li and Hu's (2013) finding that participants were more familiar with and confident in their own tasks than simulated tasks. This also mirrors Poddar and Ruthven's observation that participants were more likely to bring tasks concerning topics about which they had a fair level of knowledge.

By contrast, the simulated task situations looked less familiar to participants. Because of this, they expected simulated tasks to be more complex and difficult than their own tasks. However, they did not actually experience differences in complexity and difficulty according to their post-task responses. Participants felt that the simulated tasks were easier than expected while their own tasks were often not as easy as they thought. This is consistent with Li and Hu (2013), who reported that participants judged simulated tasks to be more difficult prior to starting and they were more likely to underestimate the level of difficulty in authentic tasks than in simulated tasks. In this study, one factor that made simulated tasks easier was that participants were only looking for generic information for simulated tasks. They recalled that the simulated tasks looked complicated because they did not have a lot of exposure to those topics before but "*there was this one site that basically had all the information*" (P28), so they were less complex than expected. The topics chosen for the two simulated tasks (i.e., tattoo removal and lupus treatment) were indeed not simple, but the lack of specificity and personal context prevented participants from digging deeper and only allowed them to take "*a superficial look*" (P30). For authentic tasks, they needed to be more specific and fulfill their personal needs, as P1 commented: "*Because I think near where I live there are too many doctors, so it was hard to narrow the results down and find someone that I was actually looking for.*" She did not have the same problem when working on simulated tasks because the scenarios were not personalized. P3 also suggested that she "*would have also looked into more local options, as opposed to just like a generic, "Which one's the best?"*" if the simulated task situations were real.

Overall, participants were more interested in and motivated to do their own tasks than simulated tasks because that they “*would have been doing this search regardless*” (P9). They thought that their own tasks were more important because those were something “*need to get done*” and “*something that I know is going to have a direct impact on my life*” (P5), while the information gathered for simulated tasks served no real purpose. The simulated tasks were not relevant to most of the participants at the time of the study even though they felt that the two situations may happen in the future so the information found may be potentially useful at a future time. Some felt that these tasks would have been more important than their own tasks if they were real. Although they were provided with a scenario for each simulated task, they had difficulties imagining themselves in those situations. There were a few participants who put themselves in the simulated scenarios for real and invested even more efforts than their own tasks, such as P4 who felt the simulated task “*is important, because it’s health, it’s a complicated issue... it’s a real life*” while his own task was not very important to him. To some participants, the simulated tasks were even more interesting than authentic task because of their inadequate knowledge on the topics: “*The lupus question I was surprisingly interested in, just because I don’t know much about it or its medications, I was surprised to see the side effects that would occur with each one*” (P3). The simulated tasks also happened to be personally relevant to a few participants who wanted to remove their tattoos or had friends or family members who were diagnosed with lupus like P10: “*I do want to go in the medical field. I want to be a PA. So, that was intriguing to learn about. And my friend’s mother died of Lupus, so it was interesting to look at.*” One participant who majored in pharmacy thought the tasks were personalized for her.

Other than perceived task characteristics, participants' task engagement was also different between authentic and simulated tasks. In general, they were more engaged in their own tasks than in simulated tasks. This was reflected in higher ratings of a few items in the post-task questionnaire, such as feeling involved and drawn into searching. Authentic tasks provided specific contexts that allowed participants to go into more depth than simulated tasks. In authentic tasks, participants had a better knowledge of what they wanted to find out than in the simulated tasks where they just looked for general information regarding the topics. Also, the outcomes of authentic tasks may affect them for real. For example, P17 believed that she "*had to know everything about it [her own task]*" because "*it's something that's going to affect me directly*". Occasionally, they seemed to forget about time when working on their own tasks. I will return to this point later when discussing the behavioral differences.

Participants felt that their search experiences were more rewarding when working on their own tasks because the information they found out could be useful in real-life scenarios, like P5 suggested, "*It's information that I can actually use rather than information that I will forget within like ten minutes*." Some of them felt that they spent more time in their own tasks because they were not just doing them for a study.

Study settings did not have a significant impact on participants' pre-task or post-task questionnaire responses for the most part. Participants felt that the tasks were more demanding in the lab setting than in the remote setting. This is probably due to the lab environment and the fact that the tasks were open-ended without definite answers. Participants could decide how much information to find and how much time to invest before making a decision. Thus, they might feel more pressured to invest effort when they were in the lab. For example, P11 commented that the lab "*was a little scrutinizing*" compared to his dorm. Their selections of information seeking problems also indicated that they were more likely to be patient when they worked in the lab than when working remotely. This corresponds to the evidence in interview data. For example, P26 brought up that she "*had more patience here [in the lab]",* but she *"wasn't as motivated to go through it*" when working remotely.

Another interesting difference between the two settings emerged from the interviews. While the tasks were about searching on the internet for information, they may require activities other than searching, which could not be finished in the lab without proper instruments or resources. For example, P28 was looking for contact lenses and she wanted to measure her eye diameter to find the best fit. She complaint about having no means of doing that when she was in the lab. Participants were also accustomed to using mobile devices to do certain tasks such as checking a dictionary. For example, P4 was an international student whose native language was not English. He habitually reached his cellphone to check the dictionary when working in the lab. After being requested to not use a cellphone, he explained that he did not realize that he could just use Google Translate because he was so used to check his cellphone. He probably would have used it if he were not in the lab. Many participants reported that they used smart phones to finish trivial search tasks such as looking for restaurants or directions, but usually not for the tasks in this study because small screens are not suitable to do intensive searching and reading.

To summarize, this section discusses the differences in subjective task characteristics and experiences between study settings and between tasks using the interview data. Perceived task characteristics and engagement were not significantly different between settings for the most part except that participants perceived the lab to be more demanding than the remote setting. This could be attributed to the laboratory effect which pushed users to invest more efforts. They were more motivated to do their own tasks because those were something to be done even without this study. They were also more engaged in authentic tasks than simulated tasks not only because they would have done those authentic tasks regardless, but also because authentic tasks were associated with their personal situations and allowed them to go into more depth. In addition, their own tasks were also more rewarding because that the information found could be used for real.

7.2 Behavioral Measurements

In this section, I will discuss RQ1 and RQ2: *Do study settings and the authenticity of study tasks affect users' Web search behavior?* I start the discussion by comparing two study settings and then authentic and simulated tasks.

7.2.1 Comparison Between Study Settings

I used the multilevel modeling approach to examine the influence of study setting and task authenticity on searching and browsing behavior. One of the advantages of MLM is that it could account for the effects of user-reported subjective task characteristics and thus sort out the effects of the study manipulations. For instance, in Table 6.9, participants' querying behaviors were not affected by the experiment variables while their subjective topic knowledge and confidence influenced some aspects of querying behaviors. It is worth pointing out that users can have various level of knowledge and confidence in real-life tasks, and search behaviors may be affected by those task characteristics regardless of task authenticity. In other words, being authentic or synthetic itself did not change querying behaviors, but those behaviors could diverge between authentic and simulated tasks due to different subjective perceptions of the tasks used in this study.

