Generalizability and Reproducibility of Search Engine Online User Studies

Zijian Xu

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science and Applications

Jiepu Jiang, Chair

Sang Won Lee

Kurt Luther

May 13, 2020

Blacksburg, Virginia

Keywords: Interactive information retrieval, online user studies, generalizability, reproducibility.

Generalizability and Reproducibility of Search Engine Online User Studies

Zijian Xu

(ABSTRACT)

Research in interactive information retrieval (IR) usually relies on lab user studies or online ones. A key concern of these studies is the generalizability and reproducibility of the results, especially when the studies involved only a limited number of participants. The interactive IR community, however, does not have a commonly agreed guideline regarding how many participants should recruit. We study this fundamental research protocol issue by examining the generalizability and reproducibility of results with respect to a different number of participants using simulation-based approaches. Specifically, we collect a relatively large number of participants' observations for a representative interactive IR experiment setting from online user studies using crowdsourcing. We sample smaller numbers of participants' results from the collected observations to simulate the results of online user studies with a smaller scale. We empirically analyze the patterns of generalizability and reproducibility regarding different dependent variables and draw conclusions related to the optimal number of participants. Our study contributes to interactive information retrieval research by 1) establishing a methodology for evaluating the generalizability and reproducibility of results, and 2) providing guidelines regarding the optimal number of participants for search engine user studies.

Generalizability and Reproducibility of Search Engine Online User Studies

Zijian Xu

(GENERAL AUDIENCE ABSTRACT)

In the domain of Information Retrieval, researchers or scientists usually require human participants to interact, test and evaluate a novel system, which is usually called user studies. However, researchers usually perform these studies with small sample size, some of them recruited fewer than 20 participants, which casts doubt on the generalizability and reproducibility of these studies. Generalizability means how reliable the results of relatively small sample size in an experimental setting can be generalized to the outcomes of a larger population. Reproducibility means whether the results from two groups with the same amount of sample size are consistent with each other. In order to examine the generalizability and reproducibility of online user studies in interactive information retrieval systems, we conducted an online user study with large sample size. We reproduced a well-recognized lab user study from Kelly et al. (2015) in an online environment. We established a simulation-based methodology for evaluating the generalizability and reproducibility of the results and then provided guidelines regarding the optimal number of participants for search engine user studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In Information Retrieval, researchers usually use lab or online user studies to examine human factors in search engines and evaluate interactive information retrieval systems. Most of these studies included relatively small numbers of human subjects, which usually ranges between 10 to 70. Therefore, we are concerned about the generalizability and reproducibility of these user studies. Also, recent research lacks sufficient understanding and powerful control regarding the relationship between the appropriate sample size and the generalizability and reproducibility of the findings. Sakai (2016) questioned the results of studies with limited sample size. In order to examine the generalizability and reproducibility of user studies in interactive information retrieval systems, we conducted an online user study with a relatively large sample size. We randomly sample results from the collected large sample to simulate experiment results with relatively smaller samples such as to help us understand the generalizability and reproducibility of results.

We reviewed articles published in two main conferences for interactive information retrieval research—SIGIR[1] (from 2014–2018) and CHIIR[2] (from 2017–2018)—and found that the most popular-used experimental design in these user studies is to employ search task type as the independent variable so that researchers can examine the user behaviors and experiences and evaluate the systems under different types of tasks. However, most literature did not specify how their search tasks are devised, so it is difficult for other researchers to reproduce

[1] ACM SIGIR Conference on Research and Development in Information Retrieval.
[2] ACM SIGIR Conference on Human Information Interaction and Retrieval.

the results.

The objective of this thesis is to specifically examine the generalizability and reproducibility through the following research questions:

1. Are the results from the online user study and lab user study consistent or not?

2. What are the trends of generalizability of results in online user study? What is the optimal number of participants for online search engine user studies in generalizability?

3. What are the trends of reproducibility of results in online user study? What is the optimal number of participants for online search engine user studies in reproducibility?

In order to solve the research questions above, we reproduce a representative lab user study using different types of tasks as independent variables to understand different search behaviors and user experience (Kelly et al., 2015) in an online environment. It includes 10 search tasks from 5 cognitive complexity levels (*Remember*, *Understand*, *Analyze*, *Evaluate* and *Create*), and the 2 search tasks in each level are related two topics (Science and Health). We collected user behaviors and asked for user experiences as the dependent variables. We recruited 500 subjects from Amazon Mechanical Turk, which is 100 subjects per cognitive complexity level. This allows us to collect a relatively large number of observations for this experiment setting.

After obtaining user behaviors and their self-reported search experience measures, we use a simulation approach to examine the generalizability and reproducibility of results for online search engine user studies. We randomly sample (without replacement) a small number of participants' results from the collected experiment results and examine how likely the findings based on the small sample (e.g., whether or not task complexity affects the number of search queries in a session) is consistent with those based on the whole collected experiment results—

this helps us understand the possibility of generalize the findings from a relatively small number of participants to a larger number. In addition, we also examine the reproducibility of results for a relatively small number of observations using a similar simulation approach. We randomly sample two small samples of users' data from the collected whole sample and examine how likely they agree with each other. We explain our methodology in detail in Section 3.6.2 and 3.6.3.

We found that, when the number of participants is around 85, the generalizability "error rate" and reproducibility "error rate" drop below 10% for most of the dependent variables, which means if a study has 85 participants, its results are more likely to be reproduced and can generalize to the outcomes of a large population for 90% of the cases. It should be noticed that changes in error rates for certain dependent variables did not follow the general trend, that is, they did not reduce to zero at the maximum sample size in the current study. We propose that this is because the current sample size is still insufficient despite the fact that it is greater than the ones in previous studies; more participants are still required to clarify the trends for these dependent variables.

Our study is the first to experimentally examine the generalizability and reproducibility of results in interactive IR user studies, which may potentially have a profound contribution to interactive information retrieval research, especially to help researchers understand the relationship between sample size and the risks of drawing conclusions.

Chapter 2

Related Work

For the discipline of Information Retrieval (IR), a substantial amount of systems needs user interactions or human behaviors to evaluate if the system can help achieve the wanted information. One of the important concerns for researchers is that whether the experimental results are reliable and reproducible in the larger population. It is necessary to explore and summarize the most popularly used experimental design among the lab user in order to examine the reproducibility and reliability of human experiments in interactive information retrieval systems. We reviewed the user studies in the papers from SIGIR(2014 - 2018) and CHIIR(2017 -2018).

2.1 User Studies

2.1.1 Participants

Generally speaking, the number of participants in these lab user studies are limited. The smallest number of participants is 10 (Garcia-Gathright et al., 2018; J. Wang & Komlodi, 2018) and the largest sample size is 72 (Ong et al., 2017). The majority of these studies recruited 20 to 55 participants, which casts doubts on the reliability and generalizability of these studies. Sakai (2016) mentioned in a systematic review of SIGIR full papers and TOIS papers from 2006 to 2015 that: 1) a sample size of 28 participants (Arguello et al.,

2012) is too small to investigate the user experience sub-scale in two systems; 2) the result from an experiment with 24 participants (Ke et al., 2009) is underpowered in comparing two algorithms. Except for those research that focuses on evaluating individual's behavior, larger sample size is likely to make the results more representative of the entire population and to get a more generalizable conclusion. Therefore, we concern that the results of lab user studies with 15 to 40 participants might be unreliable and difficult to be reproduced with a different population.

2.1.2 Independent Variables

Independent variables usually depend on the purpose of the specific study. For information retrieval studies involving lab users, researchers usually manipulate interfaces to evaluate and test new algorithms or systems , such as interactive multilingual search interfaces (Ling et al., 2018), interfaces with different snippet length (Maxwell et al., 2017), and intelligent assistants (Kiseleva et al., 2016). Therefore, interface can is shown to be one of the widely employed independent variables in lab user studies. In addition to the interfaces, search task is another frequently used independent variable in lab user studies; researchers can then approve if the systems or algorithms are justifiable. More details about search tasks will be discussed in Section 2.2.

2.1.3 Dependent Variables

Because most lab user studies implemented searching tasks for the research, user behaviors are collected as the dependent variable. User behaviors are the interactions between participants and the tested interfaces or systems, usually including the number of search queries, the number of clicks, and the number of information viewed by the participants. These

aspects recorded by user behaviors are fact-based information that can be directly extract from user's interaction with the system. Researchers also require another type of information such as how engaged people feel while interacting with the system, to what extent they are satisfied with the searching experience and the solution, and whether they experience any difficulty during the interaction. This type of information relates to the participant's subjective opinions; questionnaires will be given or interviews will be organized to obtain information around these topics.

2.1.4 Within-subject vs. Between-subject

In these lab user studies, both Within Subjects Design and Between Subjects Design were used depending on the particular research purpose. For those with within subjects design, participants were asked to perform the experiments on all the types of tasks, systems or interfaces; researchers then compare the different behaviors of one individual across different types of tasks. For the between-subjects studies, researchers normally separate the participants into different groups, and participants in the same group were asked to perform a particular number of tasks, systems or interfaces. Researchers then compare their data across different groups. The choice of the design is dependent on the topic of the research.

2.2 Search Tasks

From all these user studies, most of them used search tasks in order to reach their purposes. Most of these search tasks can be summarized into different complexity levels, such as simple and complex (Sarrafzadeh & Lank, 2017), fact-finding and exploring (Hienert et al., 2018; J. Kim et al., 2017; Kotzyba et al., 2017; Luo, Li, et al., 2017; Y. Wang et al., 2018; Xie

et al., 2017). However, there is no specific standard for these search tasks. Here are two papers, which categorized the search tasks into different levels or types.

Wu et al. (2012) devised the tasks into the 5 levels according to its cognitive complexity, which are Remember, Understand, Analyze, Evaluate and Create level. Remember level is to retrieve, recognize or recall relevant information from participant's long-term memory. Understand level is to search for the answers based on some specific messages or information. Analyze level is to separate, differentiate, and reorganize the materials. Evaluate level is to make judgments on a particular topic; and Create level is to put all elements together to plan and form a complete result. In general, the search tasks in Remember level and Understand level are relatively simple such as fact-finding tasks that are with little cognitive demands. The search tasks in the Analyze, Evaluate and Create levels are relatively difficult as they need the participants to explore the given topic and reorganize the obtained information.

The other paper that explained their categorization of the tasks is done by Li and Belkin (2008).They reviewed the tasks used by other researches in the literature and categorized them into two main categories: work task and information-related task. Work tasks required people to compute information or to work on the information they obtained, and information-related search task is to give the fact of a particular topic. The author also suggested that different aspects of work tasks may influence people's behaviors. Although Li and Belkin (2008) defined their categorization of tasks, their method was not as generalizable as the one devised by Wu et al. (2012). It is still task-dependent and the two categorization levels are strongly inter-related.

2.3 List of Papers

Table 2.1, 2.2, 2.3, 2.4, 2.5 and 2.6 listed a few papers published in the past five years (2014 - 2018) investigating user behaviors within the domain of information retrieval. The majority of the research is set in a lab condition. Sample size is also reported ranging from 10 to 72 participants. The corresponding independent and dependent variables are summarized too.

Table 2.1: Summaries of User Studies

Author/Year	Setting	Design	Sample Size	IV	DV
Y. Wang et al. (2018)	Online	Within	53	Four types of tasks	User experiences
Ling et al. (2018)	Lab	Within	25	1) Four Multilingual Search Interfaces, 2) three tasks of three different types per interface	1) User preferences, 2) User behaviors
Hienert et al. (2018)	Lab	Within and Between	40	1) Two types of task topics, 2) Two types of tasks per topic	1) User behaviors, 2) User experiences
Ghosh et al. (2018)	Lab	Within	31	Four types of tasks	1) User behaviors, 2) User experiences
Trippas et al. (2018)	Lab	Between	26	Observational information-seeking process	Observational interaction between user and intermediary
Avula et al. (2018)	Lab	Between	54	1) Three search tasks, 2) Three searchbot conditions	User experiences

Table 2.2: Summaries of User Studies Contd.

Author/Year	Setting	Design	Sample Size	IV	DV
Dinneen et al. (2018)	Lab	Between	62	1) Topic tree, 2) Ten different topics of tasks	1) User behaviors, 2) User Experiences
Salminen et al. (2018)	Lab	Within	29	Six different type of persona profiles	User behaviors
H. Zhang et al. (2018)	Lab	Within	60	1) A search interface, 2) Twelve search tasks	User behaviors
J. Wang and Komlodi (2018)	Lab	Between	10	Searching code-switching	User experiences
Klouche et al. (2017)	Lab	Within	20	1) Two types of tasks, 2) Two systems	1) User behaviors, 2) Quality of task outcomes
J. He and Yilmaz (2017)	Diary	Between	23	Daily search on website	User behaviors
Hoeber et al. (2017)	Lab	Between	14	3 search tasks	1) User behaviors, 2) User experiences
Kotzyba et al. (2017)	Lab	Within	19	14 search tasks	1) User behaviors, 2) User experiences
J. Kim et al. (2017)	Lab	Within	24	1) 12 search tasks, 2) 3 different snippet lengths	1) User behaviors, 2) User experiences
Singh et al. (2017)	Lab	Within	25	Structure and content in the give dataset	Users' evaluation
Luo, Li, et al. (2017)	Lab	Within	50	9 ad-hoc search tasks	1) User behaviors, 2) User experiences

2.4 Crowdsourcing

Although most of the user studies in the previous years are lab user studies, however, we conducted an online user study by using crowdsourcing, therefore the reliability and generalizability of results by using crowdsourcing is an important part to research. Amazon Mechanical Turk is a relatively new and reliable website for crowdsourcing, researchers can obtain more diverse data from different target group than college students (Buhrmester et al., 2011; Casler et al., 2013). In Amazon Mechanical Turk, researchers can post tasks (i.e. Human Intelligence Tasks) to the workers, and when workers finished the task, they can get the reward from researchers. This payment is usually cheaper than the payment in the lab user study (Callison-Burch, 2009), however, the lower payment did not affect the quality of data (Buhrmester et al., 2011; Rouse, 2015), therefore, we do not need to set a large payment to get better quality data. There is one possible thing can affect the quality of results by using Amazon Mechanical Turk, which is participants' attentiveness. Rouse (2015) mentioned the quality of data is more reliable when participants put more attentiveness during the tasks. However, Thomas and Clifford (2017) said online participants are as attentive as the participants in lab user studies. Another advantage by using Amazon Mechanical Turk is researchers can collect data faster than conducting lab user studies (Callison-Burch, 2009). When researchers use online fast survey, the accuracy is similar to the results in large traditional surveys (Bentley et al., 2017).

In many human factor experiments, there are no differences in the results of human involving tasks between study with Amazon Mechanical Turk participants and laboratory participants (Casler et al., 2013; Horton et al., 2011). Komarov et al. (2013) conducted an experiment in evaluating the user interfaces from online user study and lab user study, and they found Amazon Mechanical Turk is a productive setting for evaluation of user interfaces. From our

research in the search task from the previous section, the method of categorizing is from cognitive process dimensions, which is from psychology. Therefore, we wondered can the data from Amazon Mechanical Turk perform better or not. H. S. Kim and Hodgins (2017) mentioned the self-report data from psychology aspect (e.g. gambling population, cannabis, alcohol) are higher quality. Stewart et al. (2017) mentioned the quality of data from Amazon Mechanical Turk is much better and the results are reproducible.

In conclusion, Amazon Mechanical Turk is a more productive platform from human factor studies. However, we still need to care about the design of the task, because the results from Amazon Mechanical Turk can be changed if researchers do not design well (Kittur et al., 2008).

2.5 Generalizability and Reliability in Information Retrieval

There are some researchers investigate the generalizability and reliability in Information Retrieval Community. Buckley and Voorhees (2000) introduced a methodology of evaluating the stability and suggested the optimal number of queries for web measurements. In another paper, Voorhees and Buckley (2002) introduced a swap method to calculate the error rate of relevant documents between TREC data and gave the suggested topic set size for information retrieval experiments. Different from the two methods before, an improved method of the one (Voorhees & Buckley, 2002) was in use in (Sanderson & Zobel, 2005) to replicate the results in (Voorhees & Buckley, 2002). Different from the swap method, Sakai (2006) introduced a Bootstrap-based method to investigate the sensitivity of Information Retrieval metrics, however, he found the similar results by using two different methods.

Our method of examining the generalizability and reproducibility is mainly from (Voorhees & Buckley, 2002) and more details are in Section 3.6.2 and 3.6.3.

Table 2.3: Summaries of User Studies Contd.

Author/Year	Setting	Design	Sample Size	IV	DV
Jiang, He, Kelly, et al. (2017)	Lab	Within and Between	28	1) Four search tasks, 2) Two stages per task	User judgements
Huang et al. (2018)	Lab	Within and Between	60	1) Three systems, 2) Ten MER tasks per system	1) User behaviors, 2) User experiences
Lu et al. (2018)	Lab	Within and Between	32	Eleven news reading tasks	1) User behaviors, 2) User experiences
F. Zhang et al. (2018)	Lab	Within	36	Twelve images search tasks	1) User behaviors, 2) User experiences
Su et al. (2018)	Lab	Within	20	Ten tasks of four different difficulty levels	1) User behaviors, 2) User experiences
Garcia-Gathright et al. (2018)	Lab	Within	10	A music discovery system	User experiences
Edwards and Kelly (2017)	Lab	Within	40	Eight search tasks	1) User behaviors, 2) User experiences
Maxwell et al. (2017)	Lab	Within	53	1) Four search interfaces, 2) Search tasks	1) User behaviors, 2) User experiences
Sarrafzadeh and Lank (2017)	Lab	Within and Between	47	1) Two interfaces, 2) Two search tasks	1) User behaviors, 2) User experiences
Harvey and Pointon (2017)	Lab	Within and Between	24	1) Two devices, 2) Two level of distractions, 3) Search tasks	1) User behaviors, 2) User experiences
J. Y. Kim et al. (2017)	Lab	Between	60	Two conditions of opt-in client application	User behaviors

Table 2.4: Summaries of User Studies Contd.

Author/Year	Setting	Design	Sample Size	IV	DV
Xie et al. (2017)	Lab	Within and Between	46	1) Twenty-one search tasks, 2) Two search stages	User behaviors
Ong et al. (2017)	Lab	Between	72	1) Six search tasks, 2) Two devices	User behaviors
Jiang, He, and Allan (2017)	Lab	Within	28	1) Twenty-Eight Search tasks of four different types, 2) Two different scenarios of judgements	1) User behaviors, 2) User experiences
Luo, Liu, et al. (2017)	Lab	Within	43	1) Twenty ad-hoc Search tasks of four different types, 2) Four SERPs from four major commercial search engines per task	User behaviors
Luo, Liu, et al. (2017)	Lab	Within	68	Collaborative tasks	User experiences
Kiseleva et al. (2016)	Lab	Within	60	1) Eight tasks, 2) Three search dialogues	1) User behaviors, 2) User experiences
Moshfeghi et al. (2016)	Lab	Within	24	1) Two types of tasks, 2) Two scenarios	Brain activity revealed by BOLD signal
Burton and Collins-Thompson (2016)	Lab	Between	44	Twelve search tasks of four topics	1) User behaviors, 2) User experiences

Table 2.5: Summaries of User Studies Contd.

Author/Year	Setting	Design	Sample Size	IV	DV
Meier and Elsweiler (2016)	Online	Between	44	Twitter daily use	User behaviors
Umemoto et al. (2016)	Lab	Within	44	1) Search tasks of four topics, 2) Two different ScentBar	1) User behaviors, 2) User experiences
Y. Liu et al. (2016)	Lab	Within	35	Thirty search tasks	User behaviors
Capra et al. (2015)	Lab	Between	22	1) Ten search tasks, 2) Participants who accepted high-quality search query training or not	User behaviors
Capra et al. (2015)	Lab	Between	36	1) Search tasks, 2) Three interfaces with different number of results per page	1) User behaviors, 2) User experiences
Z. Liu et al. (2015)	Lab	Within	35	1) Thirty search tasks, 2) Five popular vertical types of interfaces	User behaviors
Barreda-Ángeles et al. (2015)	Lab	Within	19	1) Four search tasks, 2) Four levels of search latency	User Experience
J. He et al. (2015)	Lab	Within	49	Search tasks	1) User behaviors, 2) User experiences
Y. Liu et al. (2015)	Lab	Within	40	Thirty search tasks	1) User behaviors, 2) User experiences

Table 2.6: Summaries of User Studies Contd.

Author/Year	Setting	Design	Sample Size	IV	DV
Turpin et al. (2015)	Lab	Between	51	Selected documents from the given topics	Judgements of the documents
X. Liu et al. (2015)	Lab	Between	51	OER relevance judgments	Satisfaction scores
Azzopardi (2014)	Lab	Between	36	1) Three interfaces, 2) Three search tasks for one participant	User behaviors
Arapakis et al. (2014)	Lab	Within	12	1) Twelve level of search latency, 2) Two levels of search site speed	User experiences
Arapakis et al. (2014)	Lab	Within	20	1) Four search tasks, 2) Different latency value for each task	User experiences

Chapter 3

Experiment Design

Our purpose is to examine the reproducibility and reliability of human experiments in interactive information retrieval systems. An online between-subjects user study was conducted to investigate whether results from a smaller sample of participants can generalize to the results from a larger sample. A total of 500 subjects from Amazon Mechanical Turk were recruited, Our web search system assigned each participant to a particular search task; the participant needs to perform the following actions in order: preview the task, fill out a pre-task questionnaire, complete the task on our system and finish a post-task questionnaire.

3.1 Search Task

The current user study used the tasks from the work by Kelly et al. (2015), since it summarized most types of search tasks from SIGIR(2014 - 2018) and CHIIR(2017-2018), and these tasks were frequently used in the literature (Capra et al., 2015; Ghosh et al., 2018).

There are five cognitive levels: remember, understand, analyze, evaluate and create level. Each cognitive level has two tasks: one is in the health domain and the other is at the science and technology level. The definitions and meanings of the cognitive levels are listed below:

1. Remember level: Previous experience and knowledge are stored as people's long-term memory so as to be retrieved and manipulated when they recognize a similar topic.

2. Understand level: People need to use strategies such as classifying and summarizing to construct the general meaning of the questions in order to search for the answers

3. Analyze level: Based on their understanding, people need to disintegrate the questions and to reorganize each part according to their inter-connections in order to find out the overall purposes and answers for the question.

4. Evaluate level: People need to critically inspect and assess the materials and to make judgement on it.

5. Create level: People need to manipulate the separated parts to reproduce a comprehensible entity by developing, calculating, or constructing.

Four conclusions from the lab user study (Kelly et al., 2015) that are relevant to the current research are summarized here: 1) The overall trends of mean search behaviors increase when the cognitive complexity level increases, except query length; 2) Remember level tasks were less complex than others; 3) Remembers level tasks were the easiest across all the items; and 4) Create level tasks and evaluate level tasks have more engaging than the remember tasks.

In the experiment, participants can choose the cognitive level they are interested in and the web search engine will randomly assign one task from the two domains. Table 3.1 and 3.2 presents the ten tasks used in the experiments.