Most participants felt that the environment had an impact on their experience. They formed a perception of the lab the second they walked in: "*I think the first thing I noticed was that everything ... I was like maybe you took stuff down so there were no distractions… I am very focused on this because there is nothing really around to distract me.*" (P20). They felt that the lab environment was formal and designed intentionally for them to focus on the tasks. P23 recalled that "*In this sort of formal environment [the lab], I had to read very carefully and just be more focused and not go on YouTube.*" Study setting had significant effects on most page-visiting behaviors. Participants not only visited more content pages, but also visited the same page more times in the lab setting than in the remote setting. However, the time they stayed on a page was shorter in the lab. More pages and shorter page dwell time in the lab could be due to that participants were more concentrated, worked more efficiently, and invested greater effort in the lab. They mentioned that they wanted to "*try harder*" and "*work more efficiently*" in the lab than in the field. For example, P28 felt that she "*had something to prove*" when working in the lab, but she did not feel the same when working remotely even though she was still in the same study. They also believed that their work was of higher quality in the lab. A few of them felt that the lab was too quiet which made them uncomfortable because they were used to work in busy areas. A difference in the number of pages visited between two settings has not been found by existing studies probably because that they mostly used known-item or fact-finding tasks in which participants' searching path was relatively similar. In the open-ended tasks used in this dissertation, participants could decide how much information to find and when to stop searching, making the impact of study settings more evident.

There were also participants who did not feel a difference between settings such as P8 who "*completely forgot about it [the environment]*" once she started working on the tasks. A few participants stated that the difference between two settings was "*very slight*" so they did not expect that it would affect their search results or experience. They liked to finish their tasks soon and move on regardless of where they were working.

Temporary distractions from various sources were very common in the remote setting. Most participants mentioned that they checked their cellphone or had their TV on when working remotely. Even in the lab setting, a few of them put their cellphone by their side. The search log also revealed that they frequently visited websites such as YouTube, Facebook, and their email. In the log data, study-related websites were often intertwined with other websites, so it is important to distinguish which activities belong to which task. Although I did not include the time spent on irrelevant websites when extracting the behavioral variables, such distractions could prolong the time spent on the study-related pages due to the extra time spent on task switching and task resumption.

Some participants were distracted by their friends, particularly when they were working in a public place, which was not uncommon in this study. They mostly worked in their apartments, but also in classrooms, libraries, dining halls, or even at work. For example, P12 was with her friends when working on the study and took the time to share her findings with her friends: "*When I was doing the tattoo removal one, I was studying with my friends at the time, so it was nice; whenever I found something new, I would share it with them and they would give me feedback.*" Some of them finished the study in multiple sessions at various places. They chose those places mostly out of convenience and basically grabbed whichever chance they had. For instance, P28 was at work and got spare time, so she decided to work on the study: "*I work as a work-study student at the [removed for anonymity] office. I was just sitting at the desk. They let us do our homework ... I basically took whichever chance I had an extended amount of time to just do whatever I wanted to. It doesn't really matter where.*" Only a few mentioned that they purposefully stayed in their dorm alone so they could finish the tasks quickly.

Most participants treated the remote setting tasks like other everyday life tasks. Because of the flexibility in the remote setting, there was greater variance in times in the remote setting. Participants' behaviors varied depending on where and when they were working on the study. For instance, P26 suggested that he was less patient at the end of the day when he worked on the study: "*it was at the end of the day when I did it so I was tired and I wasn't as motivated to go through it [the study].*" They were also more likely to be distracted and to spend more time when they were with somebody else than when they worked alone at home or in a library. This finding suggests that behavioral data alone may not be sufficient for task predictions in real-life. Information about users' locations and surroundings may help in determining task characteristics.

Although participants were told to spend as much time as they needed, they often felt more rushed in the lab than in the remote setting, like P18 commented: "*I felt like I* *had to rush because I felt like I was going to run out of time*." They suggested that they wanted to "*get it done in a certain amount of time*" (P17) in the lab even when no time limit was given. Such internal time pressure pushed them to work more efficiently. One interesting reason that contributed to this internal time pressure was that they did not want another person to wait. The fact that I needed to wait for them to finish also triggered their desire to work faster. P30 explained that "*at home it's like, I'll be comfortable in my bed, and if I get distracted it doesn't matter because I'm not spending anybody's time here*" while they were more conscious about time when working in the lab because "*you're taking up someone else's time too.*"

The MLM results showed that participants' task completion time was not significantly affected by study settings because its effects on page dwell time and the number of pages visited were in the opposite direction. In other words, while participants, on average, spent a longer time visiting content pages in the remote setting than in the lab, they also visited fewer total pages when working remotely. The descriptive statistics did show that the average task time was longer in the remote setting. There may be a few participants who had too much distraction in the field.

Aside from the amount of time participants spent in different settings, how the time was spent was also worth noting. While participants had to finish everything in one sitting in the lab, about one third of them (n=11) chose to work on at least one of their tasks at multiple time points when working remotely. This did not count temporary distractions such as checking emails and SSNs. If there was a gap longer than 30 minutes in their search log within a task, I determined that the person left the task temporarily and did not complete it in one session.

Several participants divided up their tasks intentionally, like P4, who "*divided the whole task by little chunks ... little bit today, little bit tomorrow, little bit next day.*" There were also participants who spontaneously did not complete a task in one sitting as P12 suggested: "*Since you did give me the option to do it at home, it's much more easier to do it, and then go on other things, and then you did say, Oh, 'you can go back to it any time you need,' so I feel like I would just do it and then I'll finish researching another half another time, so it wasn't really consistent.*" This finding disagrees with Greifeneder's (2016) study in which all but one remote participant completed the whole study in one session. She found the magnitude of distractions in remote sessions to be low and participants were able to resist distractions for the most part. The differences between Greifeneder's results and the results of the current research could be attributed to the different task types used in two studies. Greifeneder asked participants to work on small known-item search tasks in digital libraries while I instructed participants to work on open-ended evaluative tasks that were more exploratory in nature. Evaluative tasks could be more complex and demanding than known-item tasks, and thus could lead participants to finish in multiple sessions. By contrast, participants may be more likely to keep focused in simple known-item tasks. The difference in time allocation between lab and field could be bigger in real life than what the research showed. Although participants were allowed to work on part of the study in the field and were given plenty of time, they still knew that they were working for a study and some of them wanted to finish the tasks soon. Several participants mentioned that they intentionally stayed in their dorms alone so they could get the study done, while they could be on the go when working on real-life tasks. In other words, some participants who finished tasks in one sitting may have done otherwise in reality.