3.2 System

The system used in this study is an online web search engine system. It has seven pages: 1) collecting participant's personal information; 2) presenting the search task for participants; 3) a pre-task questionnaire page; 4) an instruction page explaining this web search engine to the

Table 3.1: Lists of Search Tasks (Kelly et al., 2015)

	Health	Science and Technology
Remember	You recently watched a documentary about people living with HIV in the United States. You thought the disease was nearly eradicated, and are now curious to know more about the prevalence of the disease. Specifically, how many people in the US are currently living with HIV?	You recently watched a show on the Discovery Channel, about fish that can live so deep in the ocean that they're in darkness most or all of the time. This made you more curious about the deepest point in the ocean. What is the name of the deepest point in the ocean?
Understand	Your nephew is considering trying out for a football team. Most of your relatives are supportive of the idea, but you think the sport is dangerous and are worried about the potential health risks. Specifically, what are some long-term health risks faced by football players?	You recently became acquainted with one of the farmers at the local farmer's market. One day, over lunch, he was on a rant about how people are ruining the soil. He was clearly upset, so you're interested in finding out more. What are some human activities that degrade soil fertility?
Analyze	Having heard some of the recent reports on risks of natural tanning, it seems like a better idea to sport an artificial tan this summer. What are some of the different types of artificial tanning methods? What are the health risks associated with each method?	You recently became involved with a conservation group that picks-up trash from local waterways. One of the group members told you that your work was important because it helps keep pollution out of the ocean. What are some of the different types of ocean pollutants? What environmental risks are associated with each pollutant?
Evaluate	One of your siblings got a spur of the moment tattoo and now regrets it. What are the current available methods for tattoo removal, and how effective are they? Which method do you think is best? Why?	You have noticed that online services such as Facebook have replaced face-to-face interactions. You can see the advantages of this style of communication, but your sibling argues that people are losing their ability to communicate face-to-face. In general, does use of computers for communication have a positive or negative impact on people's face-to-face social skills?

Table 3.2: Lists of Search Tasks Contd. (Kelly et al., 2015)

	Health	Science and Technology
Create	Your great granny's doctor has told her that getting more exercise will increase her fitness and help her avoid injuries. Your great granny does not use the Internet and has asked you to create an exercise program for her. She is 90-years old. Put together two thirty-minute low-impact exercise programs that she could alternate between during the week.	After the NASCAR season opened this year, your niece became really interested in soapbox derby racing. Since her parents are both really busy, you've agreed to help her build a car so that she can enter a local race. The first step is to figure out how to build a car. Identify some basic designs that you might use and create a basic plan for constructing the car.

participant; 5) a web search page; 6) a post-task questionnaire page; and 7) a confirmation page. The most important pages are the instruction page and the web search page. Figures of the instruction page and the web search page are shown below, and the figures of other interfaces of our systems are shown in the Appendix A.

3.2.1 Instruction Interface

As shown in Figure 3.1 and 3.2, the instruction page informs about how to use the web search page later. Participants can type their answers in the text box provided (more details will be shown in Section 3.2.2). Time restriction is set for each cognitive level according to its difficulty, meaning that the users must stay on the web search page for a minimum time. According to our pilot user study with 100 participants, task engagement and the quality of the performance decreases as the cognitive level increases. Therefore, the time restriction is hoped to help avoid this situation and to control the task quality by making sure that participants actually do the related search on the given topic. From the remember level to the create level, the minimum-time restrictions are 2 minutes, 4 minutes, 7 minutes, 8 minutes, and 9 minutes. The particular number of the time restriction is consistent with

Kelly et al. (2015). Moreover, the users are informed that we will manually check the path as they perform the task; it is also told that they will have the chance to receive a bonus if they complete the task perfectly.

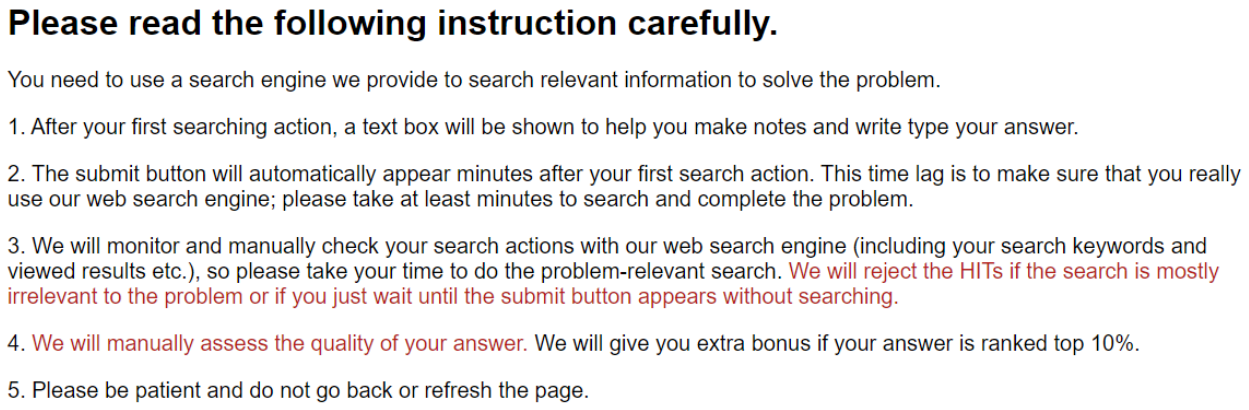


Figure 3.1: Instruction page part I

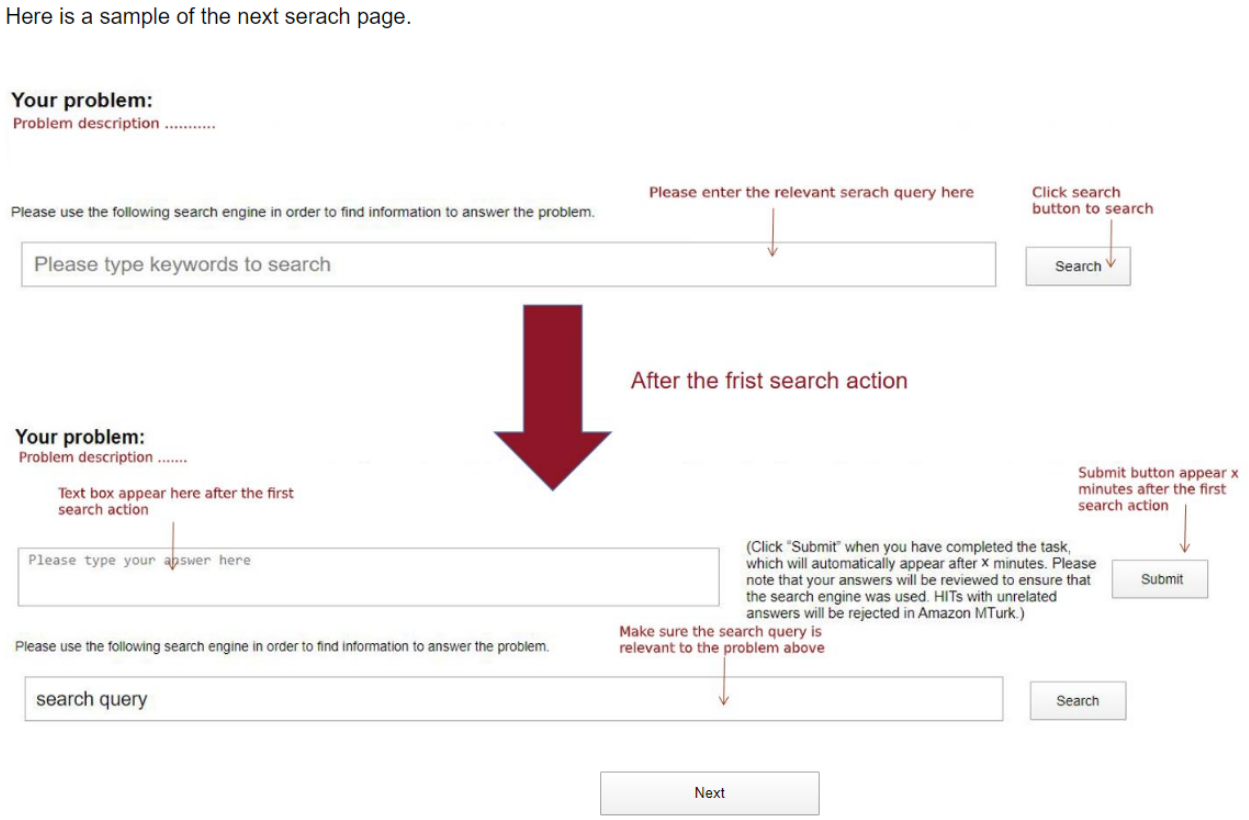


Figure 3.2: Instruction page part II

3.2.2 Web Search Interface

Figure 3.3 presents the beginning stage of the web search interface. Description of the problem is shown and a search bar for participants to type the search query beneath the text box. We used Bing Search API to retrieve results for the participants. Figure 3.4 shows the interface when the participants start searching. A text box is provided for the participants to make notes and to type their answers. The reason to have this text box in this web search page is also for quality control. If we do not ask for the answer, participants may randomly search or just search for the irrelevant information, and we could not check whether they do the relevant search or not. A total of ten results are presented in each page, and participants can click the page number at the bottom of the page for more search results, which are shown in Figure 3.5.

The task description and text box cost some space and this will make the user review less search results at the first time, and we believed this will affect the analysis a little bit, however, our purpose is to investigate can the results from a small sample of participants generalize to the results from a large sample of participants. So this effect will both affect a small sample of participants and a large sample of participants, and the task description and text box will not affect our results based on our purpose.

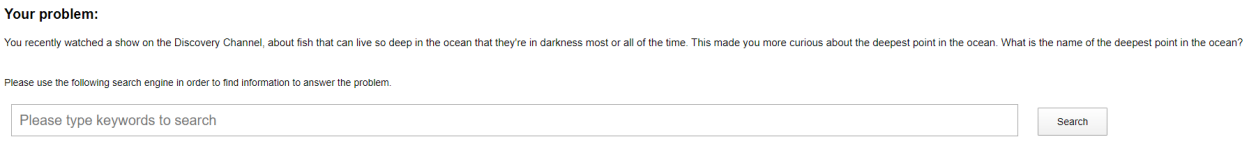


Figure 3.3: Web search page part I

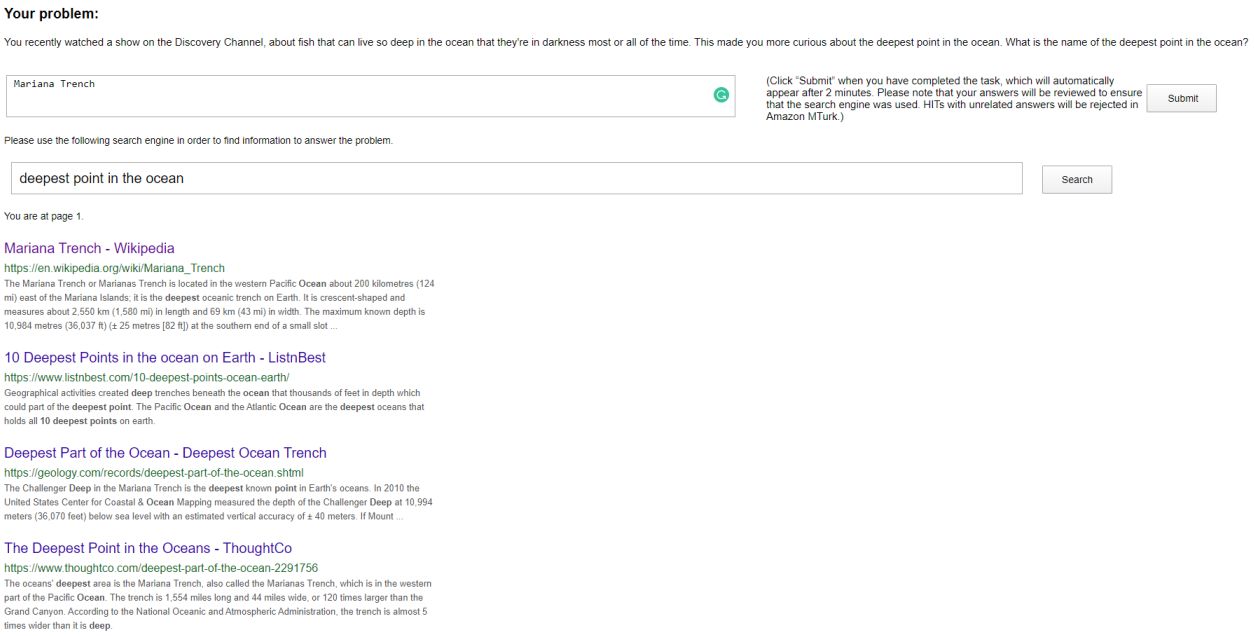


Figure 3.4: Web search page part II

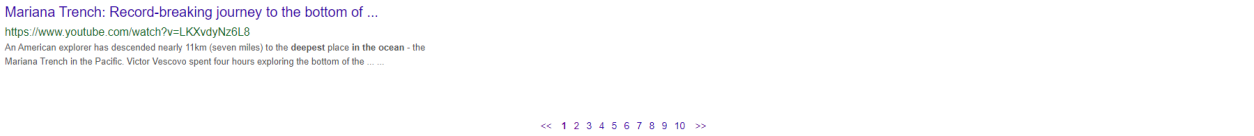


Figure 3.5: Web search page part III

3.3 Pre-Task Questionnaire

A pre-task questionnaire from Kelly et al. (2015) is given after participants previewed the search task. This questionnaire has three parts: 1) Interest and Knowledge about the task; 2) Task Complexity; and 3) Expected Task Difficulty. Participants rate most of the questions on a scale from 1 to 5 (1 = not at all, 2 = slightly, 3 = somewhat, 4 = moderately and 5 = very) except two questions. The scale for "How many times have you searched for information about this task?" is never, 1 - 2 times, 3 - 4 times and 5+ times. For "How much do you know about the topic of the task?", participants select one of the choses from Nothing, Little, Some and Great deal. The questionnaire is shown in Table 3.3.

Table 3.3: Lists of Pre-Task Questionnaire (Kelly et al., 2015)

Interest & Knowledge	How interested are you to learn more about the topic of this task? How many times have you searched for information about this task? How much do you know about the topic of the task?
Task Complexity	How defined is this task in terms of the types of information needed to complete it? How defined is this task in terms of the steps required to complete it? How defined is this task in terms of its expected solution?
Expected Task Difficulty	How difficult do you think it will be to **search** for information for this task using a search engine? How difficult do you think it will be to **understand** the information the search engine finds? How difficult do you think it will be to **decide** if the information the search engine finds is useful for completing the task? How difficult do you think it will be to **integrate** the information the search engine finds? How difficult do you think it will be to determine **when you have enough** information to finish the task?

3.4 Post-Task Questionnaire

The post-task questionnaire, also from Kelly et al. (2015), is given to the participants after finishing the search task. This questionnaire has five parts: 1) Engagement: how participants feel while completing the task; 2) Interest: how much their interest and knowledge has changed after the task; 3) Experienced Task Difficulty: same questions from the Expected Task Difficulty questions in pre-task questionnaire are used to assess the actual experience regarding task difficulty; 4) Overall Difficulty of the Task and 5) Overall Satisfaction about the Task. All questions are rated on the scale from 1 to 5 (1 = not at all, 2 = slightly, 3 = somewhat, 4 = moderately and 5 = very). The questionnaire is shown in Table 3.4.

Table 3.4: Lists of Post-Task Questionnaire (Kelly et al., 2015)

Engagement	How enjoyable was it to do this task? How engaging did you find this task? How difficult was it to concentrate while you were doing this task?
Interest	How much did your interest in the task increase as you searched? How much did your knowledge of the task increase as you searched?
Experienced Task Difficulty	How difficult was it to **search** for information for this task using a search engine? How difficult was it to **understand** the information the search engine finds? How difficult was it to **decide** if the information the search engine finds is useful for completing the task? How difficult was it to **integrate** the information the search engine finds? How difficult was it to determine **when you have enough** information to finish the task?
Overall Difficulty	Overall, how difficult was this task?
Overall Satisfaction	Overall, how satisfied are you with your solution to this task? Overall, how satisfied are you with the search strategy you took to solve this task?

3.5 Search Behaviors

We collected participants' search behaviors through our experiments. There are nine search behavior measurements in total. The first six are recorded directly by our web search engine and the rest three are computed after all the participants completed the task. The measurements and their definitions are listed below.

1. Number of Queries: The total number of unique queries when the participant performs a task.

2. Query length: The average number of terms in all unique queries when the participant performs a task.

3. Number of Unique Query Terms: The total number of unique query terms in all unique

queries when the participant performs a task.

4. Number of Clicks: The total number of clicks when the participants on the web search pages.

5. Number of Visited URLs: The total number of unique URLs visited by the participants when they perform a task.

6. Number of Queries without a Click: The total number of unique queries without a single click by the participants when they perform a task

7. Query Diversity: The total number of unique queries that were not used by the other participant when they complete the same task.

8. Query Term Diversity: The total number of unique query terms that were not used by the other participant when they complete the same task.

9. URLs Diversity: The total number of URLs visited that were not visited by the other participant when they complete the same task.

3.6 Analysis method

3.6.1 ANOVA

As the null hypothesis suggests that cognitive complexity will not affect people's searching behavior, the current experiment manipulated the task difficulty according to different cognitive demands to test if the means for each condition are significantly different from one another or if they are relatively the same. Therefore, we used one-way ANOVA to evaluate the relationship between cognitive complexity level and dependent variables that are the

search behavior, pre-task questionnaire response and post-questionnaire response. We set alpha value to 0.01.

3.6.2 Generalizability Error Rate

In order to investigate the generalizability of online user study, we used generalizability error rate to find out the best number of participants for online user study.

Since the current study recruited 100 participants for each cognitive level, which is a relatively robust sample size, it can be assumed that the current statistical test result can be the basic standard. To compute error rate for each dependent variable, we randomly selected N participants' data from the 100 participants data set per cognitive complexity level with no replacement. Then we had a small sample with $5N$ participants' data from all five cognitive complexity level. We repeated this sampling for 1000 times to create 1000 samples. As it is repeatedly sampled for 1000 times, we would have 1000 statistical test results. We compared these 1000 statistical test results with the basic standard. The generalizability error rate is calculated by the equation 3.1, where n is the number of samples where the statistical test result is different from the basic standard.

We computed generalizability error rate with the different N, number of participants per cognitive complexity level to plot the error rate figure. The range of N is from 5 (the minimum) to 100 (the basic standard) with a step of 5.

3.6.3 Reproducibility Error Rate

We further employed reproducibility error rate to find out the desired number of participants for online user study to investigate the reproducibility of online user study.

To compute error rate for each dependent variable, we randomly selected N participants' data from the 100 participants data set per cognitive complexity level with no replacement twice firstly. Then we compared the statistical test result of the two samples we got from sampling. If they are different, we counted this comparison as an error. This is because if the sample size of N is reliable enough, the results from any N participants should be the same. We repeated this sampling and comparison for 1000 times. The reproducibility error rate is calculated by the Equation 3.1, where n is the number of errors, which the statistical test results of two small samples are different.

We computed error rate with the different N, number of participants per cognitive complexity level to plot the error rate figure. The range of N is from 5 (the minimum) to 100 with a step of 5.

$$ErrorRate = \frac{n}{1000} \tag{3.1}$$

Chapter 4

Results

This section has two main parts: one is the statistical results between cognitive complexity level and dependent variables (search behaviors, pre-task questionnaire responses and post-task questionnaire responses); the other one is the error rates of the dependent variables to examine the reproducibility and reliability of human experiments in interactive information retrieval systems. The current study further analyzed and reported the generalizability for each item that was found statistically significant.

4.1 ANOVA results for cognitive complexity level and dependent variables

4.1.1 Search Behaviors

For each search behavior, we conducted a one-way between-subjects ANOVA to compare the effects of different cognitive demands. As shown in Table 4.1, we found significant effects of cognitive complexity on Unique Queries [$F = 12.36, p = 1.38 \times 10^{-9}$], Unique Query Term [$F = 5.31, p = 3.42 \times 10^{-4}$], Clicks [$F = 6.65, p = 3.24 \times 10^{-5}$], URLs Visited [$F = 5.43, p = 2.74 \times 10^{-4}$], Queries without Clicks [$F = 5.26, p = 3.72 \times 10^{-4}$], and Query Diversity [$F = 7.06, p = 1.57 \times 10^{-5}$] at the p<.01 level. It shows that these search behaviors

will be significantly influenced and changed if the task demands different levels of cognitive resources.

Post hoc comparisons were used to identify the specific differences. For Unique Query, it indicated that the mean scores for the remember level (M = 1.22, SD = 0.50) and the understand level (M = 1.27, SD = 0.63) were significantly different from the analyze level (M = 1.93, SD = 1.57) and the create level (M = 2.13, SD = 1.43) and the evaluate level (M = 1.58, SD = 1.12) was significantly different from the create level. For Unique Query Terms, participants used the most unique query terms in create level tasks (M = 10.37, SD = 9.48) and the least unique query terms in understand level tasks (M = 7.52, SD = 2.72), and significant difference was also found between the create level tasks and the remember level tasks (M = 7.65, SD = 2.67). In terms of Clicks, participants clicked significantly in create level tasks (M = 3.51, SD = 4.17) than remember level tasks (M = 1.64, SD = 1.68), understand level tasks (M = 1.78, SD = 1.41) and evaluate level tasks (M = 1.95, SD = 2.35). For the URLs Visited, significant differences were found between the remember level tasks (M = 1.16, SD = 1.53), the understand level tasks (M = 1.3, SD = 1.81) and the create level tasks (M = 2.36, SD = 2.68). Post hoc analysis also revealed that participants used more queries without clicks when the complexity level increases: significant differences were found between analyze level tasks (M = 1.06, SD = 1.50) and remember level tasks (M = 0.51, SD = 0.64), and understand level tasks (M = 0.53, SD = 0.62). Moreover, for Query Diversity, people searched more diverse for create level tasks (M = 1.39, SD = 1.33) than for remember level tasks (M = 0.69, SD = 0.72) and understand level tasks (M = 0.62, SD = 0.86).

No significant difference was found for Query Length by cognitive levels ($p = 0.021$), showing that task's cognitive demand may not affect people typing longer or shorter queries. Changes in Query Term Diversity ($p = 0.39$), and URL Diversity ($p = 0.02$) were also non-significantly

affected.

Table 4.1: Results of different search behaviors (Mean, Standard Deviation) of the five cognitive levels. * $p < 0.01$

	Remember	Understand	Analyze	Evaluate	Create
Unique Queries*	(1.22, 0.50)	(1.27, 0.63)	(1.93, 1.57)	(1.58, 1.12)	(2.13, 1.43)
Query Length	(7.49, 2.52)	(6.85, 2.27)	(6.57, 3.95)	(8.31, 4.86)	(7.29, 4.96)
Unique Query Terms*	(7.65, 2.67)	(7.52, 2.72)	(8.27, 4.19)	(10.04, 6.84)	(10.37, 9.48)
Clicks*	(1.64, 1.68)	(1.78, 2.41)	(2.55, 3.50)	(1.95, 2.35)	(3.51, 4.17)
URLs Visited*	(1.16, 1.53)	(1.30, 1.81)	(1.87, 2.39)	(1.46, 1.77)	(2.36, 2.68)
Queries w/out Clicks*	(0.51, 0.64)	(0.53, 0.62)	(1.06, 1.50)	(0.75, 1.11)	(0.99, 1.35)
Query Diversity*	(0.69, 0.72)	(0.62, 0.86)	(1.03, 1.55)	(0.99, 1.10)	(1.39, 1.33)
Query Term Diversity	(0.35, 0.84)	(0.49, 1.31)	(0.55, 1.44)	(0.8, 2.36)	(0.57, 1.63)
URL Diversity	(0.23, 0.58)	(0.37, 0.86)	(0.67, 1.86)	(0.56, 1.13)	(0.74, 1.24)

4.1.2 Task Complexity and Task Difficulty

Task complexity is determined by three domains, Types of Information Needed, Steps to Complete, and Expected Solution. Task difficulty is evaluated by the pre-questionnaires and post-questionnaires, which asked the participants to rate on Search, Understand, Decide, Integrate, and Enough levels.