7.2.2 Comparison Between Tasks

In respect to the impact of task authenticity, simulated tasks had a positive effect on average page dwell time and average unique page dwell time. They also negatively affected the number of pages and unique pages visited. In other words, participants spent more time on reading content pages but visited fewer pages and had less diverse page visits in simulated tasks. This partly echoes Liu and Belkin's (2010) finding that users took less time to determine usefulness when they had greater knowledge. A closer examination of the interview data provided an explanation to why participants visited less pages in simulated tasks. Most participants recalled that they stopped working on the simulated tasks once they found similar information on multiple sources or found a site that listed and explained various options (i.e., different options for tattoo removal or lupus treatment). The results of simulated tasks did not directly affect their lives, so they were not motivated to invest a lot of effort, as commented by P34: "*I went on one website and it listed out most of the things … I tried to just go off of the data on that website*." They tried to finish the task by visiting as few sources as possible, which led them to visit less sources than authentic tasks. Also, since the scenarios were not real, participants did not have specific personal context when working on simulated tasks. By contrast, authentic tasks were more specific to them. For example, P1 suggested that she "*would have to consider her [the hypothetical friend in the simulated scenario] situation*" if the task was real, and she was not able to do that for a simulated task. Therefore, reporting the information repeated on multiple website seemed to be a solution, as P30 explained:

"*When I went through a couple websites, it seems they all had almost the exact same answer, so it made me think that I couldn't really find anything different*." When they were asked if they would have done anything differently if the simulated tasks were real, some admitted that they "*would probably have gone through a few more sources.*"

For authentic tasks, most participants had very specific goals and requirements regarding aspects such as budget, geographical locations, or styling. Thus, they were more likely to visit more sources until they were satisfied. For example, a few participants pursued health-related tasks, and they had specific conditions, constraints, or needs, such as doctors in a specific area or financial constraints. However, for the simulated tasks which were also health related, I was not able to provide this level of details (e.g., location, personal conditions). They found it difficult to find a solid answer for the simulated tasks even with the cover story provided because "*it's kinda just one of those things like depends on who, like what your body needs, and everything varies from person to person*" (P21). Participants recalled that they gave more thoughts to the prices when working on real-life tasks due to financial constraints and information about prices was difficult to locate, making authentic tasks more difficult in this aspect. By contrast, they did not really think about prices when working on simulated tasks. P21 recalled that, for authentic tasks, she "*was looking for more price point, but it doesn't have to be the best thing in the world*." However, for the simulated tasks, she "*was just looking for what is the most effective and safe*." In brief, it was good enough for them when they found websites that "*had all the information*" or had similar information for a simulated task. But for authentic tasks, they often needed to consult more sources to satisfy their specific requirements before moving on. Their selection of information seeking problems in the post-task questionnaire also verifies this point. The problem "*the information was scattered*" was selected much more often in authentic tasks (28%) than in simulated tasks (14%), which indicated that they were less likely to find everything needed in one or a few sources for authentic tasks.

In the interviews, I asked the participants to offer their suggestions to improve the simulated task descriptions. Although they did feel that the simulated tasks were less specific and realistic than their own tasks, they did not think that the simulated task descriptions could be improved unless they were personalized to each person (e.g., "*I don't know what more you could have given without literally taking away any variables from our side*", P26). They believed that the simulated task descriptions were as good as they could be.

7.2.3 First Query Measurements & Questionnaire Completion Time

Contrary to my expectations, study settings and task authenticity did not have as much effect on first query measures as on whole-session measures. At least on the first query measures examined in this dissertation, study setting had no significant effects, meaning that we may take users' first query behaviors in the lab as their natural behaviors. Users may start the task in the same way regardless of settings and may be affected more by the setting later in a task session than earlier. The influences of task authenticity somewhat mirror results from the whole-session data. Simulated tasks negatively affected the number of unique pages visited in the first query. They also positively affected the first decision time probably because that participants were working on something unfamiliar and thus took more time to decide which source to click on first.

Participants’ questionnaire completion time was not significantly affected by study settings or tasks for the most part except that the pre-task questionnaire completion time was positively influenced by simulated tasks. This was expected since they were less familiar with the simulated tasks than their own tasks, and thus might take more time to think when completing the pre-task questionnaires.

7.3 Summary

This chapter discusses the quantitative analysis results with evidence found in the interview data. Most participants felt a difference between the lab and the remote settings. They wanted to work harder and be more focused in the lab and found the lab to be more demanding than the remote setting. As a result, they visited more content pages in the lab but spend less average time on pages. This is partly consistent with Borlund et al.’s (2012) finding that the lab may have a ‘pleasing effect’ that pushes users to invest more effort to ‘please’ the researcher. Different from Borlund et al.’s study, this ‘pleasing effect’ was not reflected in overall task completion time. The presence of a researcher also contributed to this laboratory effect as participants did not want to keep another person waiting. When working outside of the lab, participants worked on the study at a variety of places and were often distracted. Finishing a task in multiple sessions was also common, which led to a greater variance in the time spent.

Their task perceptions and engagement were very different between authentic and simulated tasks. Authentic tasks were more specific and personally relevant. This led participants to consult more pages and unique pages. In the simulated tasks, even when provided with scenarios that put them into a context, some participants still did not connect with the simulated tasks like they did in their own tasks due to a lack of knowledge, interest, confidence, etc. They believed that the simulated task situations may potentially happen in the future, but not immediately relevant. It is worth noting that the difference in knowledge, interest, or confidence between tasks were not necessarily attributed to the authenticity of a task, but at least partly to users' tendency of bringing familiar tasks to finish in a study. In real-life task situations, users could also have various levels of knowledge, interest, confidence, etc.

Participants' first-query behaviors were less affected by the study variables than whole-session behaviors. Thus, their first-query behaviors in the lab and simulated tasks may be a better representation of their natural behaviors than whole-session behaviors.

7.4 Limitations

Before discussing the implications and concluding in the next chapter, I would like to acknowledge the study limitations. First, despite the fact that participants were allowed to work on their own tasks and work remotely for part of the study, they were not placed in a purely naturalistic setting. They were given a three-day limit to finish the remote session tasks. Some participants expressed that they wanted to finish the tasks soon and tried to avoid distraction when working remotely since they were in a study and had a time limit. If they were not participating in the study and their authentic tasks were not urgent in reality, they could have postponed them if they were busy. The influences of study settings or task authenticity may be more significant in reality. Also, participants consciously realized that they needed to finish a study so they might stop at a point earlier than when they would have stopped in reality, just to get the study done. Some participants expressed that although three days were plenty, they still felt the need to finish the tasks as soon as they could. It was also occasionally reflected in the search log that participants continued working on the same task after they submitted the post-task questionnaire, meaning that they have not finished the task yet, but they wanted to finish the study first.

Second, there was no guarantee that participants submitted real tasks, though they were advised to work on something that they needed to do even without the study. All participants indicated in the interview that the authentic tasks were something they really needed to do, though not necessarily in an urgent manner.

Third, participants' authentic tasks varied greatly in topics. Therefore, not all authentic tasks are comparable with each other or with the simulated tasks despite that I requested evaluative tasks only. There was a tradeoff between collecting real-life tasks and collecting comparable tasks. The more restrictions I set, the less likely that participants would have submitted real tasks. In order to solicit tasks that participants were actually going to do, I made the choice to not limit task topics, but sacrificed on the comparability. The simulated tasks were also more structured than authentic tasks. I tried to alleviate this limitation by controlling the effects of subjective task characteristics when estimating multilevel models.

Fourth, I only used one type of task in this study, open-ended evaluative tasks, which gave participants flexibility to decide how much information to find. Users' understanding and approaches in open-ended tasks may depend more on their personal experiences and context than in factual tasks. In this research, participants pointed out that it would have been helpful for them to have a more specific context for the simulated tasks. The influence of the study variables may be different for other types of tasks, particularly factual tasks that have definite answers. I plan to test other task types such as fact-finding tasks in future research.

In addition, participants were not allowed to use mobile devices in this study while mobile devices are increasingly used by users to search for information. Some of them habitually put their smartphone by their side when working in the lab but felt obliged to not check it during the lab session. If they were not in the lab, they might have checked their phone or even used it to assist in searching (e.g., checking a dictionary). When asked about mobile device usage, most participants reported that they used smart phones for simple tasks that did not require extensive searching and reading. Most of them believed that they would not have used mobile devices to work on the search tasks for this study even if they were allowed to because of the amount of reading needed. Nevertheless, it would have been possible that some participants preferred to do part of the searching on mobile devices, especially when they were on the go. Future research may also take mobile device usage into consideration.

Chapter 8 : Conclusion

User study is a common method adopted by researchers in interactive information retrieval to collect users' search behaviors and evaluate IR systems. In particular, researchers have devoted considerable research efforts in investigating the relationships between users' online behaviors and other components in a search session, such as tasks and barriers. The design of a user study involves several components that can either be artificially created by researchers or be real parts of users' lives; and users' search behaviors and experiences are directly influenced by the methods employed to conduct a study. Past research in behavioral sciences has discovered the existence of laboratory effects which lead users to behave unnaturally in a controlled lab environment. However, how users look for information and experience search tasks differently between study settings and between authentic and simulated tasks remain unclear.

For this dissertation, I designed a 2x2 repeated-measures study to examine the influences of two essential components of task-based information seeking research design—study setting and task authenticity—on users' online behaviors and task perceptions. Thirty-six undergraduate participants finished 144 tasks (half authentic and half simulated) in two study settings. The results have demonstrated that the synthetic setting and tasks did alter some aspects of users' behaviors—such as page visits and page dwell time—that are frequently used in TBIS research. Users devoted more efforts and worked more efficiently in the lab. Also, it was not possible to observe some common phenomena in real-life searching, such as distractions, multitasking, and other influences of users' context, in the lab. In addition, simulated tasks that were not personalized to users' personal situations prevented them from finding more than generic information, which is a major difference between authentic and simulated tasks. However, first-query measures were not influenced as much as whole-session measures and can be considered as better representations of users' natural behaviors.

Despite the found differences between authentic and synthetic set-ups, the purpose of this study is not to argue that synthetic study components trigger unnatural search behaviors or should not be used. Instead, I tried to understand how and why users behave differently when different methods are employed so future research can account for those differences. The next section summarizes several primary implications for research in task-based information seeking and human-computer information retrieval in general.

8.1 Implications

First, this research has implications for interpreting and using results from studies that collect online search behavioral data. Researchers who conducted lab studies or used simulated tasks may often wonder if the behaviors collected represent users' real behaviors. Past studies have confirmed the existence of laboratory effects but have not provided enough evidence regarding which aspects of user behaviors are affected. From this research, I learned that some aspects, particularly measures related to content page visits (e.g., page dwelling time, number of pages), are affected by the synthetic setting and/or task, while some aspects (e.g., number of queries, query segment time) are not. In other words, we can consider some behaviors collected in synthetic settings or tasks—

such as query segment time, number of queries issued, and query length—to be users' natural behaviors.

Regarding the behaviors that were affected, although the interviews demonstrated that the effects were mostly due to the inherent differences between authentic and synthetic set-ups, the interview data also pointed out directions for reducing those differences. One major difference between authentic and simulated tasks is about the level of specificity and personalization. Participants mostly looked for general information for simulated tasks because the tasks were not personalized to each person. To reduce this difference, Rieh's (2002) approach in utilizing "generic tasks" may serve as an alternative to Borlund's (2000) simulated work task approach. Rieh allowed participants to define part of a task within a certain scenario (e.g., "the next conference that you are going to attend"). In this approach, even though the researchers control the overall task type and topic, they give participants the opportunity to personalize a task, so participants would have their own specific requirements and goals. This kind of tasks are personally relevant to participants, and thus may motivate them to go into more depth. There are also differences that may not be avoided by simply adjusting the study design. For instance, searching in a lab pushed users to work more diligently and efficiently. Future research could consider building models that use the behaviors collected using synthetic set-ups to predict users' behavior in naturalistic settings. This approach allows one to exploit the strength of lab experiments while also observing natural behaviors.

Second, not all behavioral differences were reflected in the measures that have been used by TBIS research. Distraction, multi-tasking, and dividing tasks into parts are prevalent in the field but can rarely be observed in a lab. Although users' overall task completion time was not significantly affected, how the time was allocated differed between settings. There was also a greater variance in times in the remote setting. When mapping users' behaviors to task characteristics, researchers should consider that a task may be mixed with other tasks and it may be paused and continued at a later time. It is crucial to untangle intertwined tasks and track unfinished tasks in real-life searching. Another implication is that users' contextual information, such as locations, may be informative in task predictions. Users' behaviors in a lab could be relatively consistent throughout a task. In real-life scenarios, however, they could work on a task at various places and at different points of time during a day. Their behaviors are affected by these contextual factors. For example, their search behaviors in morning or in a coffee shop may be different from what they do at the end of a day or at a friend's house. Search behaviors alone may not be sufficient for determining task characteristics in real-life scenarios. The knowledge of users' environment can be a useful addition to online behavioral data.

It is worth noting that I observed more interruptions to tasks brought by distractions or multitasking in the field compared to past research, particularly Greifeneder's (2016) comparison between lab and remote studies. The latter used known-item tasks with concrete answers, indicating that the effects of study setting could be different for different types of tasks. An implication for study design is that we may have different designs when different types of tasks are studied. For example, users' behaviors may not be affected significantly by study settings in simple tasks that can be finished quickly or easily. Thus, observing their behaviors in a lab may be sufficient when only simple tasks are studied. However, researchers may want to conduct more naturalistic studies or have at least part of a study done in the field if they want to observe users' behaviors in complex tasks that may span several search sessions or over a period of time. Real-life searching for complex tasks is more likely to be affected by users' real-life context, such as their locations and surroundings, which are almost impossible to be included in a lab study.