In Figure 4.1, all items in remember level tasks score higher than those in other cognitive levels ($p < .01$), showing that participants in the remember level group rated their topics to be more clearly defined than those in other groups. This means that the task's cognitive complexity of the task has an impact on participant's evaluation of how complex the task would be even before doing it. Post hoc analyses revealed significant differences in types of info needed between remember level tasks (M = 4.15, SD = 0.97) and understand level tasks (M = 3.63, SD = 0.97) and analyze level tasks (M = 3.65, SD = 0.95). Significant differences were also found in steps to complete between remember level tasks (M = 4.13, SD = 1.07) and analyze level tasks (M = 3.57, SD = 1.09).

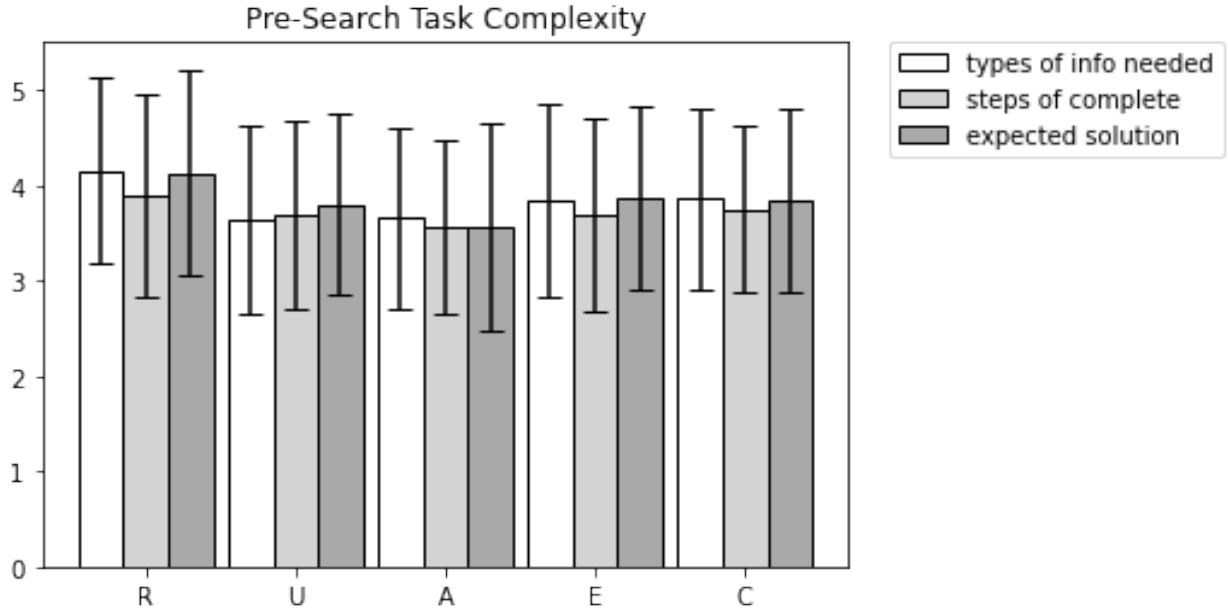


Figure 4.1: Pre-Search task complexity

Figure 4.2 reports the results for pre-search task difficulty, one-way ANOVA shows that all items in the remember level score significantly less than those in other cognitive complexity levels ($p < .01$). This finding is consistent with the one obtained from task complexity. Both of them show that participants would expect the task with lower cognitive demands to be less complex and less difficult before conducting searching.

Post-questionnaire collected participant's evaluations on task difficulty after they finished the searching behaviors. As shown in Figure 4.3, cognitive complexity has significant influences on understanding information [$F = 4.51, p = 1.37 \times 10^{-3}$], integrating the found information [$F = 4.73, p = 9.32 \times 10^{-4}$] and when to stop searching [$F = 6.56, p = 3.78 \times 10^{-5}$]. Regarding understanding information, create level tasks (M = 2.31, SD = 1.32) and understand level tasks (M = 2.28, SD = 1.33) score significantly higher than remember level tasks (M = 1.66, SD = 1.07), indicating that people find it easier to decide useful information for tasks at the remember level. For integrating the found information, people consider it to be more

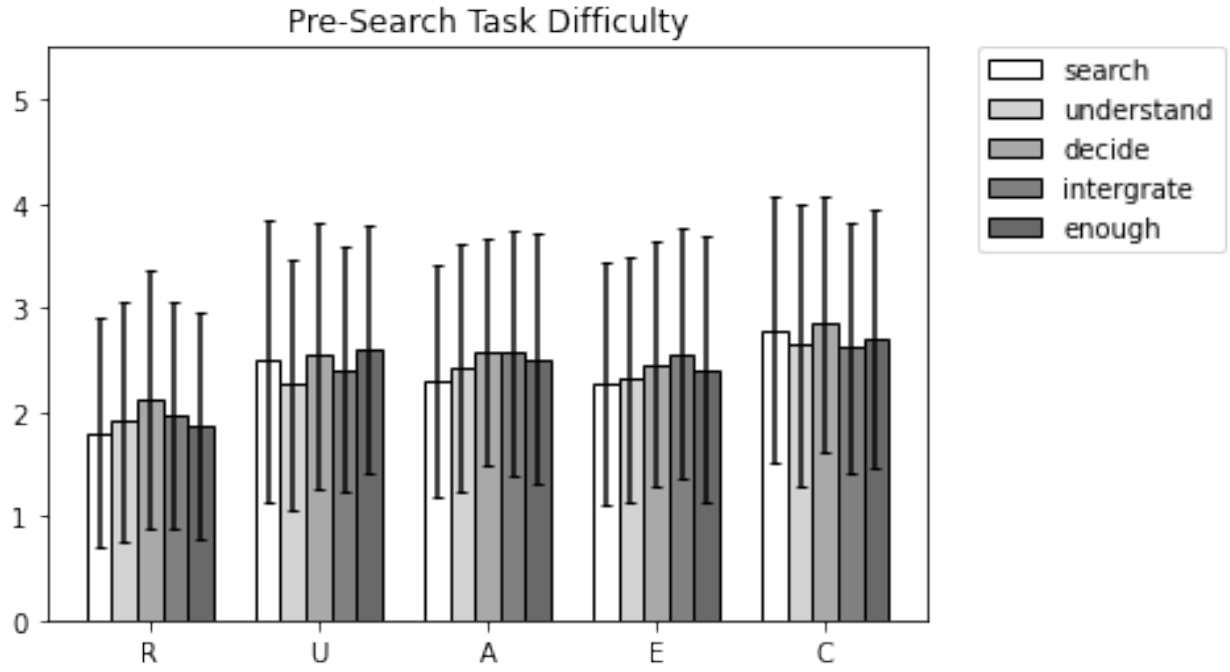


Figure 4.2: Pre-Search task difficulty

difficult for create level tasks (M = 2.43, SD = 1.32) and evaluate level tasks (M = 2.25, SD = 1.24) comparing to remember level tasks (M = 1.76, SD = 1.08). Also, remember level tasks (M = 1.73, SD = 1.03) is significantly easier to stop searching than all other cognitive levels tasks ($p < .01$). The overall trend of post-search task difficulty follows the trend of pre-search task difficult, showing that as the cognitive level increases, participants feel the task to be more difficult to complete.

4.1.3 Enjoyment and Engagement

In terms of participant's enjoyment, engagement and concentration, Figure 4.4 shows that participants reported moderated enjoyment and engagement regardless of the cognitive level of the search tasks. For concentration $[F = 4.48, p = 1.44 \times 10^{-3}]$, the result gives a significant difference between the create level and the remember level, indicating that participants

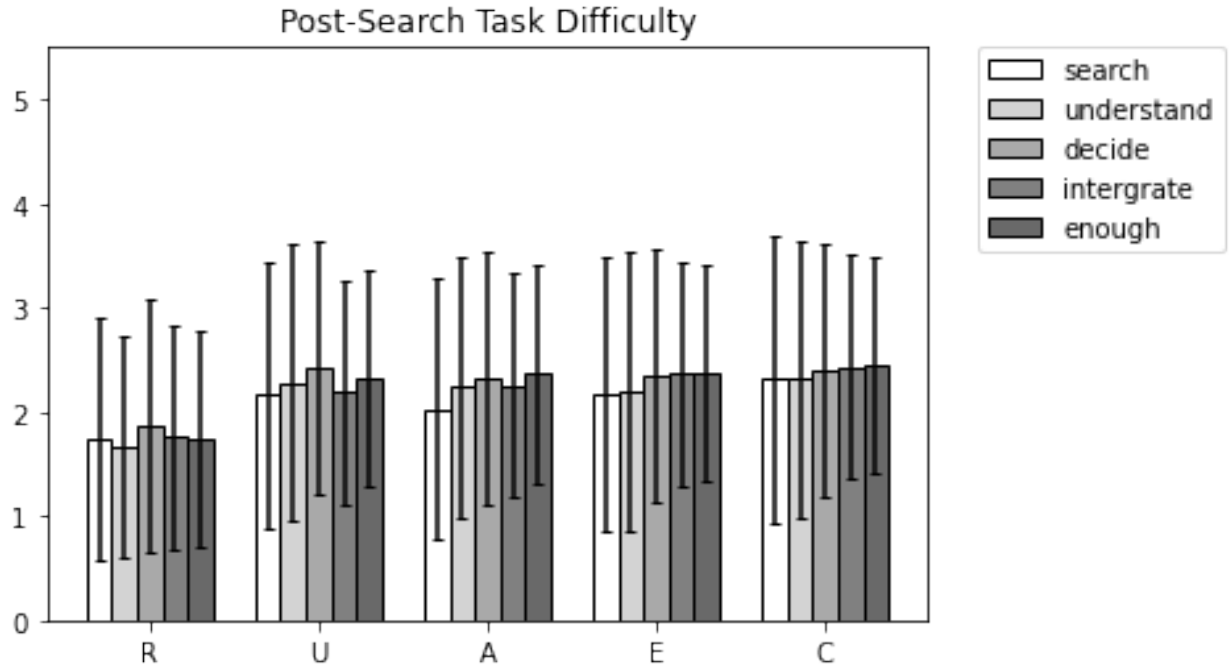


Figure 4.3: Post-Search Task Difficulty

completing tasks at the create level felt more concentrated than those completing remember level tasks. This difference indicates that create level tasks might be more difficult than remember level tasks and is consistent with the increased cognitive demands.

4.1.4 Overall Difficulty and Satisfaction

Results of overall difficulty and satisfaction are shown in Figure 4.5. Participants rated that they were equally satisfied with their solutions and strategies with no significant differences between the five cognitive levels. Overall difficulty $[F = 8.96, p = 5.42 \times 10^{-7}]$ at the remember level was rated significantly lower than that of the create, evaluate and analyze level.

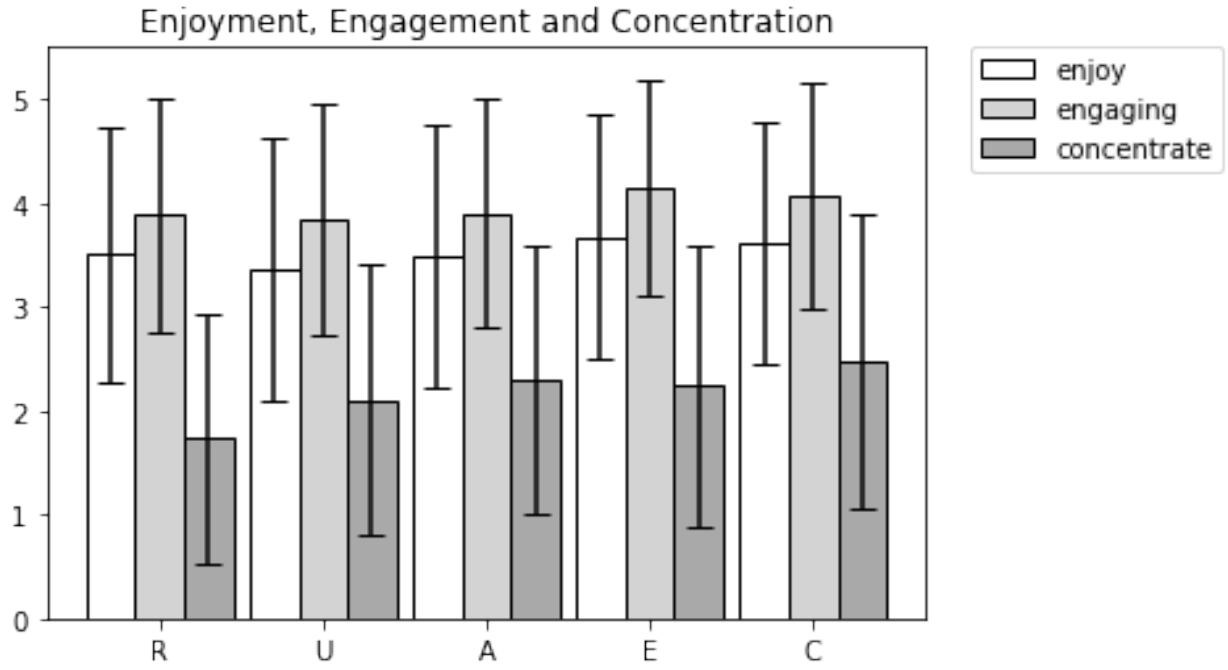


Figure 4.4: Enjoyment, Engagement and Concentration

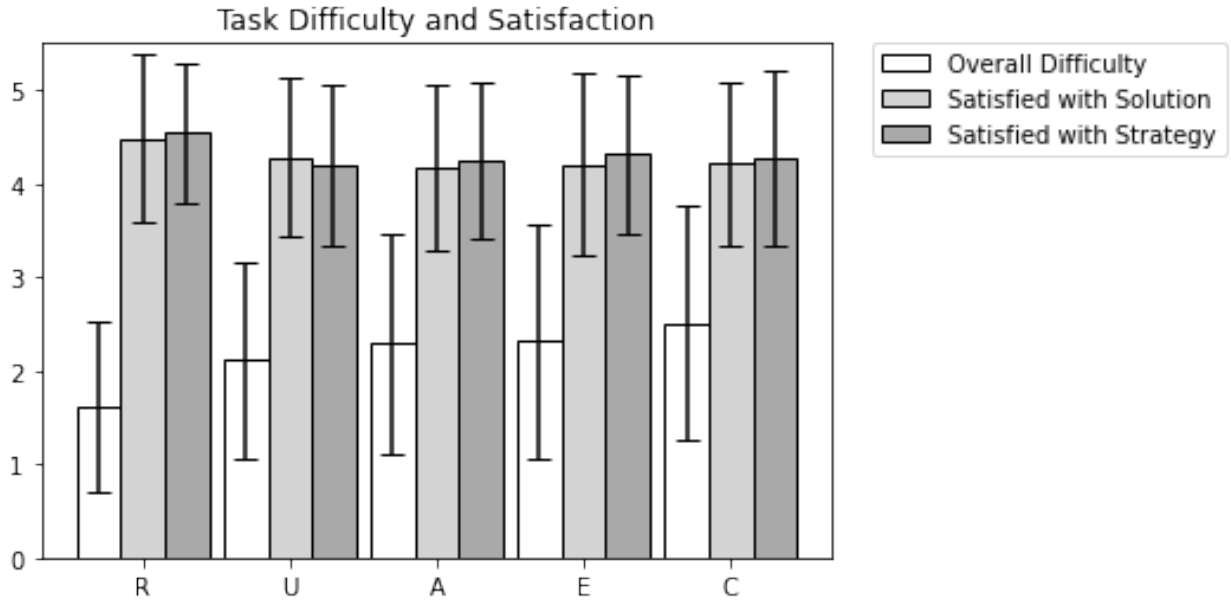


Figure 4.5: Task difficulty and satisfaction

4.2 Generalizability Error Rates of Dependent Variables

In order to evaluate the generalizability of user studies, we generate generalizability error rates figures of search behaviors and user experiences. The red dashed line in the generalizability error rate figure is where generalizability error rate is 0.1; if a figure does not have the red dashed line, it means its generalizability error rate is constantly below 0.1. To make sure the current results are qualified to compute generalizability error rates for evaluating generalizability, they were compared to the results from Kelly et al. (2015). As shown in Figure 4.6, when the number of participants per complexity level is larger than 65, the generalizability error rate is less than 0.1 and kept decreasing to 0 as the number of participants increases, indicating that the current sample with 100 online participants per cognitive complexity level is statistically comparable with offline study in Kelly et al. (2015). This finding supports that the current results are allowed to be used for evaluating generalizability.

4.2.1 Search Behaviors

Figure 4.7 shows the changes in error rate of the five search behaviors' one-way ANOVA results, which are all significant ($p < .01$). The changes in generalizability error rate of the five search behaviors revealed a similar pattern. Generalizability error rate is relatively high and approaching 1.0, when the sample size is less than 20 per cognitive complexity level. As the number of participants increases, generalizability error rate gradually decreases. For Unique Queries (Figure 4.7a), the generalizability error rate is less than 0.1 when sample size is up to 40 per cognitive complexity level. For Unique Query Terms (Figure 4.7b), URLs Visited (Figure 4.7d) and Queries w/out Clicks (Figure 4.7e) the generalizability error rates

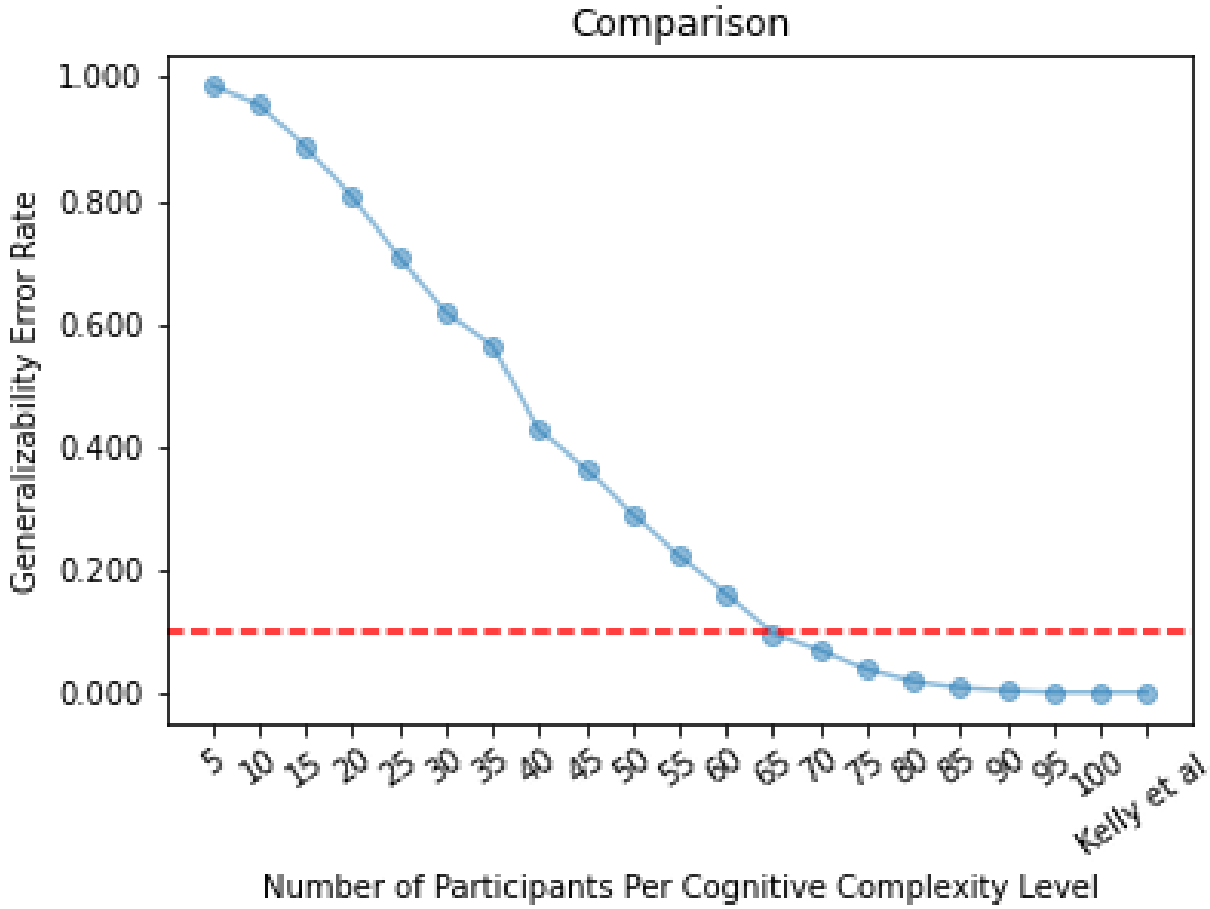
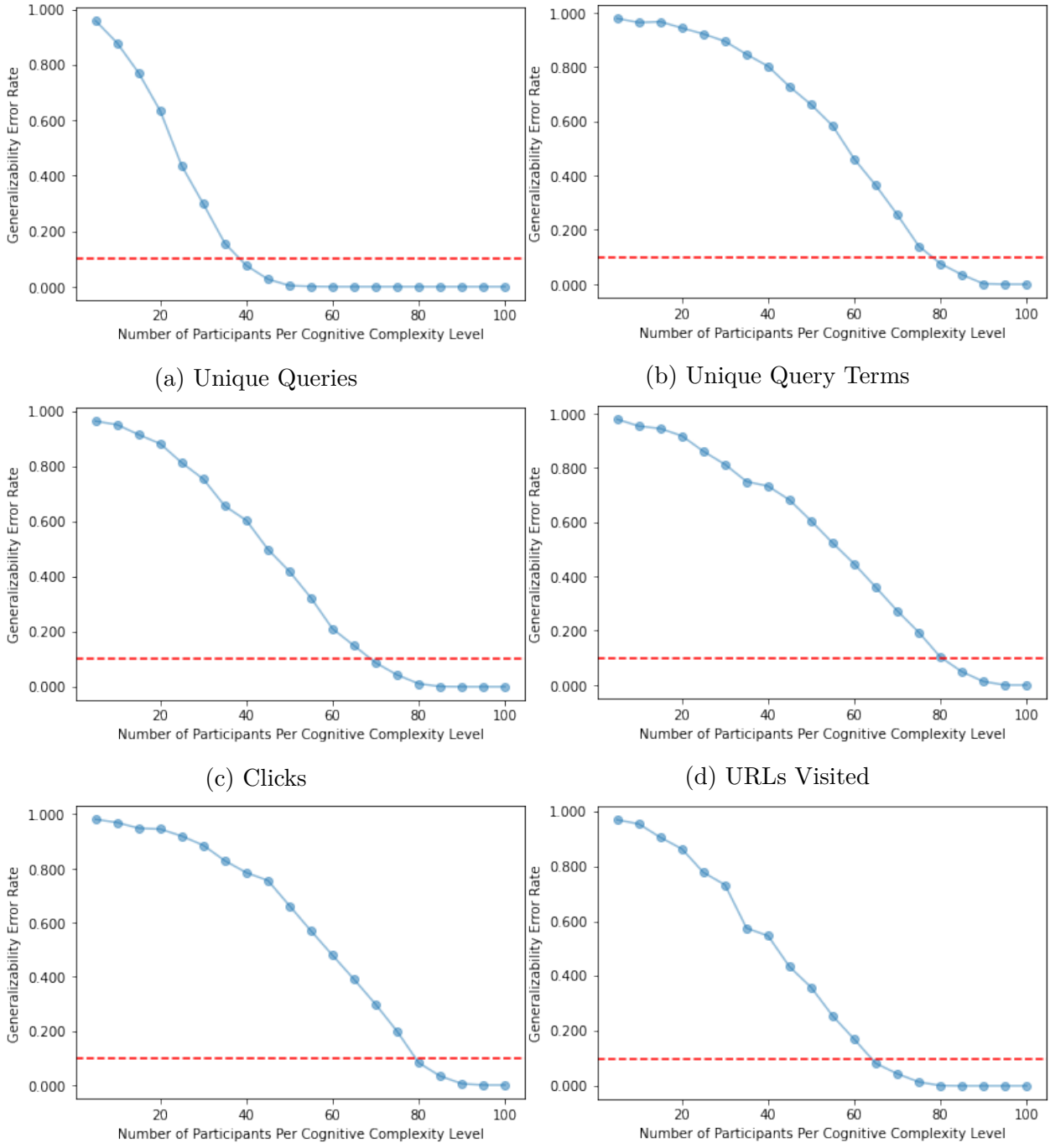


Figure 4.6: Comparison

are less than 0.1 when sample size is around 80 per cognitive complexity level. For Clicks (Figure 4.7c) and Query Diversity (Figure 4.7f), the generalizability error rates are less than 0.1 when the number of participants is over 65 per cognitive complexity level.