Third, users' behaviors in the first query were much less affected by study setting and task authenticity than the whole-session behaviors. In particular, at least the first-query measures used in this study were not affected by study settings at all, and thus may be used to represent users' natural behavior. This is particularly meaningful for researchers who use early session data to run real-time predictions and recommendations because study settings and task authenticity would be less of a concern than when whole-session measures are used.

In addition, I echo Borlund's (2000) suggestion that real information needs can be used as a baseline against simulated tasks in TBIS research and IR evaluation, so researchers would know where differences occur and take those differences into consideration. One additional suggestion would be that future researchers could request users to bring real information needs comparable to the simulated tasks provided since behaviors may vary greatly in different types of tasks. In this study, I requested all participants to bring only evaluative tasks, so the tasks were relatively more comparable to each other than to other types of tasks. Similarly, future research may also ask participants to work on a portion of the study remotely and use that data as their baseline behaviors.

8.2 Future Research

I plan to continue this line of research that investigates the impact of study methods on users' behaviors and experiences and seek solutions to quantify the impact. One future direction is to use the data collected in synthetic settings or tasks to predict how users may behave in reality. Although lab experiments have been criticized for being artificial and not evoking natural behaviors, they have advantages in isolating the effects of one or a few variables and reducing confounding variables. Using behavioral data collected in lab settings to predict users' search behaviors in the field may allow one to combine the strength of lab studies and field studies.

One study is indeed not enough to confirm the influences of study setting or task authenticity. Future research should validate the results of this study and continue building our understanding of the impact of various research methods. Apart from study setting and task authenticity, there are certainly other components to consider such as participant population and time limit. I hope to more comprehensively examine the impact of study design on user behaviors by exploring other study components and task types (e.g., fact-finding tasks) that were not examined in this dissertation.

Appendix A: Recruitment Message

Below is the recruitment email/message that were used to recruit participants on email lists, Facebook groups, and in various buildings on campus. The message for each channel had a slightly different format, but the content was the same.

Come take part in a Web search user study at SC&I!

The InfoSeeking Lab at School of Communication and Information (SC&I) is conducting a study on people's search behaviors on the internet. We are recruiting undergraduate students to participate in the study. The study will take approximately 2-2.5 hours spanning around one week. You will finish four search tasks (two remotely and two in our computer lab), a few questionnaires, and be interviewed at the end of the study. You will receive $40 for your participation. If you are interested in participating in this study, please follow the link to register or for more information:
http://coagmento.org/tbis/signup_intro.php

To be a participant in our study, you need to meet the following requirements:

- You must be at least 18 years old to participate.
- You must be an undergraduate student at Rutgers University.
- You must use Chrome browser on a regular basis.

If you have any questions or need additional information, please email Yiwei Wang at yw498@scarletmail.rutgers.edu.

The study has been approved by the Institutional Review Board (IRB) at Rutgers (protocol #Pro2018000197), and is supervised by Dr. Chirag Shah at School of Communication & Information. Thank you for your interest!

Appendix B: Registration Form

Welcome! This is the sign-up form to register for the research study. The research project, Task-based Search, is supervised by Dr. Chirag Shah in the School of Communication and Information (SC&I).

The study will include two sessions. In one of the two sessions, you will finish two search tasks (finishing tasks through searching for information online) on your personal device (laptop or desktop computers) remotely at a location of your choice (Remote Session). In the other session (Lab Session), you will finish two search tasks in our interaction lab located at School of Communication & Information. The order of these two sessions will be randomly determined; and you will receive detailed instruction after registering. You will also be asked to complete a few questionnaires regarding your task perceptions and experiences. You will need to install a Chrome browser extension on your browser and log into our study system to finish those tasks and questionnaires. At the end of the study, you will be interviewed by a researcher regarding your search experiences. The entire study will take about 2-2.5 hours in a one-week period depending on when your lab session and follow-up interview are scheduled.

Requirements:

- You must be at least 18 years old to participate.
- You must be an undergraduate student at Rutgers University.
- You must use Chrome browser on a regular basis.

Please fill out the following form if you wish to register. You will receive a confirmation email within 24-48 hours with details about the study procedure.

First Name

Last Name

Rutgers Email

Confirm Email

Age

How did you find out about this study? (Facebook, Twitter, Mailing list, etc.)

Gender

Field of Study

Year in College

Is your native language English? If not, what is your native language?

Appendix C: Instruction for Preparing Authentic Tasks

Before you start the main study, you need to describe two search tasks that you plan to work on. Search tasks are the things you are trying to accomplish by searching for information. They can be anything in your daily life, not necessarily related to school or work. They also need to be something that you need to search for in reality, even without being requested by this study. Your search tasks will be reviewed by the research group. After they are approved, you will start the main study.

Please think of two tasks in which you need to 1) evaluate several options belonging to the same category, and 2) make a selection among the options. Examples include but are not limited to comparing different products, solutions, methods of doing something, etc. These two tasks need to fulfill the following requirements:

- For the most part, the information needed to finish the tasks can be found by searching online. For example, tasks that require reading physical books/documents, or communication with other people, do not qualify.
- The tasks should not be too simple, such as selecting a restaurant from several restaurants. Also, it should not be too complex or time-consuming, such as comparing several graduate programs, which would require more than an hour.
- The two tasks should not be related to each other.

Please note that you do not need to search for these two tasks now. You will be given time to work on these two tasks during the main study.

Please describe your first task here:

Please provide more details about this task. Why do you need to do this?

Please describe your second task here:

Please provide more details about this task. Why do you need to do this?

After your tasks are reviewed and approved by a researcher, you will be granted access to the next step. You will be notified by email when your tasks are approved.

Appendix D: Pre-task Questionnaire

Below are the questions asked in the pre-task questionnaire. For each question, participants selected their response on a five-point scale ranging from strongly disagree to strongly agree.

1. I feel this task is difficult
2. I feel this task is complex
3. I am knowledgeable in this topic
4. This task is important to me
5. I am interested in this topic
6. I feel motivated to work on this task
7. I feel confident in gathering the information to complete this task

Appendix E: Post-task Questionnaire

Below are the questions asked in the post-task questionnaire. For all except the last question, participants selected their response on a five-point scale ranging from strongly disagree to strongly agree.

1. I felt this task was difficult
2. I felt this task was complex
3. I felt discouraged while searching
4. I felt frustrated while searching
5. search experience did not work out the way I had planned
6. This search experience was demanding
7. I felt in control of the searching experience
8. I felt interested in this search task
9. My search experience was fun
10. I felt involved in the search task
11. My search experience was rewarding
12. I was really drawn into my search tasks
13. I lost myself in this searching experience
14. I was so involved in my search task that I lost track of time
15. I consider my search experience as a success
16. Have you encountered any difficulties when working on the task? If so, please select all that apply:

- Lack of sufficient patience

- Unable to articulate information needs
- Unaware of relevant information sources (e.g., websites)
- Too much information
- Information was too scattered
- Information was not up-to-date
- Poor quality display of text or graphics
- Information was unreliable
- Time constraints
- Technological problems (e.g., web page errors)
- Too many irrelevant results
- Look at wrong sources
- Unconfident about finding information
- Financial constraints (e.g., need subscription to read a paper)
- Poor search skills
- Unable to understand the information found
- Other (please specify):

Appendix F: Exit Interview Guide

Now that you've finished the entire study, I have a few more questions regarding your experiences with working on the tasks.

1. How was your overall experience of working on this study? Is there anything in particular that you want to talk about? *[Warm-up question]*
2. Where did you finish the remote session? Follow-up: Why did you choose that location?
3. Now, let's talk more about the tasks that you completed remotely. *[Questions were prepared in advance based on their questionnaire responses]*
    - 3.1 *[Discussing each task separately]* In your first task, the task assigned by me, you were asked to [e.g., look for methods for tattoo removal], you indicated in the questionnaire that [e.g., you were unconfident about finding the information], could you talk more about that? Why did you feel [e.g., unconfident]? How did [e.g., feeling unconfident] affect searching?
    - 3.2 *[Discussing each task separately]* In the second task, the task brought by you, you were looking for [e.g., the best sun block for combination skin], you indicated that [e.g., there were too many irrelevant results] could you elaborate on that? How did that problem affect searching?
    - 3.3 Are you satisfied with the information you found out? Why or why not?
    - 3.4 You indicated in the questionnaire that you were more [e.g., motivated] to do your own task than the task I gave you, could you talk a bit about that?

   3.5 If you were working on the assigned task in real life, would you have done anything differently? What would you have done differently?

4. Now, let's move on to the tasks you finished here in our lab.
   4.1.[Repeating 3.1-3.5]
   4.2.Do you think the lab environment affected your search? How?
   4.3.Did you do anything differently when you searched at the locations you chose than when you searched in the lab?
5. In this study, you were restricted to use desktop or laptop computers. Do you often use mobile devices - such as tablets or cellphones - to search? If you were not doing those tasks for my study, would you have used mobile devices?
6. I'd like to know a little more about the two search tasks brought by you. How easy was it for you to think of these two tasks? *[Asking more about the context of their real tasks here if not enough details were provided in their submission.]*
7. Do you have any additional things you'd like to share with me about your experience in this study before we finish speaking?
8. Do you have any questions for me?

For pilot tests only:

- Are there anything about the study procedure or task descriptions that look confusing to you? Anything that we need to clarify?
- Did the study take you more or less time than you expected?
- How do you think about the compensation?

Appendix G: Authentic Tasks Submitted by the Participants

- The first task would be to find a doctor near me, which would require me to search for doctors in the area and select one that covers my insurance. I would also have to make the decision based on where their office is located, what they specialize in, the timings of the office, and the reviews that other people have left about the doctor.
- The second task would be to compare different backpacks while online shopping and selecting the best one for my needs. I would have to consider the brand of the backpack, the features that it has, the amount of space inside, the design, the price, and customer reviews.
- Comparing protein powders.
- I’m looking for a bank that is the best for you to open a savings account.
- Finding different options of travel from Princeton to New Brunswick. I recently moved to a new town and am exploring travel options.
- I'm looking for scholarship/funding opportunities and selecting scholarship/funding programs to apply for.
- I’m looking for ingredients in serums that claim to even skin tone, and I also want to look for serums that have those ingredients.
- Making a selection between different kinds of protein powders.
- Comparing different houses to live in during my college semesters.
- Compare different types of shoes and select a pair to buy.
- I’m looking for fall clothing online.

- I’m looking for a birthday gift for my sister.
- I plan to look for a physician that does lip augmentation because I have been wanting to get dermal lip fillers. I need to compare doctors in terms of patient reviews, their specialties, costs, and distance.
- Choosing what type of phone I should buy to replace my current one.
- Choose exercise to do in order to improve upper body strength.
- Best facial sunscreens for combination skin.
- Low fixed interest rate student loans.
- Finding suitable flight from JFK to Iraq.
- Organ recital near me.
- New tires for my car.
- Determining the over-ear headphones to buy for college students.
- Comparing different credit cards.
- Comparing different speakers. My current Bluetooth speaker is having connection issues so a new speaker, if reasonable, isn’t a bad idea.
- I plan to look for jobs directed for my major that I am qualified for because I am graduating in December and need to start applying around this time.
- Running shoes. To learn the types of brand, shoe styles, cost, etc. especially for running long distance.
- Shopping for a new pair of running shoes. I just joined a running club here on campus and do not have proper shoes.
- Shopping around for a phone to replace my current one.

- Comparing current flagship phone models. (Andriod, apple, blackberry etc.). I would like to know which current flagship phone suits my needs for personal use.
- Bluetooth earbuds.
- Comparing different bicep exercises.
- Searching and comparing ingredients in dairy-free recipes to suit new diet.
- Search for a new smart phone.
- I need to buy new clippers for cutting hair.
- What is the best moisturizer for dry skin?
- What kind of foods should I be eating for thyroid health?
- I would like to search for air pod headphones that I want. I am a runner, and my wired headphones often come out and bother me. A friend recommended getting the air pods, so I was looking for ones that seemed authentic, but were cheaper than the apple ones.
- My first task will be to look for which shampoo and conditioner will be the best for me to purchase in order to begin my natural hair journey.
- My second task will be to look for which iPhone will be the best for me to buy since I am thinking of upgrading. I want to compare the general specs of the phones, the price per gigabytes, the camera quality, and the colors offered.
- I would like to search for the best activities to do while in different cities I will pass through on a road trip next month.
- I’m looking for a certified nursing assistant course in my area.
- What is the best product on the market for acne? I would like to make my complexion better looking, so I will be looking into which products help cure or

lessen acne, taking into account cost and ingredients and reviews of the product to make my decision.

- I will look for a present for my dad's birthday.
- Selecting the cheapest and optimal option to go to Disney land including flight, and Disney land tickets.
- I need to buy a couch for my dorm.
- I am going to check hotels in New York.
- I'm evaluating between mac and PC. I need a new laptop and I've been using pc for a long time so I was thinking of a mac or pc because 90% of college students have mac laptops and I was wondering if they were more convenient.
- What would be good and healthy meals for people with busy schedules.
- I'm looking for concert tickets between multiple websites by considering multiple aspects such as prices and seat locations.
- Compare different dog kennels in my area.
- Comparing skincare products that are suitable for my skin type on different websites such as sephora, ulta, etc.
- Comparing different salons, techniques, and prices to color treat hair.
- I this task, I'm deciding if I should follow a gluten free diet and/or do intermittent fasting
- Figure out what area of finance I want to go into, like corporate or investments.
- Choosing a pair of color contact lenses.
- Buying DIY hair dye.