Figure 4.8 reported the changes in generalizability error rates for the search behaviors with insignificant ANOVA results. For Query Length (Figure 4.8a) and URL Diversity (Figure 4.8c), generalizability error rates increased as the sample size grows, but they started to decrease when the sample size is close to 100 per cognitive complexity level. For Query Term Diversity, generalizability error rate was close to zero regardless of the changes in the number of participants, and becomes zero when sample size reaches 50 per cognitive



(a) Unique Queries

(b) Unique Query Terms

(c) Clicks

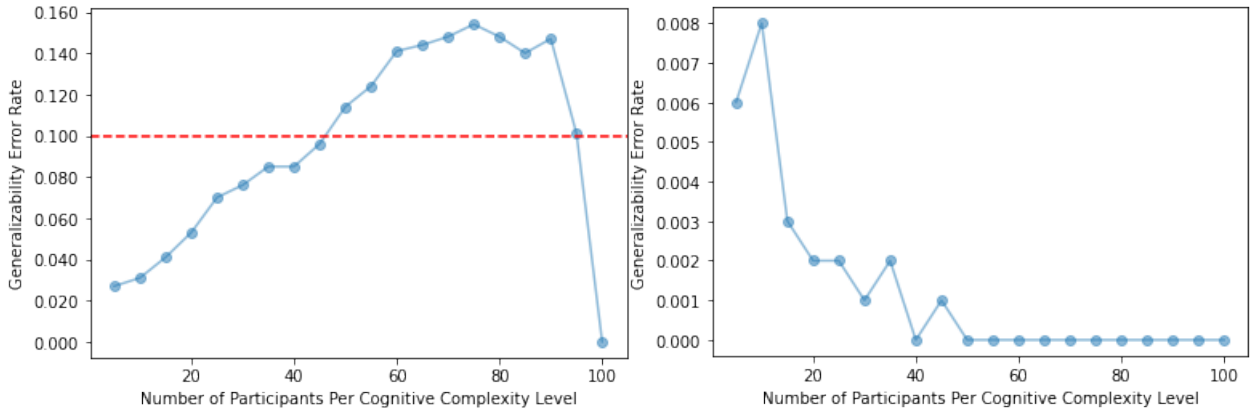
(d) URLs Visited

(e) Queries w/out Clicks

(f) Query Diversity

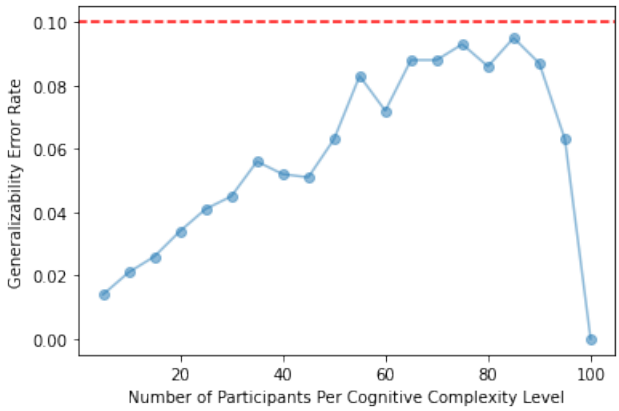
Figure 4.7: Generalizability error rate figures of 6 Search Behaviors with significant influences from one-way ANOVA test

complexity level.



(a) Query Length

(b) Query Term Diversity

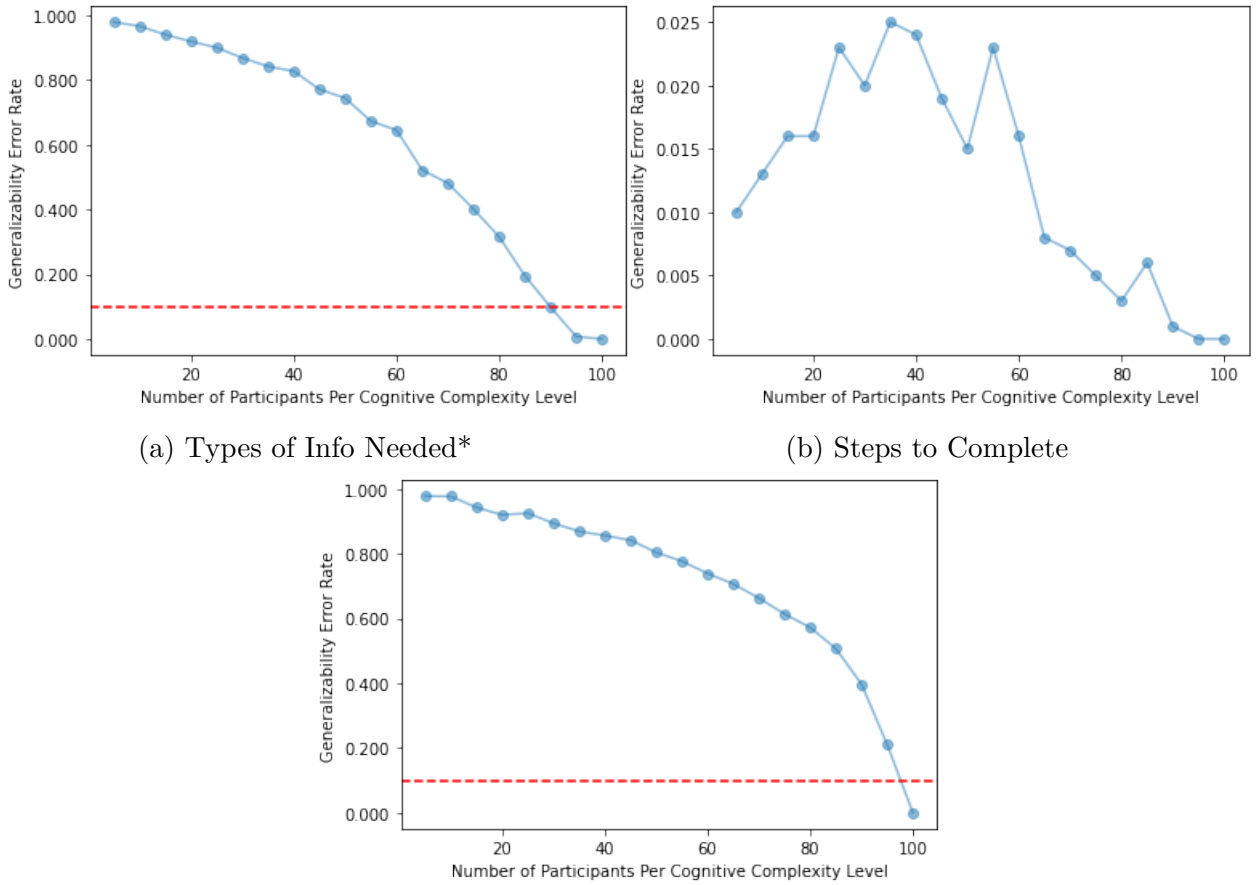


(c) URL Diversity

Figure 4.8: Generalizability error rate figures of 3 Search Behaviors without significant influences from one-way ANOVA test

4.2.2 Task Complexity

Figure 4.9 demonstrates the changes in generalizability error rates for Task Complexity. Error rates for two items (Figure 4.9a and Figure 4.9c) with significant ANOVA results decreased from 1 to 0 as the sample size grows. Changes in generalizability error rates for Steps to Complete (Figure 4.9b) fluctuate between 0.03 and 0.00.



(a) Types of Info Needed*

(b) Steps to Complete

(c) Expected Solution*

Figure 4.9: Generalizability error rate figures of Pre-Search Task Complexity from one-way ANOVA test. * means $p < 0.01$ at 100 participants per cognitive complexity level

4.2.3 Task Difficulty

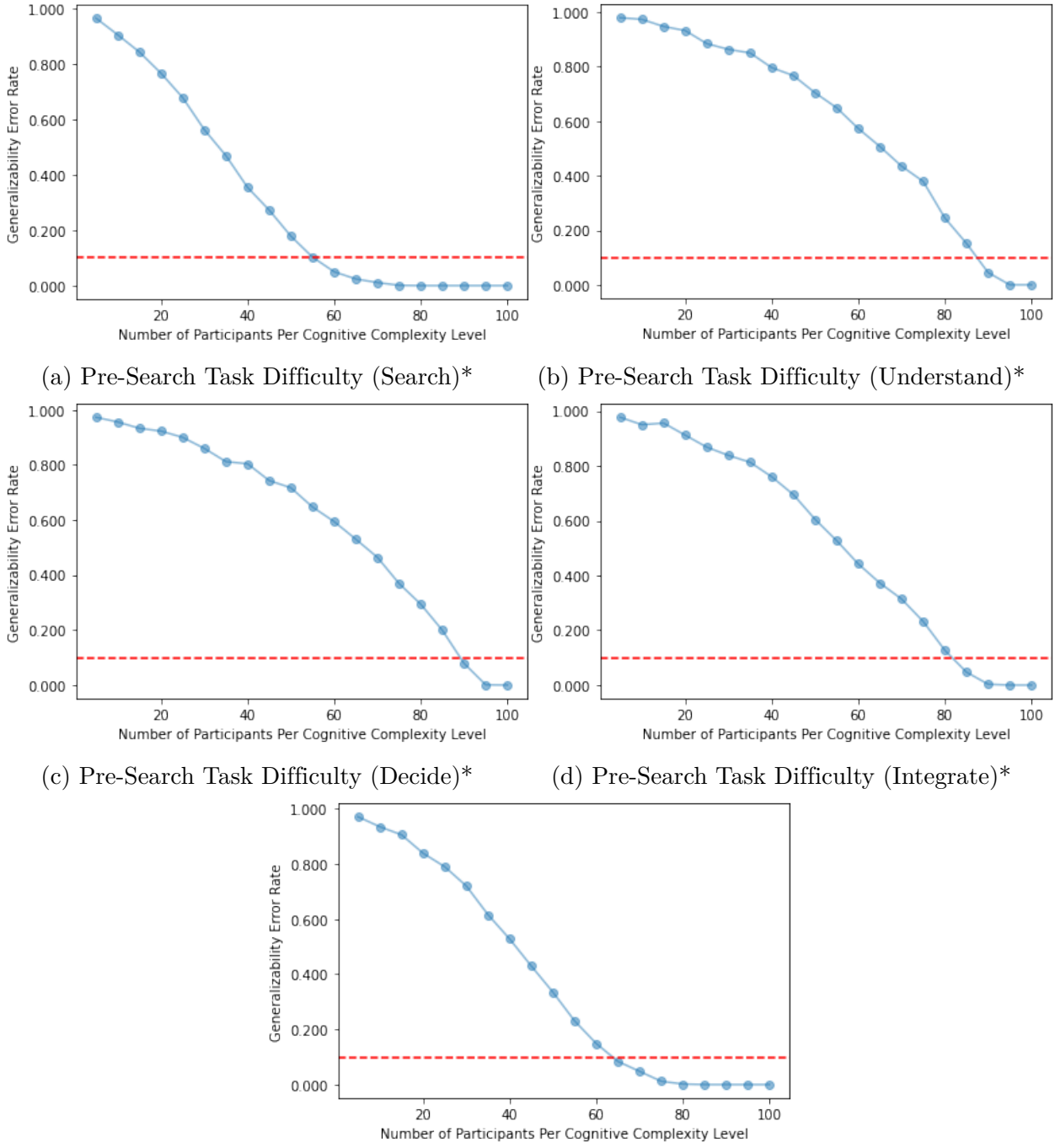
Generalizability error rates of all items in pre-search task difficulty decrease from 1 to 0 as the number of participants increases, but the changing pattern differs. For difficulty in searching information (Figure 4.10a) and difficulty of determining when to stop (Figure 4.10e), generalizability error rates decline faster and are below 0.1 when the number of participants is over 60. For the difficulty in understanding searched information (Figure 4.10b), difficulty in deciding searched information (Figure 4.10c) and difficulty in integrating searched information (Figure 4.10d), the generalizability error rates decrease to the 0.1 level

when the number of participants is around 85.

Figure 4.11 shows generalizability error rate figures for the five items of post-search task difficulty. The ANOVA results were significant for understanding the searched information, integrating the searched information and determining when to stop searching, so the changes in generalizability error rates of these three variables follow the general pattern: decreasing as the number of participants increases, and their generalizability error rates decline below 0.1 when the number of participants per cognitive complexity level is around 90, 90, 70 respectively. Generalizability error rate fluctuates when participants per cognitive complexity level is fewer than 85 for searching information (Figure 4.11a), but the maximum generalizability error rate is around 0.12, which is not very large. For deciding the useful information (Figure 4.11c), generalizability error rate shows an increasing trend when the number of participants increases, but it diminishes to 0 when the number of participants per cognitive complexity level is 100 .

4.2.4 Enjoyment, Engagement and Concentration

Similar to other significant ANOVA results, changes in generalizability error rate for Concentration also show a decreasing pattern from 1 to 0 as the sample size accumulates (Figure 4.12c). The other two figures demonstrate the changes in error rates for the insignificant results, which fluctuates between 0.02 and 0. For Enjoyment (Figure 4.12a) and Engagement (Figure 4.12b), generalizability error rates are unstable, but they decrease to 0 at 70 participants and 90 participants per cognitive complexity level respectively.



(a) Pre-Search Task Difficulty (Search)*

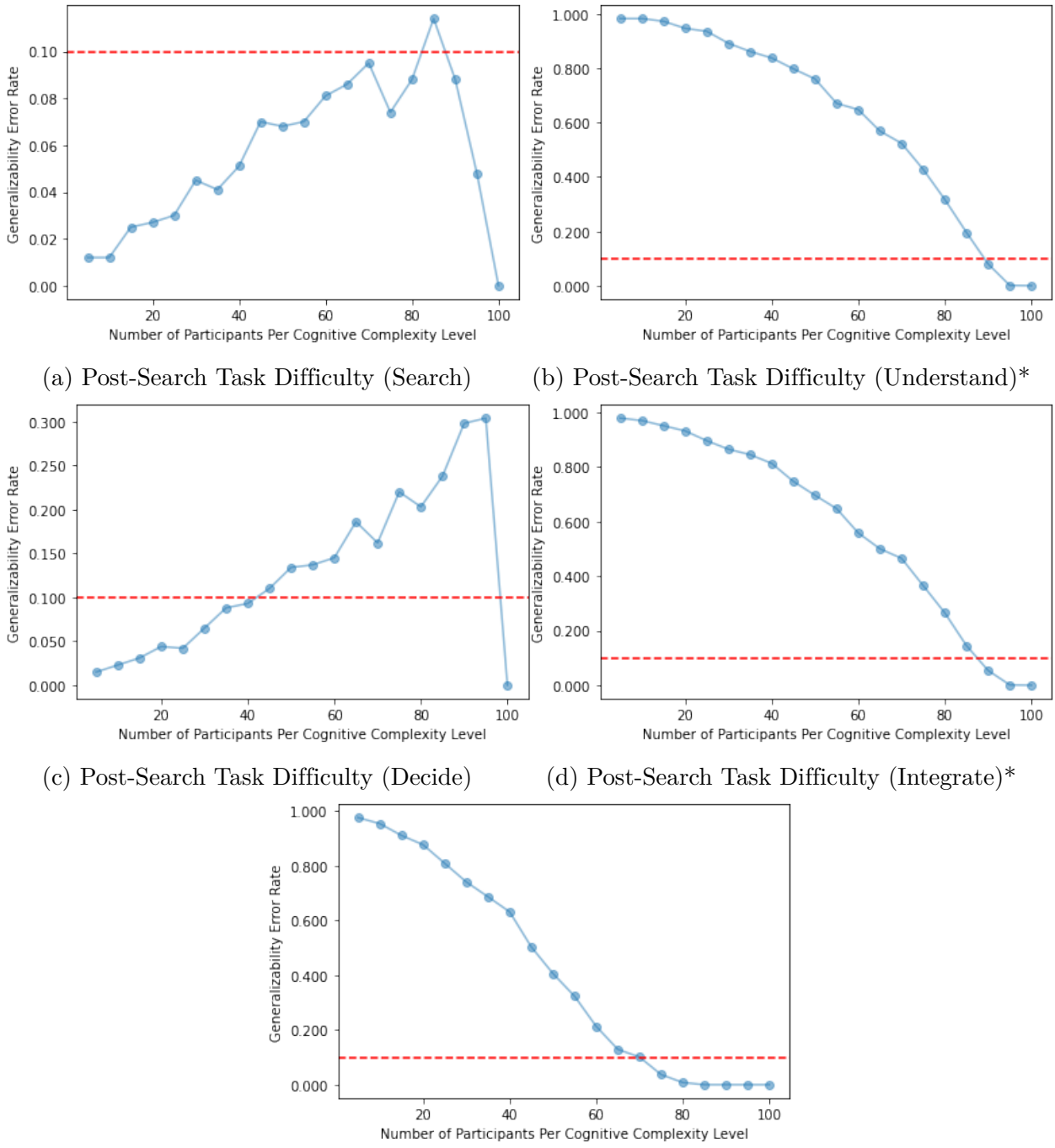
(b) Pre-Search Task Difficulty (Understand)*

(c) Pre-Search Task Difficulty (Decide)*

(d) Pre-Search Task Difficulty (Integrate)*

(e) Pre-Search Task Difficulty (Enough)*

Figure 4.10: Generalizability error rate figures of Pre-Search Task Difficulty from one-way ANOVA test. * means $p < 0.01$ at 100 participants per cognitive complexity level



(a) Post-Search Task Difficulty (Search)

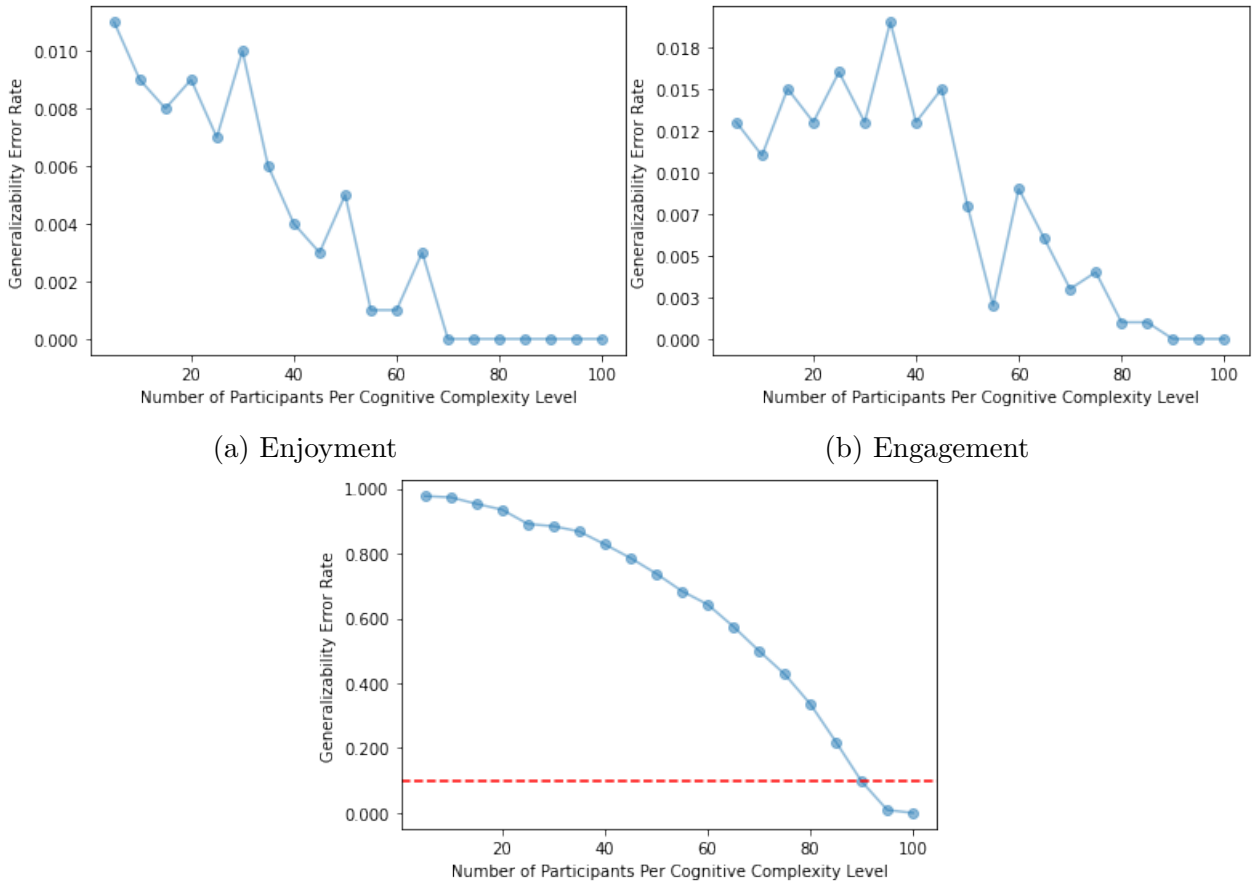
(b) Post-Search Task Difficulty (Understand)*

(c) Post-Search Task Difficulty (Decide)

(d) Post-Search Task Difficulty (Integrate)*

(e) Post-Search Task Difficulty (Enough)*

Figure 4.11: Generalizability error rate figures of Post-Search Task Difficulty from one-way ANOVA test. * means $p < 0.01$ at 100 participants per cognitive complexity level.