- Searching for a new laptop case for my Macbook pro. I will factor in safety, cost, color/design, customer reviews, etc.
- Researching best job offer to take.
- Finding flights from NJ to San Dieg, CA from June 23rd/24th.
- Looking for a fannypack.
- Finding recipes for meal prep.
- Finding a nice pair of boots.
- Searching up specific types of jobs at several different locations suit my interests and qualifications.
- Shopping for a handbag.
- I want to look up cybersecurity internships in New Jersey.
- I'm trying to decide if I want to go into the field of Computer Science or Information Technology.
- I want to find the best companies that fit my interests.
- I want to find a new comprehensive workout plan.
- I'm looking for a gym that is low-cost and close to my home.
- I'm looking for a saving's account to invest.
- Searching for an online service to file my taxes.
- To compare TV online for a 46" screen size.

References

Alharbi, A., & Mayhew, P. (2015). Users' performance in lab and non-lab environment through online usability testing: A case of evaluating the usability of digital academic libraries' websites. In *Proceedings of the 2015 Science and Information Conference*, 151-161.

Andrzejczek, C., & Liu, D. (2010). The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience. *Journal of Systems and Software*, *83*(7), 1258–1266.

Aula, A., Khan, R., & Guan, Z. (2010, April). How does search behavior change as search becomes more difficult? In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, 35–44.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379-384.

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

Blomgren, L., Vallo, H., & Byström, K. (2004). Evaluation of an information system in an information seeking process. *Research and Advanced Technology for Digital Libraries: 8th European Conference*, 57–68.

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

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3), 1–34.

Borlund, P. (2016). A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. *Journal of Documentation, 72*(3), 394-413.

Borlund, P., & Dreier, S. (2014). An investigation of the search behaviour associated with Ingwersen's three types of information needs. *Information Processing & Management, 50*(4), 493–507.

Borlund, P., Dreier, S., & Byström, K. (2012). What does time spent on searching indicate? In *Proceedings of the 4th Information Interaction in Context Symposium*, 184–193.

Borlund, P., & Ingwersen, P. (1999, April). The application of work tasks in connection with the evaluation of interactive information retrieval systems: Empirical results. In *Mira Conference*.

Borlund, P., & Schneider, J. W. (2010). Reconsideration of the simulated work task situation. In *Proceeding of the Third Symposium on Information Interaction in Context*, 155–164.

Bowerman, B.L., & O'Connell, R.T. (1990). *Linear Statistical Models: An Applied Approach (2nd ed.)*. Belmont, CA: Duxbury.

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

Bruun, A., Gull, P., Hofmeister, L., & Stage, J. (2009). Let your users do the testing: A comparison of three remote asynchronous usability testing methods. *SIGCHI Conference on Human Factors in Computing Systems*, 1619–1628.

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

Byström, K. (2007). Approaches to "task" in contemporary information studies. *Information Research*, *12*(4), 12-4.

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

Byström, K., & Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology*, *56*(10), 1050–1061.

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

Capra, R., Arguello, J., & Zhang, Y. (2017). The effects of search task determinability on search behavior. *European Conference on Information Retrieval*, 108–121.

Chowdhury, S., Gibb, F., & Landoni, M. (2014). A model of uncertainty and its relation to information seeking and retrieval (IS&R). *Journal of Documentation*, *70*(4), 575-604.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Denvir, C. (2017). Remote control: Evaluating the potential of virtual desktops as a data collection tool in studies exploring how people use the Internet. *International Journal of Social Research Methodology, 20*(5), 533-546.

Faraway, J. J. (2005). *Linear Models with R.* Boca Raton, FL: Chapman & Hall/CRC.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R.* Dordrecht, The Netherlands: SAGE.

Finch, H. W., Bolin, J. E., & Kelley, K. (2014). *Multilevel Modeling Using R.* Boca Raton, FL: CRC Press.

Greifeneder, E. S. (2012). Does it matter where we test? Online user studies in digital libraries in natural environment. Ph.D. Dissertation. Humboldt University of Berlin, Berlin, Germany.

Greifeneder, E. S. (2016). The effects of distraction on task completion scores in a natural environment test setting. *Journal of the Association for Information Science and Technology, 67*(12), 2858-2870.

Gwizdka, J., & Lopatovska, I. (2009). The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology, 60*(12), 2452–2464.

Gwizdka, J., & Spence, I. (2007). What can searching behavior tell us about the difficulty of information tasks? A study of web navigation. In *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1-22.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407.

He, J., & Yilmaz, E. (2017). User behavior and task characteristics: A field study of daily information behavior. In *Proceedings of The ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR),* 67-76.

Hendahewa, C., & Shah, C. (2015). Implicit search feature based approach to assist users in exploratory search tasks. *Information Processing and Management, 51*(5), 643–661.

Hox, J. (2010). *Multilevel Analysis (2nd ed.).* New York, NY: Routledge.

Ingwersen, P. (1986). Cognitive analysis and the role of the intermediary in information retrieval. In *Intelligent Information Systems (Davies, R., editor)*, 206-237. Chichester, West Sussex: Horwood.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR Theory. *Journal of Documentation, 52*(1), 3-50.

Ingwersen, P. (2000). Users in context. In M. Agosti, F. Crestani, & G. Pasi (Eds.), *Lectures on information retrieval. Third European Summer-School, ESSIR, Varenna, Italy* (pp. 157–178). Heidelberg: Springer-Verlag.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Dortrecht, NL: Springer.

Jiang, J., He, D., & Allan, J. (2014, July). Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 607-616).

Jiang, J., He, D., Kelly, D., & Allan, J. (2017). Understanding ephemeral state of relevance. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval (CHIIR)*, 137-146.

Kellar, M., Watters, C., & Shepherd, M. (2007). A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology*, *58*(7), 999–1018.

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

Kelly, D., Arguello, J., Edwards, A., & Wu, W. C. (2015, September). Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (pp. 101-110).

Kelly, D., & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, 377-384.

Kelly, D., & Gyllstrom, K. (2011, May). An examination of two delivery modes for interactive search system experiments: remote and laboratory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1531-1540).

Kim, J. (2008). Task as a context of information seeking: An investigation of daily life tasks on the web. *Libri*, *58*(3), 172–181.

Kim, K.-S., & Allen, B. (2002). Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology*, *53*(2), 109–119.

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

Kumpulainen, S. (2014). Trails across the heterogeneous information environment. *Journal of Documentation, 70*(5), 856–877.

Landsberger, H. A. (1958). *Hawthorne revisited: Management and the worker, its critics, and developments in human relations in industry*. Ithaca, NY: Cornell University.

Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals, and lawyers. *The Library Quarterly*, *66*(2), 161-193.

Li, Y. (2009). Exploring the relationships between work task and search task in information search. *Journal of the American Society for Information Science and Technology, 60*(2), 275-291.

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management, 44*(6), 1822-1837.

Li, Y., & Belkin, N. J. (2010). An exploration of the relationships between work task and interactive information search behavior. *Journal of the American Society for Information Science and Technology*, *61*(9), 1771-1789.

Li, Y., & Hu, D. (2013). Interactive retrieval using simulated versus real work task situations: Differences in sub-facets of tasks and interaction performance. In *Proceedings of the Association for Information Science and Technology Annual Meeting, 50*(1), 1-10.