(a) Enjoyment

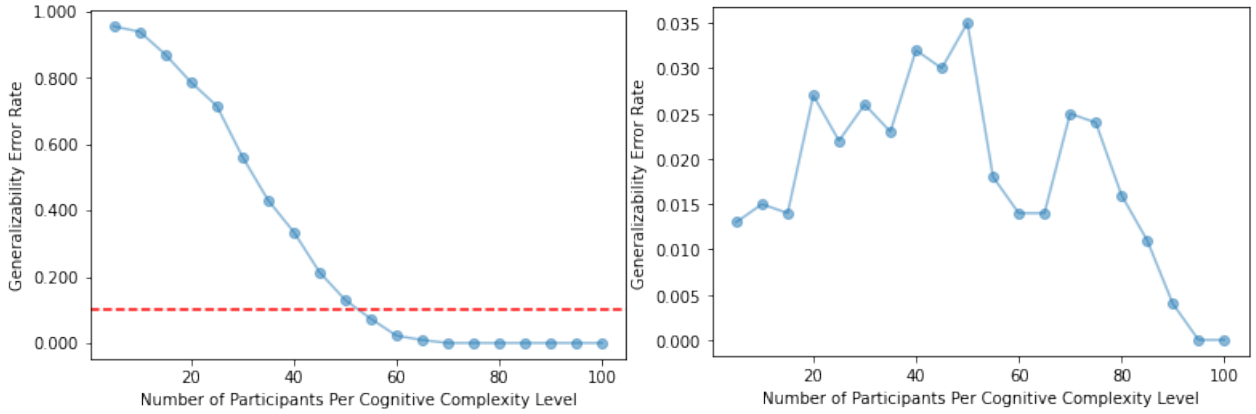
(b) Engagement

(c) Concentration

Figure 4.12: Generalizability error rate figures of Enjoyment, Engagement and Concentration from one-way ANOVA test. * means $p < 0.01$ at 100 participants per cognitive complexity level

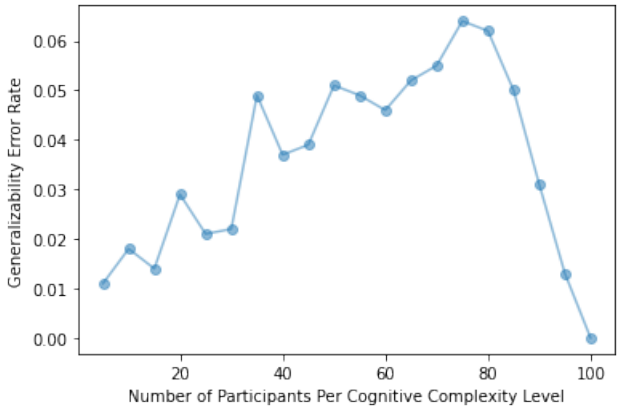
4.2.5 Overall Difficulty and Satisfaction

Figure 4.13 shows the changes in generalizability error rate for Overall Difficulty, Satisfied with Solution and Satisfied with Strategy. For Overall Difficulty, the rate decreases from 1 to 0 as the number of participants per cognitive complexity level increases, which becomes less than 0.1 when the sample size is 55 per cognitive complexity level. The changes in generalizability error rates fluctuate for the other two items, but the maximum are around 0.05.



(a) Overall Difficulty

(b) Satisfied with Solution



(c) Satisfied with Strategy

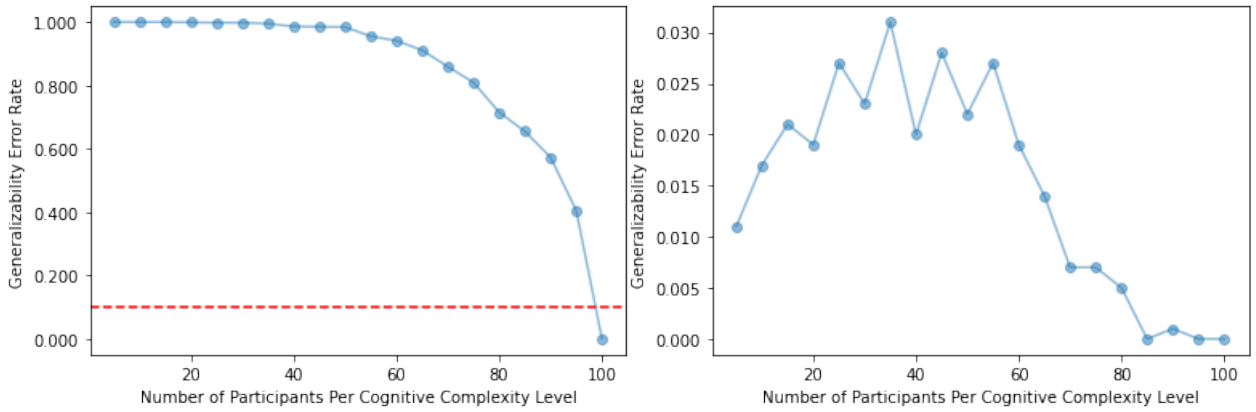
Figure 4.13: Generalizability error rate figures of Overall Difficulty and Satisfaction from one-way ANOVA test. * means $p < 0.01$ at 100 participants per cognitive complexity level

4.3 Generalizability Error Rates for Individual Items

In Section 4.1, we have reported the one-way ANOVA results, and some dependent variables are statistically significant. In this section, we developed the generalizability error rate figures for these variables' post hoc results. Depending on the significance of the post hoc results, we computed two figures for each variable: a) figures of which cognitive complexity levels having significant influences , and b) figures of which cognitive complexity levels showing non-significant influences.

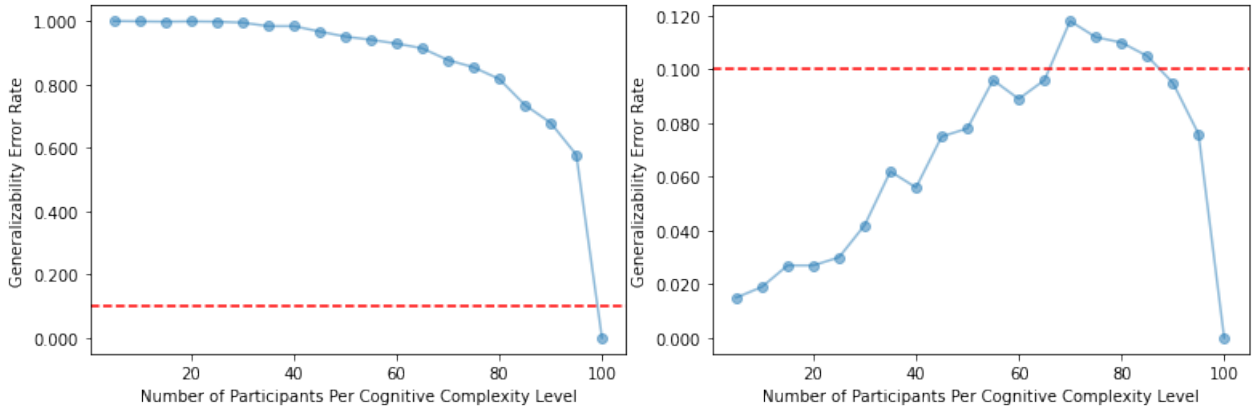
From Figure 4.14a to Figure 4.31a, all of the changes in generalizability error rate have the same pattern, in which the generalizability error rates decrease from 1 to 0 as the number of participants increases. However, they are not exactly the same. For Number of Unique Queries (Figure 4.14a), Unique Query Terms (Figure 4.15a), Queries without Clicks (Figure 4.19a), Difficulty in Integrating Information after Search Task (Figure 4.28a), and Difficulty in Determining when to Stop after Search Task (Figure 4.29a), the generalizability error rates slowly decrease until the number of participants is close to 100 per cognitive complexity level, followed by an acute decline to 0 when the number of participants is 100 per cognitive complexity level. For the rest of the variables, generalizability error rates decrease more gradually and become under than 0.1 when number of participants is between 85 and 95 per cognitive complexity level.

For the ones that post hoc results are not significant, (Figure 4.14b to Figure 4.31b), the changes in generalizability error rate are inconsistent. For Number of Unique Queries (Figure 4.14b), Clicks (Figure 4.16b), Query Diversity (Figure 4.19b), Type of Info Needed (Figure 4.20b), Expected Solution (Figure 4.21b), Difficulty in Deciding Information after Search Task (Figure 4.24b), and Difficulty in Determining when to Stop after Search Task (Figure 4.30b), the changing patterns of generalizability error rate are unstable, but the maximum values are less than 0.1. The generalizability error rates of the other variables increase from 0 as the number of participants increases, then suddenly drop to 0 when the number of participants is at 100 per cognitive complexity level.



(a) Number of Unique Queries (b) Number of Unique Queries

Figure 4.14: Generalizability error rate figures of Number of Unique Query from post hoc test

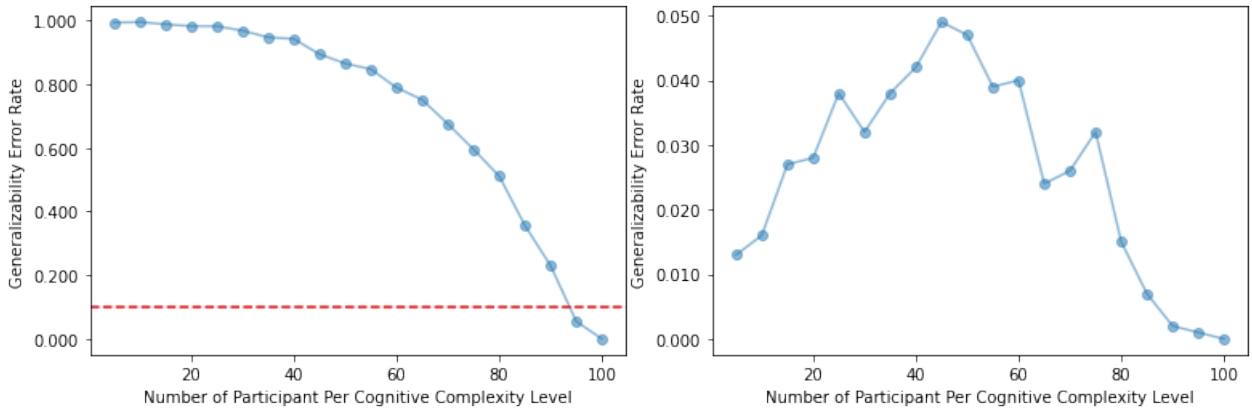


(a) Unique Query Terms (b) Unique Query Terms

Figure 4.15: Generalizability error rate figures of Unique Query Terms from post hoc test

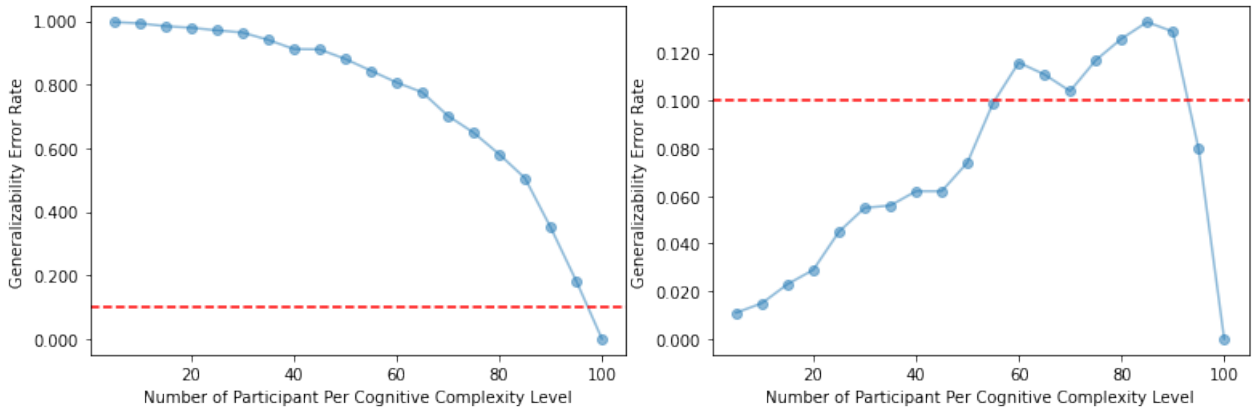
4.4 Reproducibility Error Rates for Dependent Variables

Generalizability concerns to what extent the results of one study can represent the outcomes of a larger population, which relies on how reliable the sample is. We have shown that a relatively satisfied amount of sample size for most IR research involving user behaviors should be around 85; however, some studies with fewer than 20 participants may argue that



(a) Clicks (b) Clicks

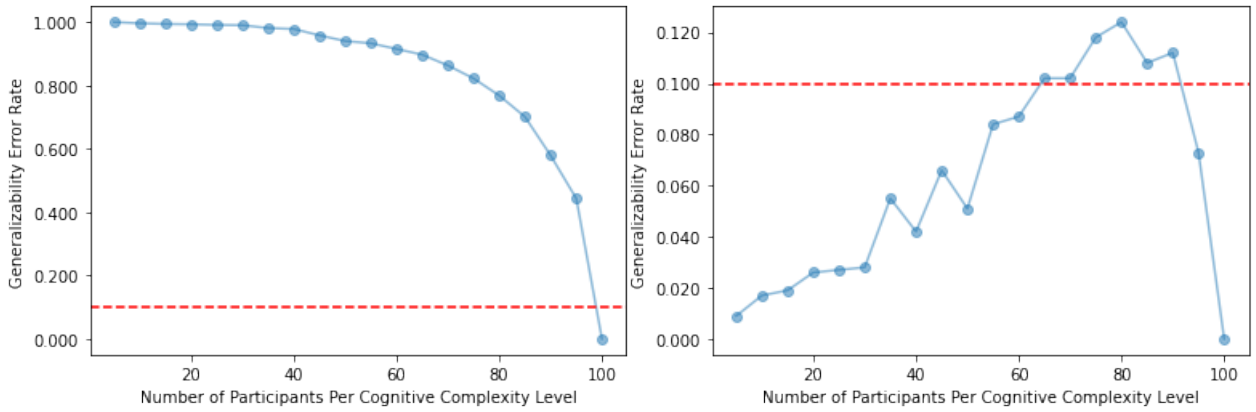
Figure 4.16: Generalizability error rate figures of Clicks from post hoc test



(a) URLs Visited (b) URLs Visited

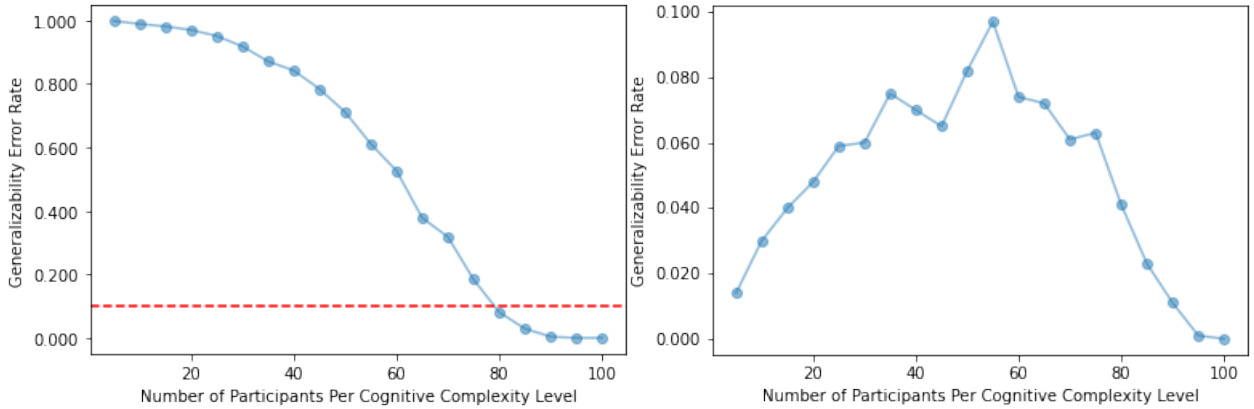
Figure 4.17: Generalizability error rate figures of URLs Visited from post hoc test

they derived the same conclusion more efficiently. The question thus follows is that whether their designs and results are reproducible? To evaluate the reproducibility of user studies, we generate reproducibility error rates figures of search behaviors and user experiences, which are mentioned in Section 3.6.3.



(a) Queries without Clicks (b) Queries without Clicks

Figure 4.18: Generalizability error rate figures of Queries without Clicks from post hoc test



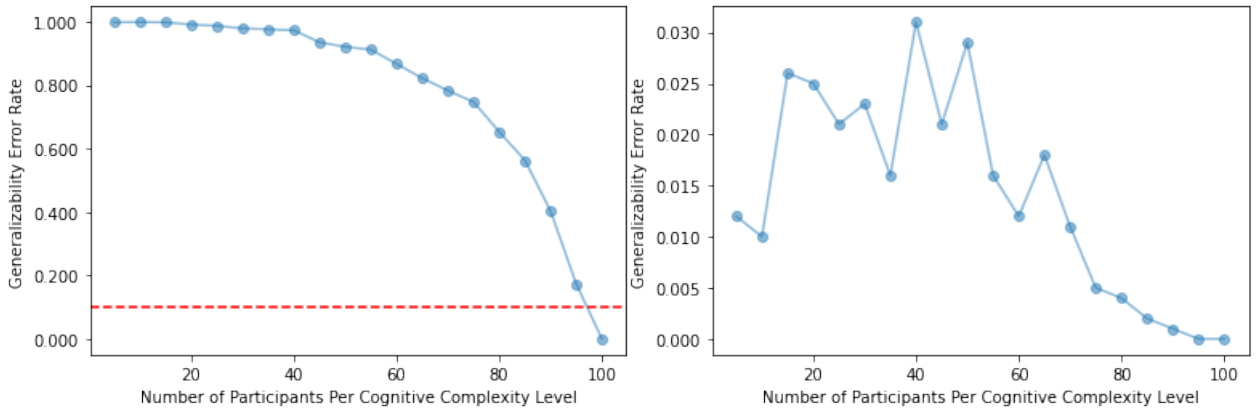
(a) Query Diversity (b) Query Diversity

Figure 4.19: Generalizability error rate figures of Query Diversity from post hoc test

4.4.1 Search Behaviors

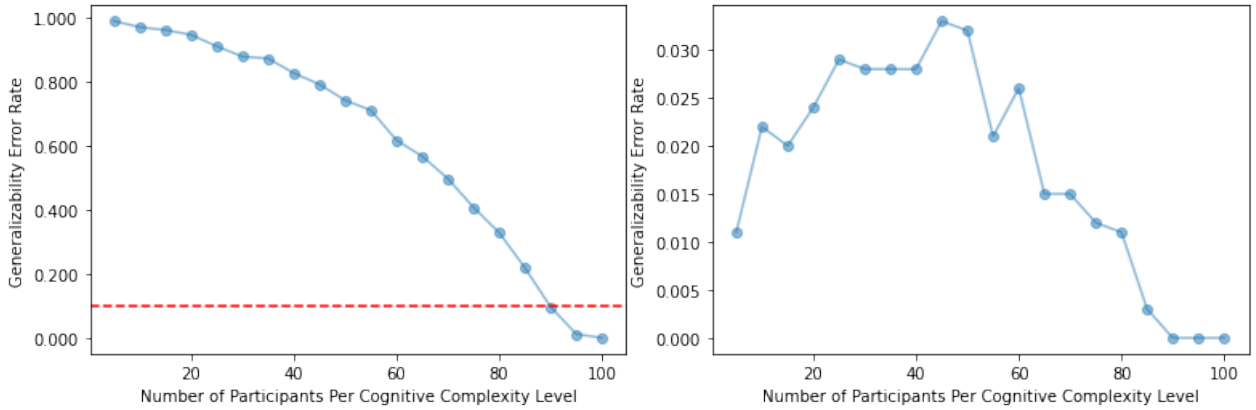
Figure 4.32 shows the reproducibility error rate figures for the search behaviors that are significant influenced by cognitive complexity. All of them demonstrate a bell-shaped curve that increasing from 0 to 0.5 at first, then dropping down to zero after a certain point.

For the four search behaviors that were non-significantly influenced by cognitive levels, Figure 4.33 demonstrates the reproducibility error rates of Query Length (Figure 4.33a) and URLs



(a) Types of Info Needed (b) Types of Info Needed

Figure 4.20: Generalizability error rate figures of Types of Info Needed from post hoc test



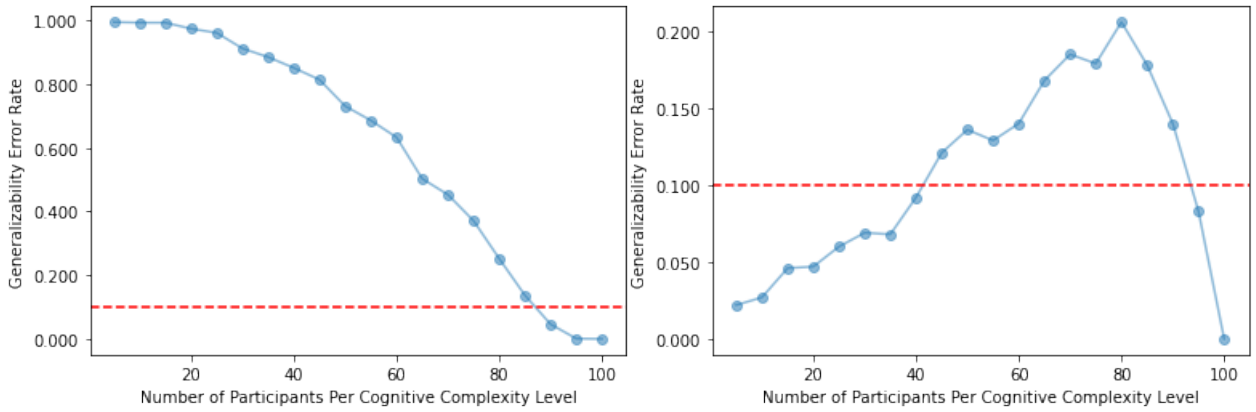
(a) Expected Solution (b) Expected Solution

Figure 4.21: Generalizability error rate figures of Expected Solution from post hoc test

Diversity (Figure 4.33c) increase as the number of participants increases, and then start to decrease after 80 and 85 participants per cognitive complexity level. For Query Term Diversity (Figure 4.33b), the rates are always approximate to zero.

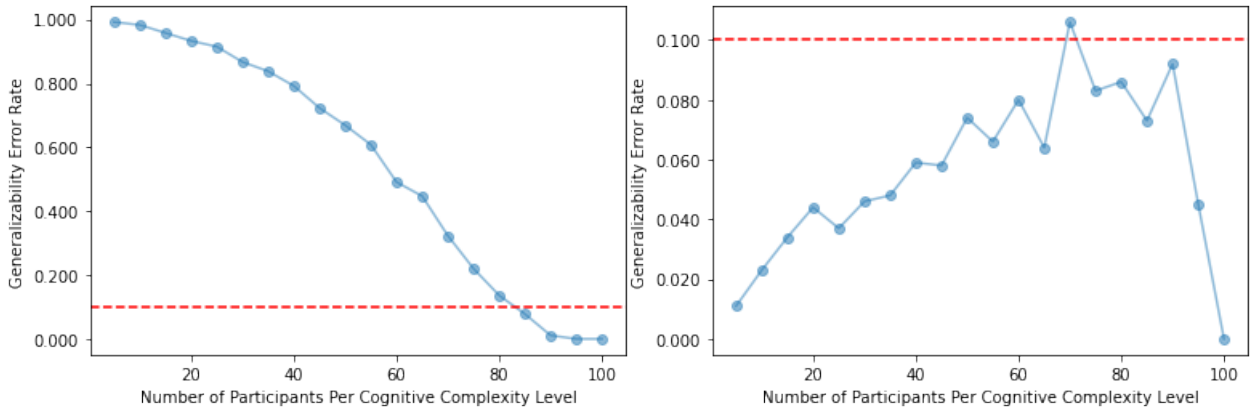
4.4.2 Task Complexity

The general trends of Types of Info Needed (Figure 4.34a) and Expected Solution (Figure 4.34c) increase as the number of participants increases, and decrease when the numbers of



(a) Pre-Search Task Difficulty (Search) (b) Pre-Search Task Difficulty (Search)

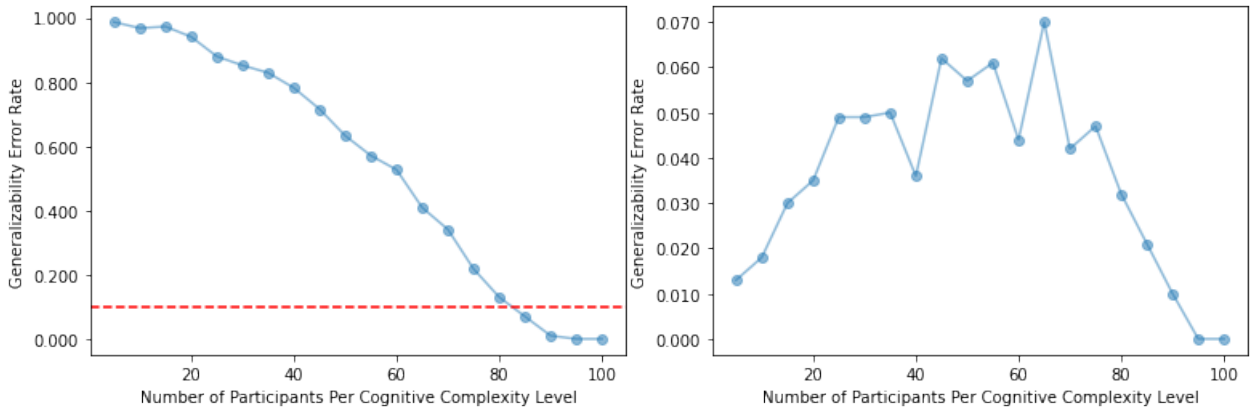
Figure 4.22: Generalizability error rate figures of Pre-Search Task Difficulty (Search) from post hoc test



(a) Pre-Search Task Difficulty (Understand) (b) Pre-Search Task Difficulty (Understand)

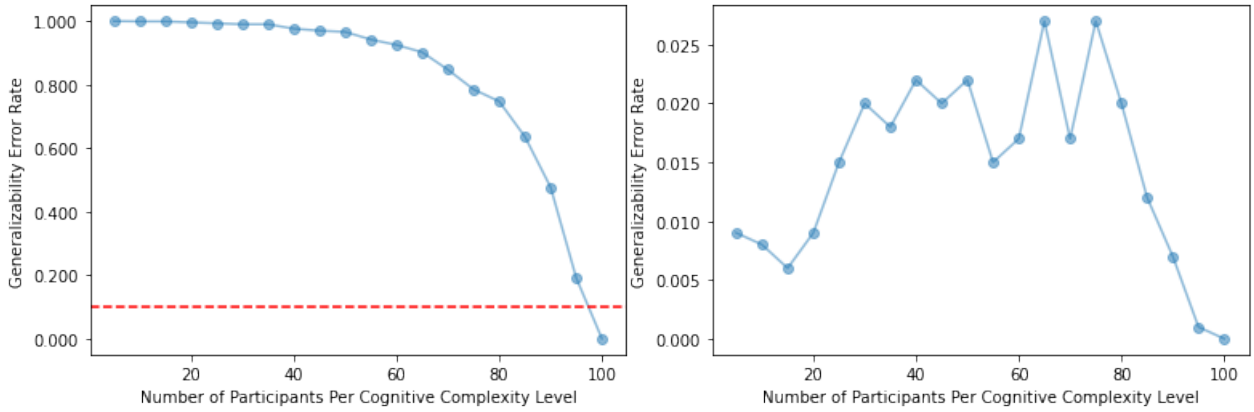
Figure 4.23: Generalizability error rate figures of Pre-Search Task Difficulty (Understand) from post hoc test

participants are 60 and 80 per cognitive complexity level, respectively. The changes in error rate for Steps to Complete (Figure 4.34b) is similar to the trend in Figure 4.9: the rate starts to decrease to zero at 60 participants per cognitive complexity level, although the overall fluctuation of the rate is small.



(a) Pre-Search Task Difficulty (Decide) (b) Pre-Search Task Difficulty (Decide)

Figure 4.24: Generalizability error rate figures of Pre-Search Task Difficulty (Decide) from post hoc test

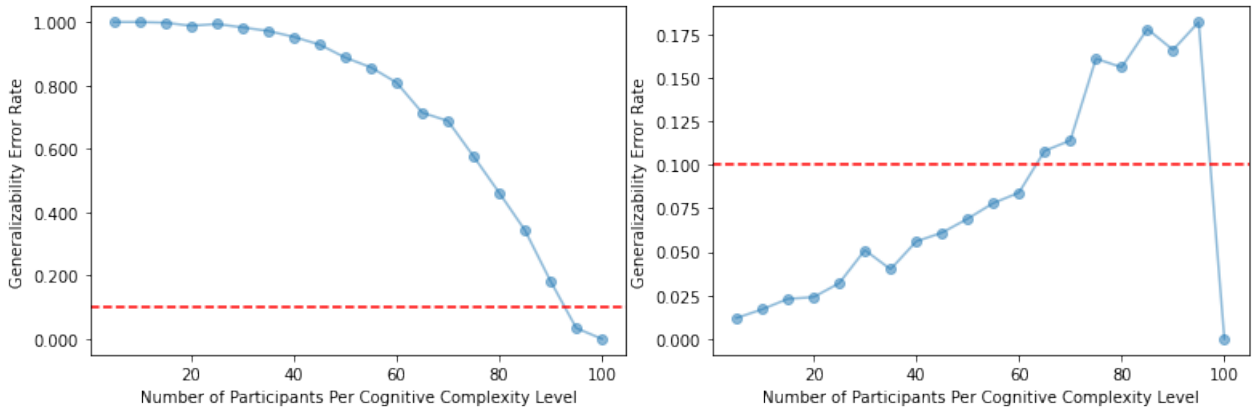


(a) Pre-Search Task Difficulty (Integrate) (b) Pre-Search Task Difficulty (Integrate)

Figure 4.25: Generalizability error rate figures of Pre-Search Task Difficulty (Integrate) from post hoc test

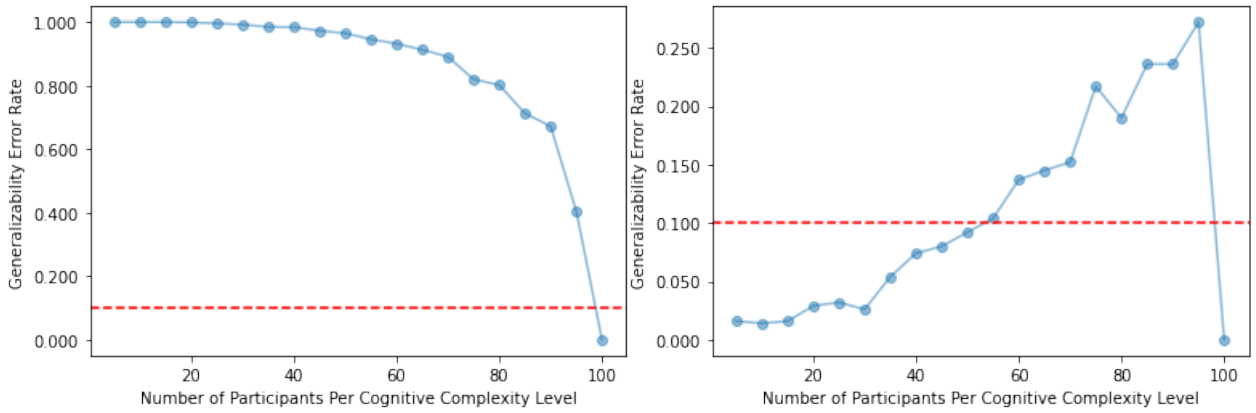
4.4.3 Task Difficulty

Cognitive complexity levels significantly influenced all items in Pre-Search Task Difficulty, and the general trends of the reproducibility error rates are consistent with the other significant results, showing an increase from 0 to 0.5 at first following with a decrease to 0. However, the turning points are different. For Difficulty in Searching Information (Figure 4.35a) and Difficulty of Determining When to Stop (Figure 4.35e), the rates start to de-



(a) Pre-Search Task Difficulty (Enough) (b) Pre-Search Task Difficulty (Enough)

Figure 4.26: Generalizability error rate figures of Pre-Search Task Difficulty (Enough) from post hoc test

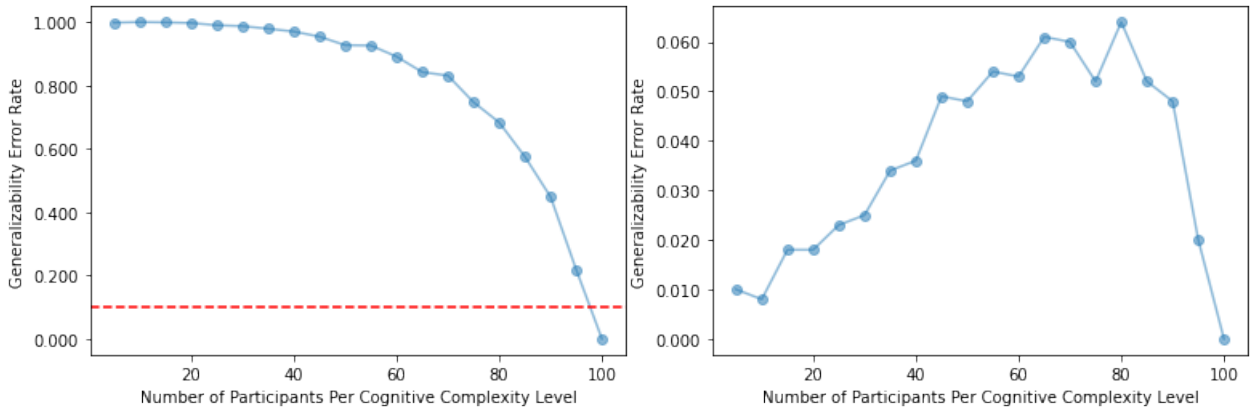


(a) Post-Search Task Difficulty (Understand) (b) Post-Search Task Difficulty (Understand)

Figure 4.27: Generalizability error rate figures of Post-Search Task Difficulty (Understand) from post hoc test

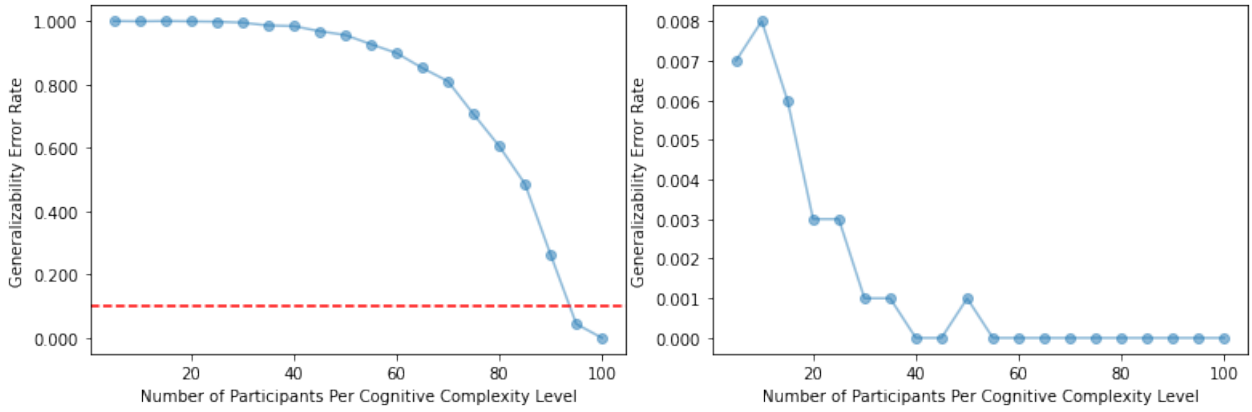
crease when the numbers of participants are around 30. For Difficulty in Understanding the Information (Figure 4.35b), Difficulty in Deciding Information (Figure 4.35c) and Difficulty of Integrating Information (Figure 4.35d), the rates start to decrease when the numbers of participants are around 65 per cognitive complexity level.

Figure 4.36 shows the reproducibility error rates of the items in post-search task difficulty. For Difficulty in Understanding Information (Figure 4.36b), Difficulty in Integrating Infor-



(a) Post-Search Task Difficulty (Integrate) (b) Post-Search Task Difficulty (Integrate)

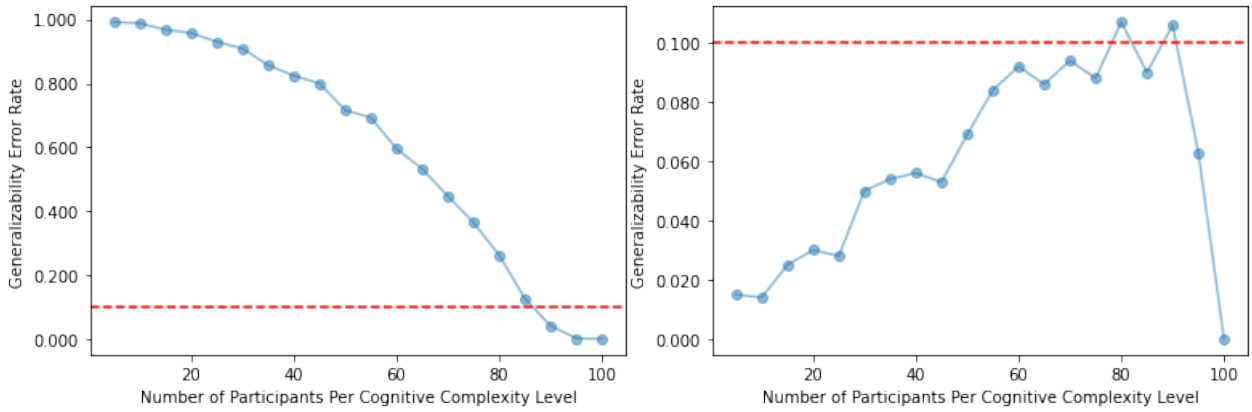
Figure 4.28: Generalizability error rate figures of Post-Search Task Difficulty (Integrate) from post hoc test



(a) Post-Search Task Difficulty (Enough) (b) Post-Search Task Difficulty (Enough)

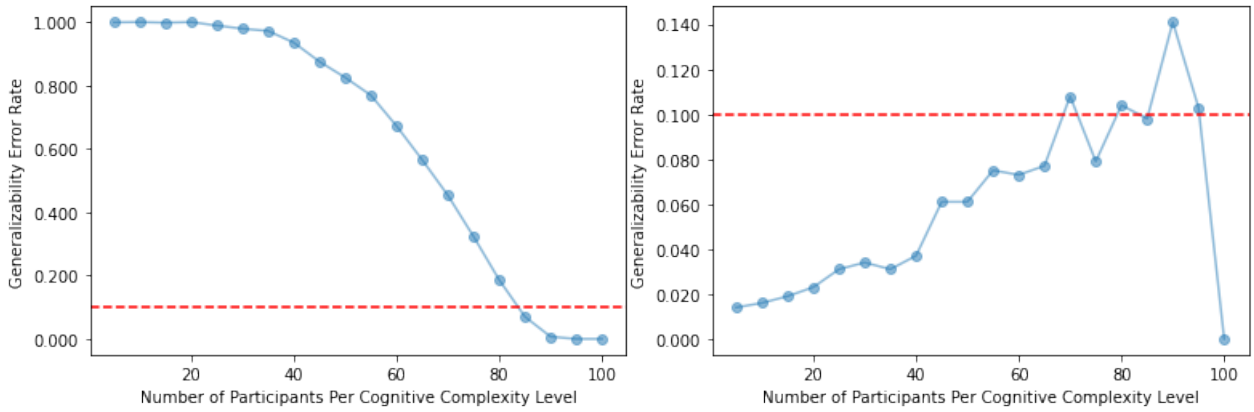
Figure 4.29: Generalizability error rate figures of Post-Search Task Difficulty (Enough) from post hoc test

mation (Figure 4.36d) and Difficulty of Determining When to Stop (Figure 4.36e), the rates increase with the number of participants and declining when the the sample size is up to 70, 60, and 40, respectively. For Difficulty in Searching Information (Figure 4.36a) and Difficulty in Deciding Information (Figure 4.36c), the general trends are similar to the rates in Figure 4.11.



(a) Concentration (b) Concentration

Figure 4.30: Generalizability error rate figures of Concentration from post hoc test

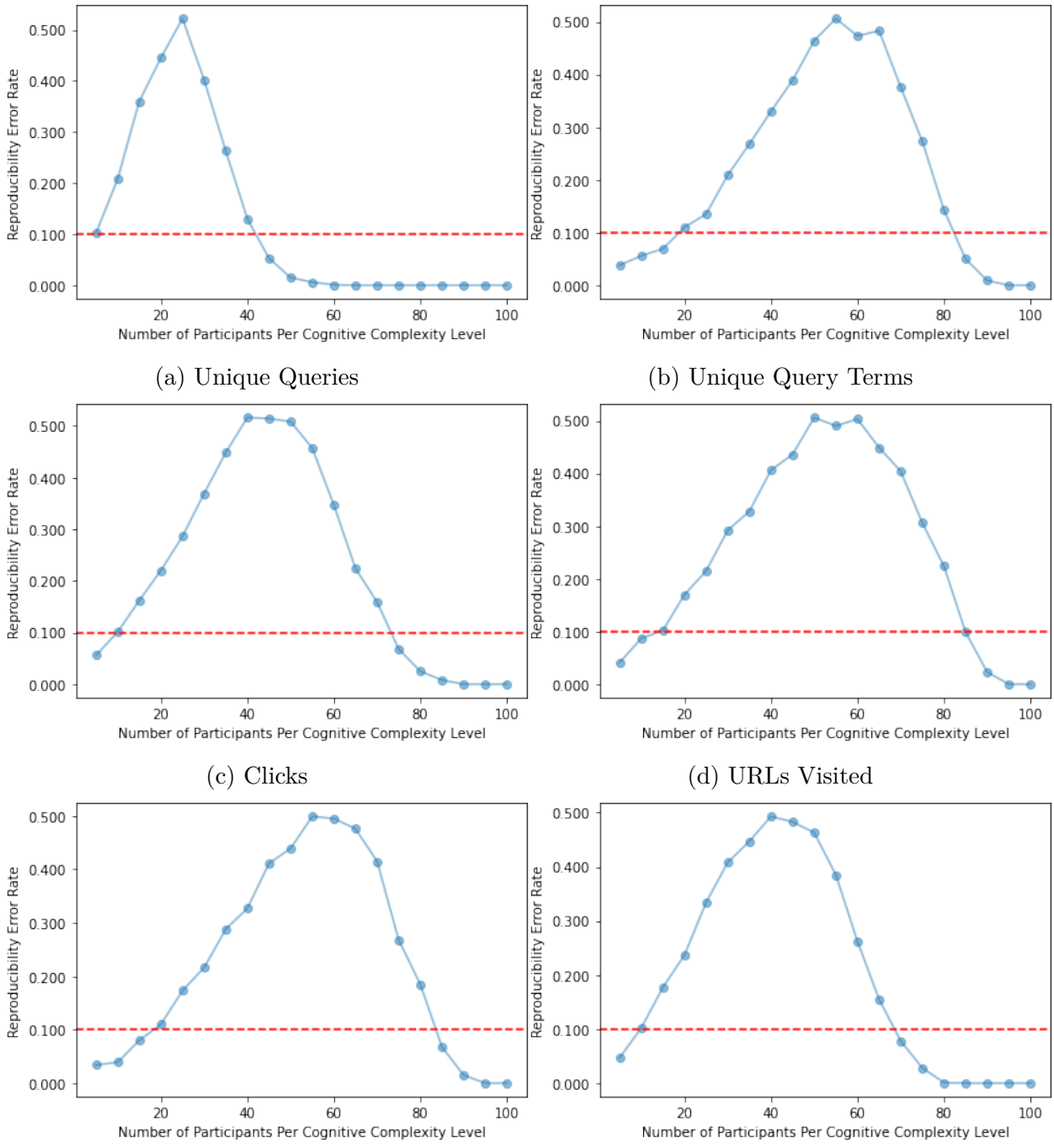


(a) Overall Difficulty (b) Overall Difficulty

Figure 4.31: Generalizability error rate figures of Overall Difficulty from post hoc test

4.4.4 Enjoyment, Engagement and Concentration

Similar to other significant results, changes in reproducibility error rate for Concentration also showed a bell-shaped curve that turning when the number of participants is 65 (Figure 4.37c). The other two figures demonstrate the changes in error rates for the non-significant results, which fluctuated between 0.05 and 0. For Enjoyment (Figure 4.37a) and Engagement (Figure 4.37b), reproducibility error rates stabilize and start to decrease when the sample size is 60 per cognitive complexity level.



(a) Unique Queries

(b) Unique Query Terms

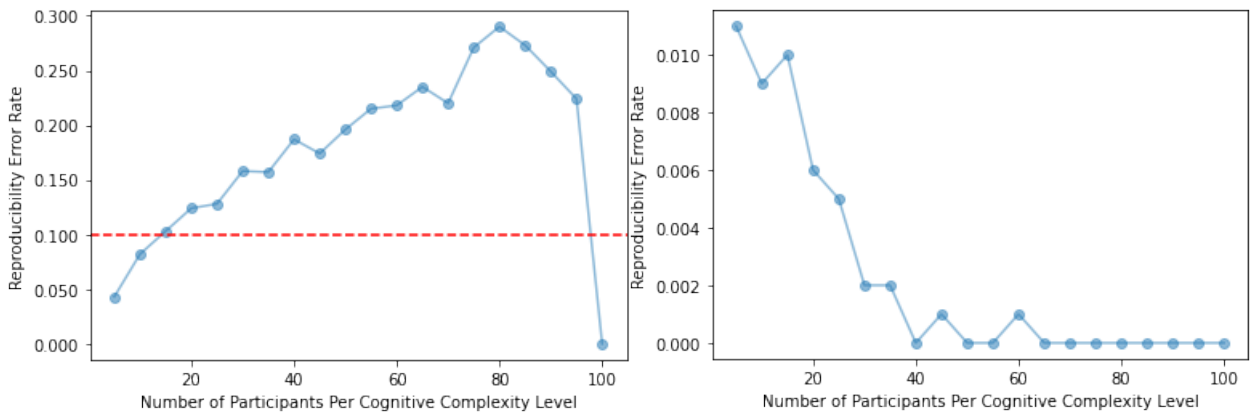
(c) Clicks

(d) URLs Visited

(e) Queries w/out Clicks

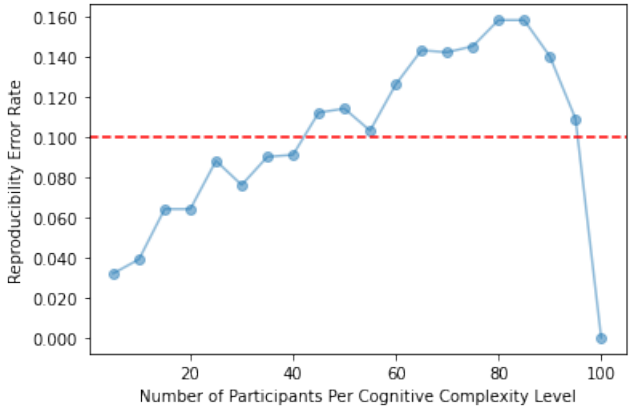
(f) Query Diversity

Figure 4.32: Reproducibility error rate figures of 6 Search Behaviors with significant influences



(a) Query Length

(b) Query Term Diversity

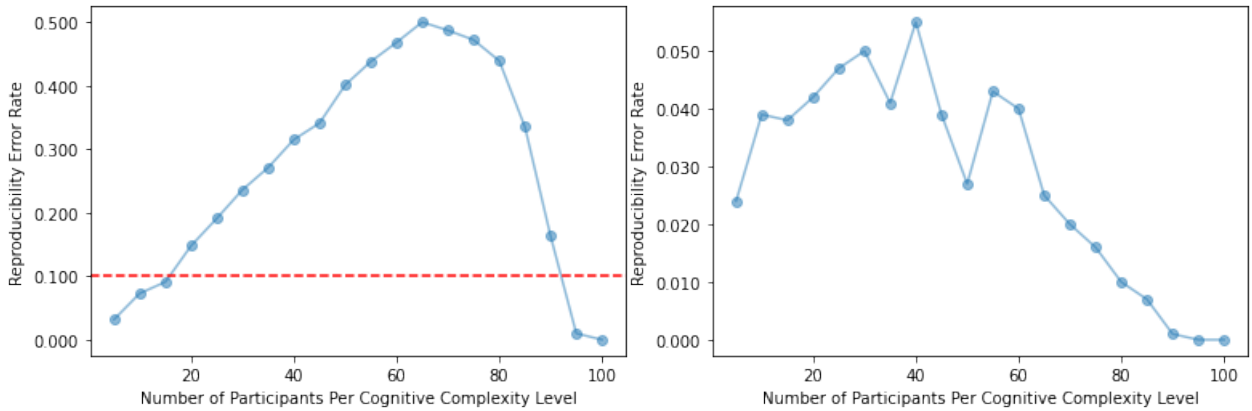


(c) URL Diversity

Figure 4.33: Reproducibility error rate figures of 3 Search Behaviors without significant influences

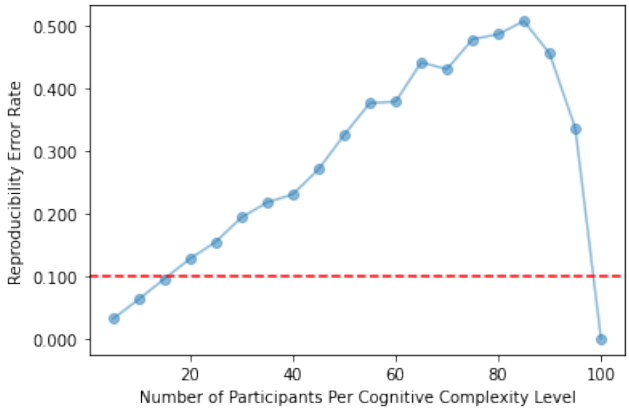
4.4.5 Overall Difficulty and Satisfaction

Overall Difficulty (Figure 4.38a) shows the similar pattern as the other significant results: The rate increases as the number of participants accumulates, and decreases after the sample size is 30. For the items of Satisfied with Solution (Figure 4.38b) and Satisfied with Strategy (Figure 4.38c), the rates are both very unstable but not very large.