Levitt, S. D., & List, J. A. (2005). What do laboratory experiments tell us about the real world? Journal of Economic Perspectives, 21(2), 152-174.

Liu, C., Gwizdka, J., Liu, J., Xu, T., & Belkin, N. J. (2010). Analysis and evaluation of query reformulations in different task types. In Proceedings of the *Association for Information Science and Technology Annual Meeting.*

Liu, C., Liu, Y., Gedeon, T., Zhao, Y., Wei, Y., & Yang, F. (2019). The effects of perceived chronic pressure and time constraint on information search behaviors and experience. *Information Processing & Management*, *56*(5), 1667–1679.

Liu, J., & Belkin, N. J. (2010). Personalizing information retrieval for multi-session tasks:

The roles of task stage and task type. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval,* 26-33.

Liu, J., & Belkin, N. J. (2015). Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. *Journal of the Association for Information Science and Technology*, 66(1), 58-81.

Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., ... & Zhang, X. (2010). Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital libraries*, 69-78.

Liu, J., Kim, C. S., & Creel, C. (2015). Exploring search task difficulty reasons in different task types and user knowledge groups. *Information Processing & Management*, *51*(3), 273-285.

Liu, J., Liu, C., & Belkin, N. J. (2016). Predicting information searchers' topic knowledge at different search stages. *Journal of the Association for Information Science and Technology, 67*(11), 2652-2666.

Liu, J., Liu, C., Yuan, X., & Belkin, N. J. (2011). Understanding searchers' perception of task difficulty: Relationships with task type. In *Proceedings of the Association for Information Science and Technology Annual Meeting*.

Liu, J., Mitsui, M., Belkin, N. J., & Shah, C. (2019). Task, information seeking intentions, and user behavior: Toward a multi-level understanding of Web search. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*, 123-132.

Liu, J., Wang, Y., Mandal, S., & Shah, C. (2019). Exploring the immediate and short-term effects of peer advice and cognitive authority on Web search behavior. *Information Processing & Management, 5*6(3), 1010-1025.

Lucchese, C., Orlando, S., Perego, R., Silvestri, F., & Tolomei, G. (2011). Identifying task-based sessions in search engine query logs. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 277-286.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science, 40*(1), 54-66.

Mitsui, M**.,** Liu, J., & Shah, C. (2018). How much is too much? Whole session vs. first query behaviors in task prediction. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'18).

Nakagawa, S., & Schielzeth H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecology and Evolution, 4,* 133-142.

O'Brien, H. L., & Lebow, M. (2013). Mixed-methods approach to measuring user experience in online news interactions. *Journal of the American Society for Information Science and Technology, 64*(8), 1543-1556.

O'Brien, H. L., & Toms, E. G., (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology, 59*(6), 938-955.

Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776-783.

Pharo, N. (2004). A new model of information behavior based on the search situation transition schema. *Information Research, 10*(1).

Pharo, N., & Järvelin, K. (2004). The SST method: a tool for analysing Web information search processes. *Information Processing & Management, 40*, 633-654.

Poddar, A., & Ruthven, I. (2010). The emotional impact of search tasks. In *Proceedings of the Third Symposium on Information Interaction in Context.*

Reid, J. (2000). A task-oriented non-interactive evaluation methodology for information retrieval systems. *Information Retrieval, 2*(1), 115-129.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology, 53*(2), 145-161.

Roethlisberger, F. J., Dickson, W. J., and Wright, H. A. (1975). Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago. Cambridge, MA: Harvard University Press.

Saastamoinen, M., & Järvelin, K. (2017). Search task features in work tasks of varying types and complexity. *Journal of the American Society for Information Science and Technology*, *68*(5), 1111–1123.

Saastamoinen, M., Kumpulainen, S., & Järvelin, K. (2012, August). Task complexity and information searching in administrative tasks revisited. In *Proceedings of the Fourth Information Interaction in Context Symposium, Nijmegen, The Netherlands* (pp. 204–213).

Saldaña, J. (2015). The Coding Manual for Qualitative Researchers. London, UK: SAGE Publications.

Savolainen, R. (1995). Everyday life information seeking: Approaching information seeking in the context of "way of life". *Library & information science research*, *17*(3), 259-294.

Soltani, D., Mitsui, M., & Shah, C. (2019, March). Coagmento: Rapid prototyping of Web search experiments. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)* (pp. 367–371).

Takahashi, L., and Nebe, K. (2019). Observed differences between lab and online tests using the AttrakDiff semantic differential scale. *Journal of Usability Studies, 14*(2), 65–75.

Tajia, S., & Nyce, J. M. (2015). The problem with problematic situations: Differences between practices, tasks, and situations as units of analysis. *Library and Information Science Research, 37*(1), 61-67.

Toms, E. G., Freund, L., & Li, C. (2004). WiIRE: The web interactive information retrieval experimentation system prototype. *Information Proceeding and Management, 40*(4): 655-675.

Vakkari, P. (2001). A theory of the task-based information retrieval process: A summary and generalization of a longitudinal study. *Journal of Documentation, 57*, 44-60.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology*, *37*(1), 413–464.

Wang, Y., Liu, J., Mandal, S., & Shah, C. (2018). Persuasion by peer or expert for Web search. In *Companion of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW),* 225-228.

Wang, Y., & Shah, C. (2017). Investigating failures in information seeking episodes. *Aslib Journal of Information Management, 69*(4), 441-459.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing (2nd ed.).* Burlington, MA: Elsevier.

Wildemuth, B.M., & Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, New York, NY.

Wildemuth, B., Freund, L., & Toms, E. (2014). Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation, 70*(6), 1118-1140.

Wirth, W., Sommer, K., von Pape, T., & Karnowski, V. (2016). Success in online searches: Differences between evaluation and finding tasks. *Journal of the Association for Information Science and Technology*, *67*(12), 2897–2908.

Xie, I. (2009). Dimensions of tasks: Influences on information-seeking and retrieval process. *Journal of Documentation, 65*(3), 339-366.

Wilson, T. D. (2000). Human information behavior. *Informing science*, 3(2), 49-56.

Vuong, T., Saastamoinen, M., Jacucci, G., & Ruotsalo, T. (2019). Understanding user behavior in naturalistic information search tasks. *Journal of the Association for Information Science and Technology*.

Wu, Z., Mao J., Liu, Y, Zhang, M, and Ma, S. (2019, February). Comparing image search user behavior on lab study and field study task settings. In *Proceedings of 12th ACM International Conference on Web Search and Data Mining.*

Zhang, Y., & Gwizdka, J. (2014). Effects of tasks at similar and different complexity levels. In *Proceedings of the Association for Information Science and Technology Annual Meeting*.

Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J. M., & Azzopardi, L. (2013). Crowdsourcing interactions: Using crowdsourcing for evaluating interactive information retrieval systems. *Information Research, 16*(2): 267–305.