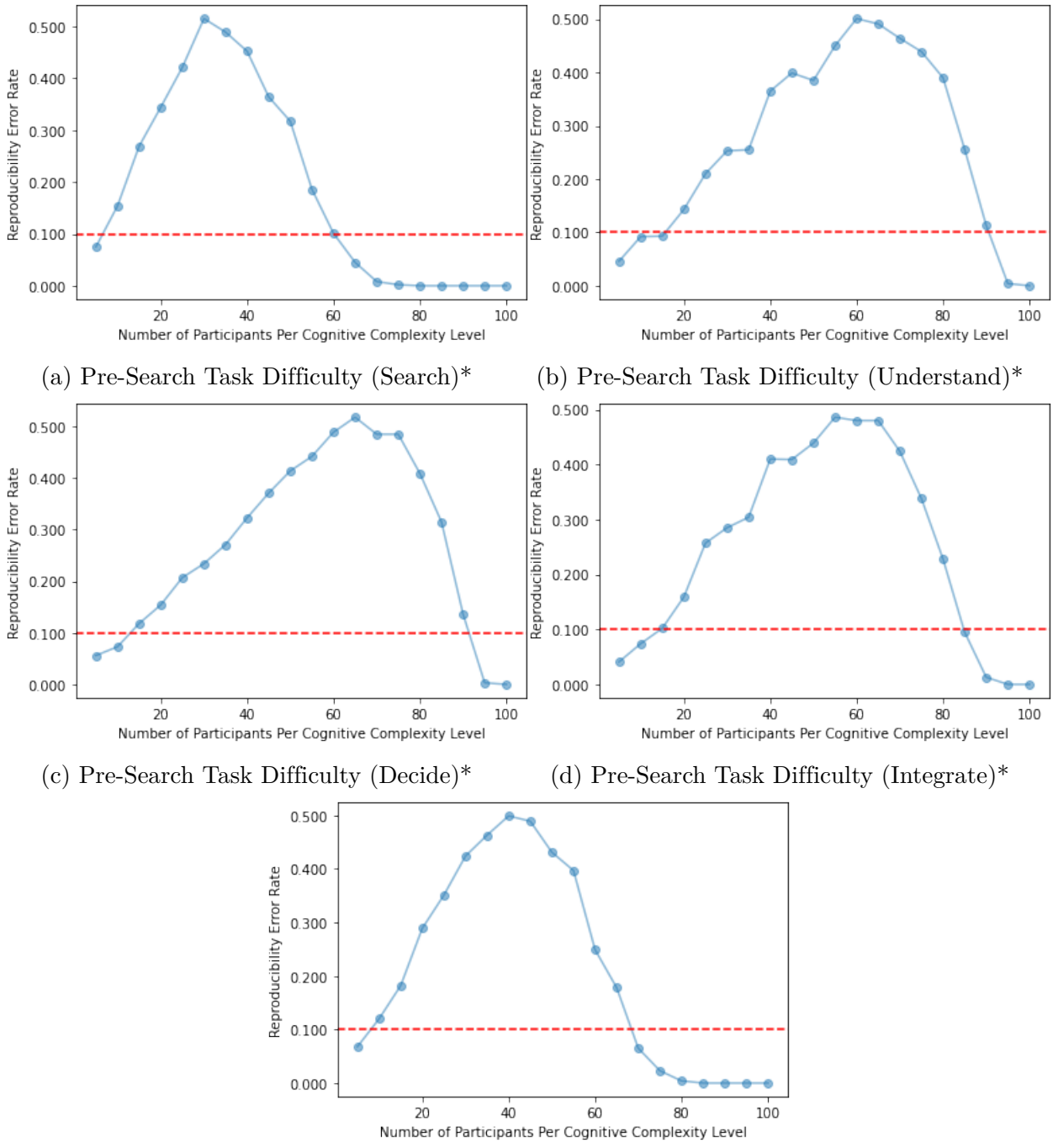
(a) Types of Info Needed*

(b) Steps to Complete



(c) Expected Solution*

Figure 4.34: Reproducibility error rate figures of Pre-Search Task Complexity. * means $p < 0.01$ at 100 participants per cognitive complexity level



(a) Pre-Search Task Difficulty (Search)*

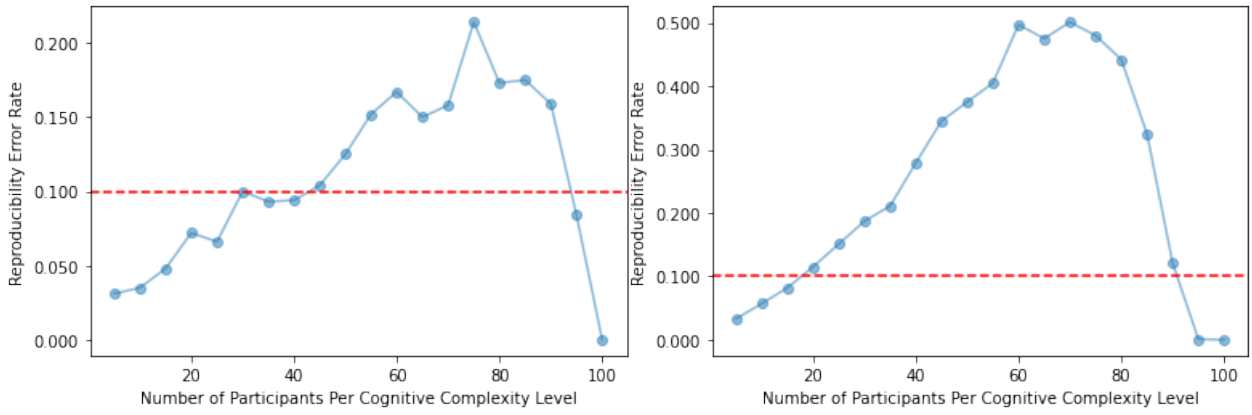
(b) Pre-Search Task Difficulty (Understand)*

(c) Pre-Search Task Difficulty (Decide)*

(d) Pre-Search Task Difficulty (Integrate)*

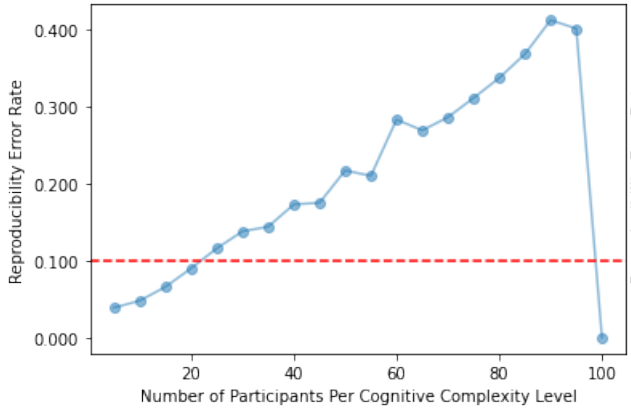
(e) Pre-Search Task Difficulty (Enough)*

Figure 4.35: Reproducibility error rate figures of Pre-Search Task Difficulty. * means p < 0.01 at 100 participants Per cognitive complexity level

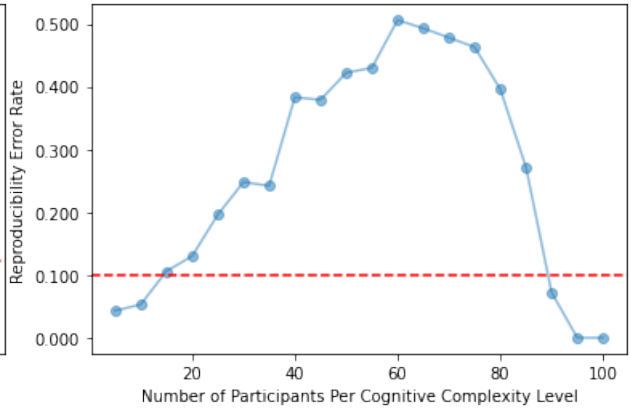


(a) Post-Search Task Difficulty (Search)

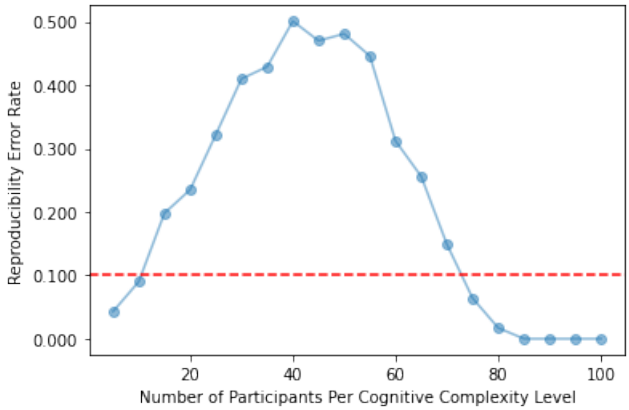
(b) Post-Search Task Difficulty (Understand)*



(c) Post-Search Task Difficulty (Decide)

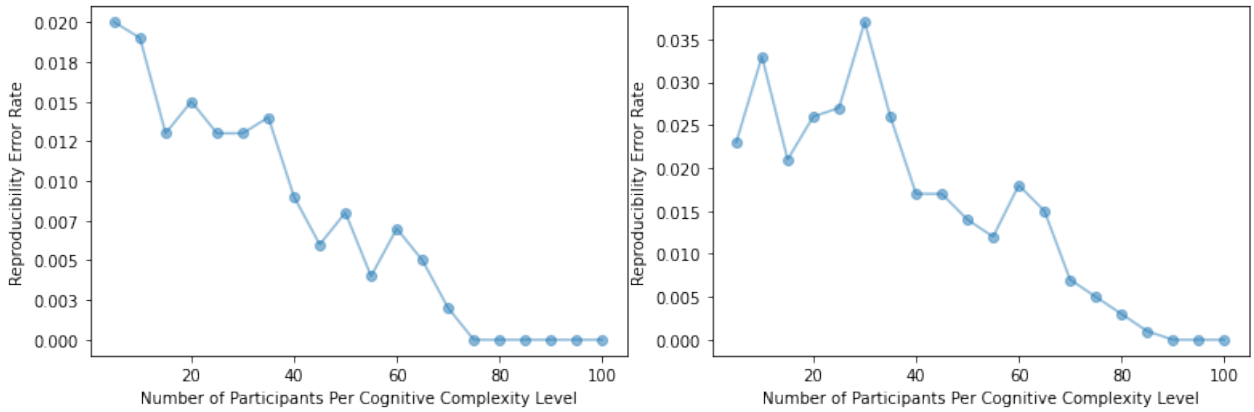


(d) Post-Search Task Difficulty (Integrate)*



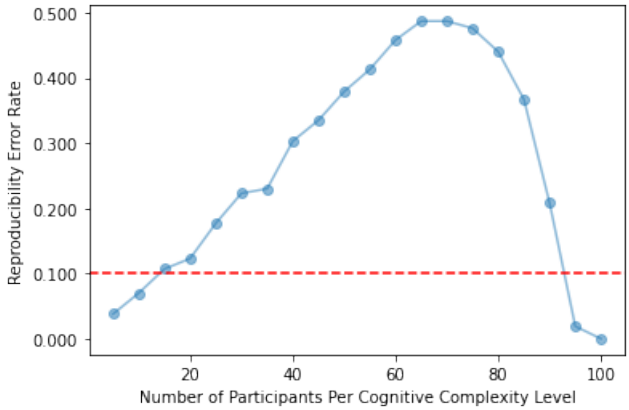
(e) Post-Search Task Difficulty (Enough)*

Figure 4.36: Reproducibility error rate figures Post-Search Task Difficulty. * means $p < 0.01$ at 100 participants per cognitive complexity level



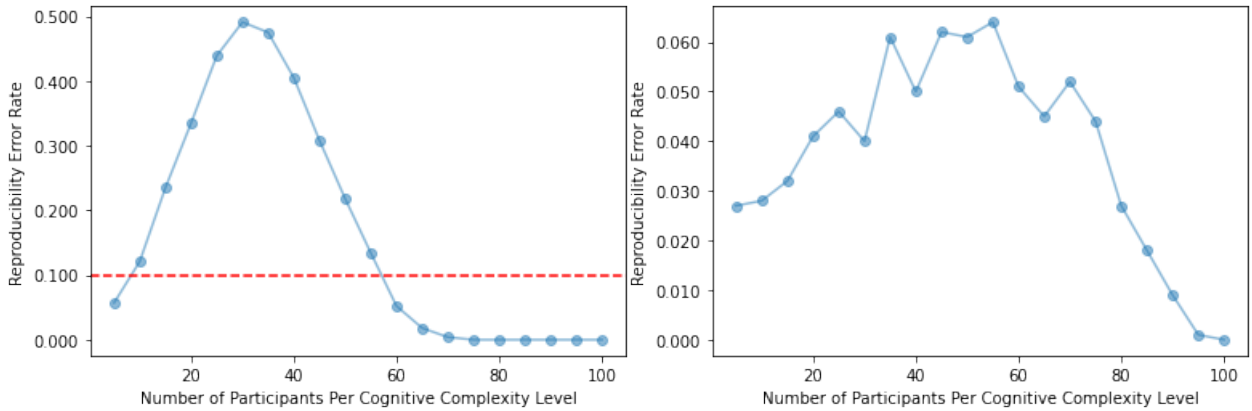
(a) Enjoyment

(b) Engagement



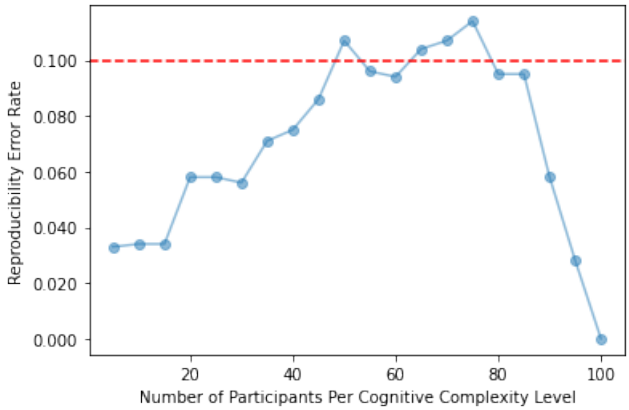
(c) Concentration*

Figure 4.37: Reproducibility error rate figures of Enjoyment, Engagement and Concentration. * means $p < 0.01$ at 100 participants per cognitive complexity level



(a) Overall Difficulty*

(b) Satisfied with Solution



(c) Satisfied with Strategy

Figure 4.38: Reproducibility error rate figures of Overall Difficulty and Satisfaction * means $p < 0.01$ at 100 participants per cognitive complexity level

Chapter 5

Discussion

5.1 Online User Study vs. Lab User Study

The results of the current research showed that cognitive complexity has a significant impact on a number of search behaviors and user experiences including Unique Queries, Unique Query Terms, Clicks, URLs Visited, Queries without Clicks and Query Diversity. Further tests revealed that significant changes do not occur between every two cognitive level, but the results for the dependent variables in the more demanding cognitive levels (i.e. Evaluate and Create levels) are always significantly differed from those in the easier levels (i.e. Remember and Understand levels).

The current study repeated the study from Kelly et al. (2015), but under an online environment. It successfully reproduced the previous results by finding the relationships between cognitive complexity levels and a number of search-related dependent variables when the experiment is conducted online. Both the current results and results in Kelly et al. (2015) showed that human search behavior becomes more complex, requiring more process such as information-integration and reorganization, as the cognitive demands of the tasks increases, and this behavioral change occurs no matter the environment is online or offline.

We found a similar change with user experiences. The results showed that participants reported greater concentration and felt they need to be more engaged in the tasks that have

greater cognitive demands. It is also consistent with conclusion in Kelly et al. (2015) and the common sense that tasks with lower cognitive complexity are much easier to resolve.

In order to make sure the results from the current study is comparable to results in Kelly et al. (2015), assuming her conclusion is correct and reliable enough to be the standard, it further computed offline results in Kelly et al. (2015) to examine the changes in error rate (Figure 4.6). The findings showed that error rate diminished below 10% level when the number of participants approaches 65, supporting that current online study is also powerful enough to be used for subsequent comparisons. This result also indicates that it is feasible for other offline studies to be conducted online.

5.2 Generalizability of Online User Studies

From the results that are statistically significant, a general trend appeared, showing that as the number of participants increases from 5 to 100, error rate decreases from 1 to 0, which means that the more the participants, the more accurate the result and the more reliable and generalizable the system. This finding is supported by Sakai (2016) who suggested that one study is underpowered and may be unreliable if its sample size is limited.

We found that when the number of participants is between 80 and 90, the error rates for most of the dependent variables becomes less than 0.1 or approaching zero. Different variables revealed two distinct patterns. For instance, the error rate decreases steeply as the number of participants accumulates, which approaches zero and changes little when the sample size is 40, as shown in Figure 4.7a (Number of Unique Queries). We suggest that this change is related to the significance level of these variables that are extremely small, so a little change in the number of participants would greatly affect its results and adjust the error rate. The other situation is that the error rate decreases slowly as the sample size grows from 5; then it

shows an accelerated decrease as the number of participants approaches 80, such as Expected Solution in Figure 4.9c.

Although people may argue that the sample size does not necessarily to be over 80 to make the results reliable, these patterns were variable-dependent and task-sensitive. In order to have error rate of most dependent variables be 10% level and lower, the number of participants should between 80 and 90, so that the results of online user studies that involving search behaviors and user experiences can be reliable and generalizable to larger population.

Admittedly, there are certain situations showing non-significant changes in error rates as the number of participants increases. For most of these cases, the error rate is already less than 0.1 even with few participants, which left the rate little room to give significant decrease significantly as the sample size grows. In fact, this situation usually happens when the dependent variables are subjective ratings, such as Engagement and Overall Satisfaction. These dependent variables are persistent for the majority of the population indicating a generally consistent attitude of the audience, which does not change with the sample size by common sense. Another pattern of the non-significance is that the change of error rates fluctuates when the sample size is small and then decreases to 0 as the number of participants is at a certain point, such as the figures of Query Term Diversity (Figure 4.8b), and Enjoyment (Figure 4.12a). It should be noticed that the unstable part of this change is still lower than 10% level and it is still a consistent with the pattern that as sample size is larger than 70, error rate approaches zero.

However, some situations showed a different trend. For example, the error rates of Query Length (Figure 4.8a) and URLs Diversity (Figure 4.8c) increments as the number of participants grows, and then it becomes stable and suddenly diminishes to zero when the number of participants approaches 100 level. The reason causing this situation is because these dependent variables do not have significant influences, and when the sample size is small, it is

already difficult to detect the significant differences, so the rates are close to 0. When the number of participants increases, there will be some cases appeared, where the significant differences were found. However, when the number of participants is large enough, the rate will become stable and then start to decrease, because the ground truth is significant differences were not found. Although the rates start to decrease when the number of participants is close to 100, we still do not know what will happen after 100 participants, so we may need larger sample size to justify our thoughts and we expect to see the rates decrease to 0 smoothly after 100 participants. Another example of a different trend is Decide in Post-Search Task Difficulty (Figure 4.11c), the error rate keeps increasing and suddenly drop to 0 when the number of participants is at 100. It is possible that the Type II error may explain for these situations, but more participants are required for further investigation. If the number of participants is over 100 per cognitive level, the changes in error rates would be more clearer. Nonetheless, an increase in error rate is not only insignificant but also rarely occurs.

Generally speaking, the current study calculating error rates for a variety of dependent variables supports and extends the previous literature. It suggests that for the research involving human search behaviors, in order to make sure the system is testable and reliable, the sample size should be over 80; otherwise, the results may be under the risk of being underpower. More details of each dependent variable are shown in Table 5.1.

5.3 Reproducibility of Online User Studies

Regarding the result in Section 4.4, the general trend for most of the variables showed that reproducibility error rates increase from 0 to 50% as the number of participants increases, up to a point when it starts to decrease to zero. It reveals that the results of these particular

Table 5.1: Generalizability: suggestion sample size for each dependent variables. * means p < 0.01

Dependent Variables	Suggestion
Unique Queries*	40 Participants
Query Length	Need Larger Sample Size
Unique Query Terms*	80 Participants
Clicks*	70 Participants
URLs Visited*	80 Participants
Queries w/out Clicks*	80 Participants
Query Diversity*	65 Participants
Query Term Diversity	No Need to Concern
URL Diversity	No Need to Concern
Types of Info Needed*	90 Participants
Steps to Complete	No Need to Concern
Expected Solution*	Need Larger Sample Size
Pre-Search Task Difficulty (Search)*	55 Participants
Pre-Search Task Difficulty (Understand)*	90 Participants
Pre-Search Task Difficulty (Decide)*	90 Participants
Pre-Search Task Difficulty (Integrate)*	85 Participants
Pre-Search Task Difficulty (Enough)*	65 Participants
Post-Search Task Difficulty (Search)	No Need to Concern
Post-Search Task Difficulty (Understand)*	90 Participants
Post-Search Task Difficulty (Decide)	Need Larger Sample Size
Post-Search Task Difficulty (Integrate)*	90 Participants
Post-Search Task Difficulty (Enough)*	70 Participants
Enjoyment	No Need to Concern
Engagement	No Need to Concern
Concentration*	90 Participants
Overall Difficulty*	55 Participants
Satisfied with Solution	No Need to Concern
Satisfied with Strategy	No Need to Concern

variables are reproducible when the number of participants is great or extremely small. Specifically, the reproducibility error rates are less than 10% when the number of participants is between 85 and 95, which is a similar finding to the generalizability results in Section 5.2. Both of them show and support that results of user studies are more reliable, generalizable, and reproducible when the sample size is around 85.

It is interesting that the reproducibility error rates are also close to zero when the numbers of participants are extremely low, but this finding is acceptable. Because when the number of participants is small, the results are highly different from the findings resulted with 100 participants sample, which can be supported by the figures in Section 4.2. The generalizability error rates are close to 1 when the number of participants is small, meaning that they are highly unreliable and questionable in generalizability. Thus, although the reproducibility seems to be positive the results are in fact dubious when the number of participants is small.

For the dependent variables (e.g. Query Length and URL Diversity) where the reproducibility error rates increase as the number of participants grows and suddenly diminish to zero at the 100 participants per cognitive complexity level, we speculate that this is because the current number of participants is still insufficient enough to show the entire changing pattern. If we can invite more participants, we should expect a pattern that reproducibility error rate increases from 0 to 50% as the number of participants increase and decreases from 50% to 0 more gradually at a certain point.

Overall, we suggest that the sample size should be over 85 to ensure the results are replicable when human search behaviors and personal experiences are involved in the research. Although results can also be reproduced when the sample size is limited (5 - 10), these findings are not reliable or generalizable. More details of each dependent variable are shown in Table 5.2.

Table 5.2: Reproducibility: suggestion sample size for each dependent variables. * means p < 0.01

Dependent Variables	Suggestion
Unique Queries*	45 Participants
Query Length	Need Larger Sample Size
Unique Query Terms*	85 Participants
Clicks*	75 Participants
URLs Visited*	85 Participants
Queries w/out Clicks*	85 Participants
Query Diversity*	70 Participants
Query Term Diversity	Need Larger Sample Size
URL Diversity	No Need to Concern
Types of Info Needed*	95 Participants
Steps to Complete	No Need to Concern
Expected Solution*	Need Larger Sample Size
Pre-Search Task Difficulty (Search)*	60 Participants
Pre-Search Task Difficulty (Understand)*	95 Participants
Pre-Search Task Difficulty (Decide)*	95 Participants
Pre-Search Task Difficulty (Integrate)*	95 Participants
Pre-Search Task Difficulty (Enough)*	70 Participants
Post-Search Task Difficulty (Search)	Need Larger Sample Size
Post-Search Task Difficulty (Understand)*	95 Participants
Post-Search Task Difficulty (Decide)	Need Larger Sample Size
Post-Search Task Difficulty (Integrate)*	90 Participants
Post-Search Task Difficulty (Enough)*	75 Participants
Enjoyment	No Need to Concern
Engagement	No Need to Concern
Concentration*	95 Participants
Overall Difficulty*	60 Participants
Satisfied with Solution	No Need to Concern
Satisfied with Strategy	No Need to Concern

5.4 Effects to Interactive Information Retrieval Community

From our study, we can see for most of figures in the generalizability, the error rate trend shows the generalizability is negative when the number of participants is relatively small,

and for the figures in reproducibility, we found the results are very bad when the number of participants is at middle-level, which is around 50 participants. This may give us a chance to question the generalizability and reproducibility from the lab user study from the previous years because from the literature review, we found most of these lab user studies recruited 20 to 55 participants. However, because we only conducted the online user study, someone may argue that the results from an online user study would be different from the results from the lab user study. Although we prove that the general results from our online user study and the general results from the lab user study (Kelly et al., 2015) are consistent, there are still slight differences in details. Therefore, we think it is still too early to confirm that the results from lab user studies from previous years are unreliable in generalizability and reproducibility, but we suggest that researchers need to use a relatively large number of participants for the interactive information retrieval user study in the future. We can say it is highly possible that these results from the lab user studies in previous years are unreliable in generalizability and reproducibility. It would be interesting to further examine the same questions using similar methods in offline lab user studies. If the similar results are found, we can confirm that the results of lab user studies from previous years are highly unreliable in generalizability and reproducibility.

5.5 Effects to Other Independent Variables in Human Subjects Studies

In our online user study, we used different five cognitive complexity levels as the independent variables, however, these are not in use in every human factor user studies. Therefore, we could not say our results can represent the results in all human factor user studies with different independent variables. However, we think the results would be similar if the

independent variables are complexity related or psychology-related because our independent variable is based on the cognitive process dimensions in psychology. It would be better if we can further investigate the same questions with different independent variables.

Chapter 6

Conclusions

Our research reproduced lab user study (Kelly et al., 2015) in an online environment. It supported the previous results by finding that the cognitive complexity level of the search task may have an impact on various searching behaviors and their personal experiences. The current results showed that this adaptation is not exclusive to offline situations but also happens with the online situation. It indicates results from two user studies are consistent, if the system is rigorously designed. Because of the consistent results from two user studies, it is more reliable to use the data to investigate the generalizability and reliability of results from our online user study, and then question the generalizability and reliability of the lab user study.

In order to investigate the generalizability and reproducibility of the results, we established two methodologies for evaluating the generalizability and reproducibility: 1) generalizability error rate, 2) reproducibility error rate. By using these two methodologies, we generate the error rate figures for generalizability and reproducibility. From these error rate figures, we can provide commonly guidelines regarding the optimal number of participants for search engine user studies. Both generalizability error rate and the reproducibility error rate reduced to zero or almost zero when the sample size is around 85 for over half of the dependent variables examined. However, it should not be neglected that the changes in error rate for the other variables are not clear in the current experiment; therefore, we recommend future study with more rigorous experimental design to recruit even more participants to clarify

the changes. Nevertheless, the current results provided a minimum range of sample size for user studies. Also with these error rate figures, researchers can understand the relationship between sample size and the risks of making conclusions. We suggest researchers recruit a relatively large number of participants for interactive information retrieval user study in the future.

However, we could not say our results can assert that the results in lab user studies from previous years are unreliable in generalizability and reproducibility because we only tried the online user studies now. Although we found the results from both online and lab user study are consistent, there are still differences in details. Therefore, we could only say it is highly possible that the results in lab user studies from previous years are unreliable in generalizability and reproducibility. If we can examine the same research questions using similar methods in offline lab user studies, it would be better to make a conclusion. Also, we only tried one kind of independent variables. Although five cognitive complexity level can represent to most independent variables because most of the independent variables are different complexity level from our research, it cannot represent all the independent variables. It would be interesting if we change the independent variables for further investigation.

In summary, we successfully established two methodologies in examining the generalizability and reproducibility, and provide guidelines regarding the optimal number of participants for search engine user studies. However, there are some future works need to be done. First, from our research, we used simulation method to calculate generalizability and reproducibility, this simulation method assumes 100 participants per complexity level are correct and enough, however, there are some cases show 100 participants per level are not enough, so we still need more participants. Then it would be interesting to further examine the same questions using similar methods in offline lab user studies or with different independent variables.

Bibliography

Arapakis, I., Bai, X., & Cambazoglu, B. B. (2014). Impact of response latency on user behavior in web search, In *Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval*, Gold Coast, Queensland, Australia, Association for Computing Machinery. https://doi.org/10.1145/2600428.2609627

Arguello, J., Wu, W.-C., Kelly, D., & Edwards, A. (2012). Task complexity, vertical display and user interaction in aggregated search, In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, Oregon, USA, Association for Computing Machinery. https://doi.org/10.1145/2348283.2348343

Avula, S., Chadwick, G., Arguello, J., & Capra, R. (2018). Searchbots: User engagement with chatbots during collaborative search, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176380

Azzopardi, L. (2014). Modelling interaction with economic models of search, In *Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval*, Gold Coast, Queensland, Australia, Association for Computing Machinery. https://doi.org/10.1145/2600428.2609574

Barreda-Ángeles, M., Arapakis, I., Bai, X., Cambazoglu, B. B., & Pereda-Baños, A. (2015). Unconscious physiological effects of search latency on users and their click behaviour, In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, Association for Computing Machinery. https://doi.org/10.1145/2766462.2767719

Bentley, F. R., Daskalova, N., & White, B. (2017). Comparing the reliability of amazon mechanical turk and survey monkey to traditional market research surveys, In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, Denver, Colorado, USA, Association for Computing Machinery. https://doi.org/10.1145/3027063.3053335

Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability, In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, Association for Computing Machinery. https://doi.org/10.1145/345508.345543

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), https://doi.org/10.1177/1745691610393980, 3–5. https://doi.org/10.1177/1745691610393980

Burton, R., & Collins-Thompson, K. (2016). User behavior in asynchronous slow search, In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, Association for Computing Machinery. https://doi.org/10.1145/2911451.2911541

Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk, In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, Singapore, Association for Computational Linguistics.

Capra, R., Arguello, J., Crescenzi, A., & Vardell, E. (2015). Differences in the use of search assistance for tasks of varying complexity, In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, Association for Computing Machinery. https://doi.org/10.1145/2766462.2767741

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? a comparison of participants and data gathered via amazon's mturk, social media, and face-to-face behavioral testing. *Computers in human behavior*, *29*(6), 2156–2160.

Dinneen, J. D., Asadi, B., Frissen, I., Shu, F., & Julien, C.-A. (2018). Improving exploration of topic hierarchies: Comparative testing of simplified library of congress subject heading structures, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176385

Edwards, A., & Kelly, D. (2017). Engaged or frustrated? disambiguating emotional state in search, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080818

Garcia-Gathright, J., St. Thomas, B., Hosey, C., Nazari, Z., & Diaz, F. (2018). Understanding and evaluating user satisfaction with music discovery, In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, Ann Arbor, MI, USA, Association for Computing Machinery. https://doi.org/10.1145/3209978.3210049

Ghosh, S., Rath, M., & Shah, C. (2018). Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176386

Harvey, M., & Pointon, M. (2017). Searching on the go: The effects of fragmented attention on mobile web search tasks, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080770

He, J., Bron, M., de Vries, A., Azzopardi, L., & de Rijke, M. (2015). Untangling result list refinement and ranking quality: A framework for evaluation and prediction, In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, Association for Computing Machinery. https://doi.org/10.1145/2766462.2767740

He, J., & Yilmaz, E. (2017). User behaviour and task characteristics: A field study of daily information behaviour, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020188

Hienert, D., Mitsui, M., Mayr, P., Shah, C., & Belkin, N. J. (2018). The role of the task topic in web search of different task types, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176382

Hoeber, O., Sarkar, A., Vacariu, A., Whitney, M., Gaikwad, M., & Kaur, G. (2017). Evaluating the value of lensing wikipedia during the information seeking process, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020178

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, *14*(3), 399–425.

Huang, J., Hu, W., Li, H., & Qu, Y. (2018). Automated comparative table generation for facilitating human intervention in multi-entity resolution, In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, Ann Arbor, MI, USA, Association for Computing Machinery. https://doi.org/10.1145/3209978.3210021

Jiang, J., He, D., & Allan, J. (2017). Comparing in situ and multidimensional relevance judgments, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080840

Jiang, J., He, D., Kelly, D., & Allan, J. (2017). Understanding ephemeral state of relevance, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020176

Ke, W., Sugimoto, C. R., & Mostafa, J. (2009). Dynamicity vs. effectiveness: Studying online clustering for scatter/gather, In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, MA, USA, Association for Computing Machinery. https://doi.org/10.1145/1571941.1571947

Kelly, D., Arguello, J., Edwards, A., & Wu, W.-c. (2015). 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*, Northampton, Massachusetts, USA, Association for Computing Machinery. https://doi.org/10.1145/2808194.2809465

Kim, H. S., & Hodgins, D. C. (2017). Reliability and validity of data obtained from alcohol, cannabis, and gambling populations on amazon's mechanical turk. *Psychology of addictive behaviors*, *31*(1), 85.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H.-J. (2017). What snippet size is needed in mobile web search?, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020173

Kim, J. Y., Craswell, N., Dumais, S., Radlinski, F., & Liu, F. (2017). Understanding and modeling success in email search, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080837

Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016). Predicting user satisfaction with intelligent assistants, In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, Association for Computing Machinery. https://doi.org/10.1145/2911451.2911521

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk, In *Proceedings of the sigchi conference on human factors in computing systems*, Florence, Italy, Association for Computing Machinery. https://doi.org/10.1145/1357054.1357127

Klouche, K., Ruotsalo, T., Micallef, L., Andolina, S., & Jacucci, G. (2017). Visual re-ranking for multi-aspect information retrieval, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020174

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, Association for Computing Machinery. https://doi.org/10.1145/2470654.2470684

Kotzyba, M., Gossen, T., Schwerdt, J., & Nürnberger, A. (2017). Exploration or fact-finding: Inferring user's search activity just in time, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020180

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing Management*, *44*(6), 1822–1837. https://doi.org/https://doi.org/10.1016/j.ipm.2008.07.005

Ling, C., Steichen, B., & Choulos, A. G. (2018). A comparative user study of interactive multilingual search interfaces, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176383

Liu, X., Jiang, Z., & Gao, L. (2015). Scientific information understanding via open educational resources (oer), In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, Association for Computing Machinery. https://doi.org/10.1145/2766462.2767750

Liu, Y., Chen, Y., Tang, J., Sun, J., Zhang, M., Ma, S., & Zhu, X. (2015). Different users, different opinions: Predicting search satisfaction with mouse movement information, In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, Association for Computing Machinery. https://doi.org/10.1145/2766462.2767721

Liu, Y., Liu, Z., Zhou, K., Wang, M., Luan, H., Wang, C., Zhang, M., & Ma, S. (2016). Predicting search user examination with visual saliency, In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, Association for Computing Machinery. https://doi.org/10.1145/2911451.2911517

Liu, Z., Liu, Y., Zhou, K., Zhang, M., & Ma, S. (2015). Influence of vertical result in web search examination, In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, Association for Computing Machinery. https://doi.org/10.1145/2766462.2767714

Lu, H., Zhang, M., & Ma, S. (2018). Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading, In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, Ann Arbor, MI, USA, Association for Computing Machinery. https://doi.org/10.1145/3209978.3210007

Luo, C., Li, X., Liu, Y., Sakai, T., Zhang, F., Zhang, M., & Ma, S. (2017). Investigating users' time perception during web search, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020184

Luo, C., Liu, Y., Sakai, T., Zhang, F., Zhang, M., & Ma, S. (2017). Evaluating mobile search with height-biased gain, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080795

Maxwell, D., Azzopardi, L., & Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080824

Meier, F., & Elsweiler, D. (2016). Going back in time: An investigation of social media re-finding, In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, Association for Computing Machinery. https://doi.org/10.1145/2911451.2911524

Moshfeghi, Y., Triantafillou, P., & Pollick, F. E. (2016). Understanding information need: An FMRI study, In *Proceedings of the 39th International ACM SIGIR Conference*

on Research and Development in Information Retrieval, Pisa, Italy, Association for Computing Machinery. https://doi.org/10.1145/2911451.2911534

Ong, K., Järvelin, K., Sanderson, M., & Scholer, F. (2017). Using information scent to understand mobile and desktop web search behavior, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080817

Rouse, S. V. (2015). A reliability analysis of mechanical turk data. *Computers in Human Behavior*, *43*, 304–307.

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap, In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*.

Sakai, T. (2016). Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015, In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, Association for Computing Machinery. https://doi.org/10.1145/2911451.2911492

Salminen, J., Jansen, B. J., An, J., Jung, S.-G., Nielsen, L., & Kwak, H. (2018). Fixation and confusion: Investigating eye-tracking participants' exposure to information in personas, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176391

Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability, In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*.

Sarrafzadeh, B., & Lank, E. (2017). Improving exploratory search experience through hierarchical knowledge graphs, In *Proceedings of the 40th International ACM SIGIR Confer-*

ence on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080829

Singh, J., Zerr, S., & Siersdorfer, S. (2017). Structure-aware visualization of text corpora, In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, Oslo, Norway, Association for Computing Machinery. https://doi.org/10.1145/3020165.3020182

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in cognitive sciences*, *21*(10), 736–748.

Su, Y., Hassan Awadallah, A., Wang, M., & White, R. W. (2018). Natural language interfaces with fine-grained user interaction: A case study on web apis, In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, Ann Arbor, MI, USA, Association for Computing Machinery. https://doi.org/10.1145/3209978.3210013

Thomas, K. A., & Clifford, S. (2017). Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197.

Trippas, J. R., Spina, D., Cavedon, L., Joho, H., & Sanderson, M. (2018). Informing the design of spoken conversational search: Perspective paper, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176387

Turpin, A., Scholer, F., Mizzaro, S., & Maddalena, E. (2015). The benefits of magnitude estimation relevance assessments for information retrieval evaluation, In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, Association for Computing Machinery. https://doi.org/10.1145/2766462.2767760

Umemoto, K., Yamamoto, T., & Tanaka, K. (2016). Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search, In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, Association for Computing Machinery. https://doi.org/10.1145/2911451.2911546

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error, In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*.

Wang, J., & Komlodi, A. (2018). Switching languages in online searching: A qualitative study of web users' code-switching search behaviors, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176396

Wang, Y., Sarkar, S., & Shah, C. (2018). Juggling with information sources, task type, and information quality, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176390

Wu, W.-C., Kelly, D., Edwards, A., & Arguello, J. (2012). Grannies, tanning beds, tattoos and nascar: Evaluation of search tasks with varying levels of cognitive complexity, In *Proceedings of the 4th Information Interaction in Context Symposium*, Nijmegen, The Netherlands, Association for Computing Machinery. https://doi.org/10.1145/2362724.2362768

Xie, X., Liu, Y., Wang, X., Wang, M., Wu, Z., Wu, Y., Zhang, M., & Ma, S. (2017). Investigating examination behavior of image search users, In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3077136.3080799

Zhang, F., Zhou, K., Shao, Y., Luo, C., Zhang, M., & Ma, S. (2018). How well do offline and online evaluation metrics measure user satisfaction in web image search?, In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, Ann Arbor, MI, USA, Association for Computing Machinery. https://doi.org/10.1145/3209978.3210059

Zhang, H., Abualsaud, M., & Smucker, M. D. (2018). A study of immediate requery behavior in search, In *Proceedings of the 2018 Conference on Human Information Interaction Retrieval*, New Brunswick, NJ, USA, Association for Computing Machinery. https://doi.org/10.1145/3176349.3176400

Appendices

Appendix A

Appendix of Screenshot of Interfaces

This section showed the screenshots of interfaces in our systems, which include 1)collecting participant's personal information; 2) presenting the search task for participants; 3)a pre-task questionnaire page; 4) a post-task questionnaire page; and 5) a confirmation page.

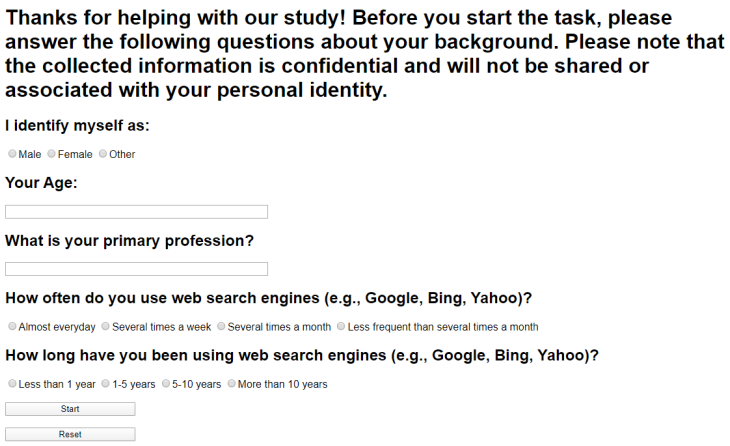


Figure A.1: Questionnaire Page for Participant's Personal Information

Please read the following problem description. Imagine you are the person in the described situation. Please use this web search engine to find information related to the described problem.

Your Problem

After the NASCAR season opened this year, your niece became really interested in soapbox derby racing. Since her parents are both really busy, you've agreed to help her build a car so that she can enter a local race. The first step is to figure out how to build a car. Identify some basic designs that you might use and create a basic plan for constructing the car.

Next

Figure A.2: Task Preview Page

Before you get started, please answer the following questions about the described problem.

Your Problem:

After the NASCAR season opened this year, your niece became really interested in soapbox derby racing. Since her parents are both really busy, you've agreed to help her build a car so that she can enter a local race. The first step is to figure out how to build a car. Identify some basic designs that you might use and create a basic plan for constructing the car.

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

Not At All | Slightly | Somewhat | Moderately | Very

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

Never | 1-2 times | 3-4 times | 5+ times

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

Nothing | Little | Some | Great deal

How defined is this task in terms of the types of information needed to complete it?

Not At All | Slightly | Somewhat | Moderately | Very

How defined is this task in terms of the steps required to complete it?

Not At All | Slightly | Somewhat | Moderately | Very

How defined is this task in terms of its expected solution?

Not At All | Slightly | Somewhat | Moderately | Very

How difficult do you think it will be to *search* for information for this task using a search engine?

Not At All | Slightly | Somewhat | Moderately | Very

How difficult do you think it will be to *understand* the information the search engine finds?

Not At All | Slightly | Somewhat | Moderately | Very

Figure A.3: Pre-Task Questionnaire Page Part I

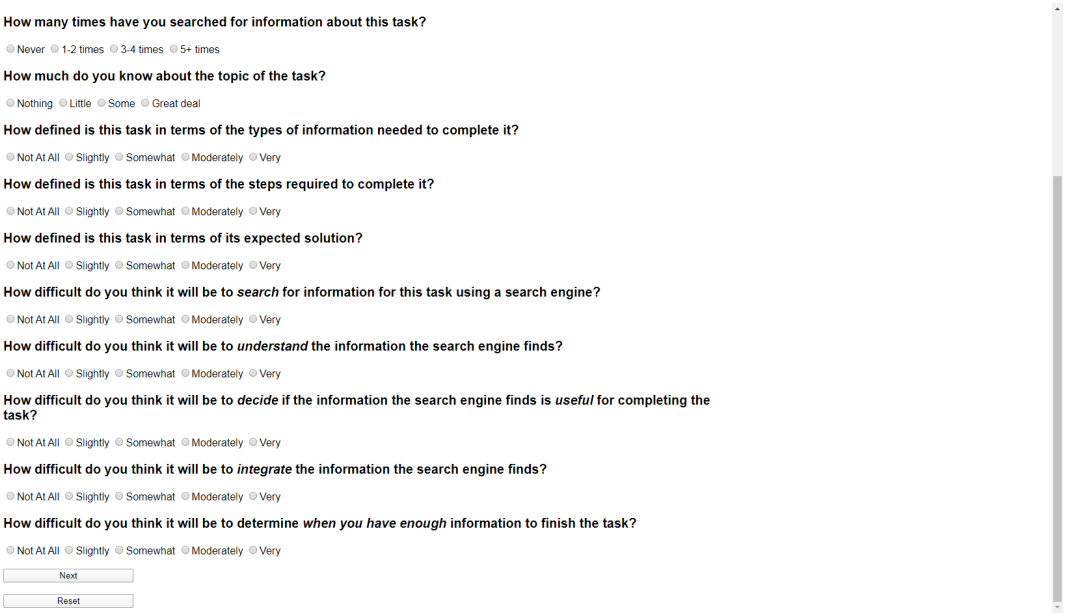


Figure A.4: Pre-Task Questionnaire Page Part II

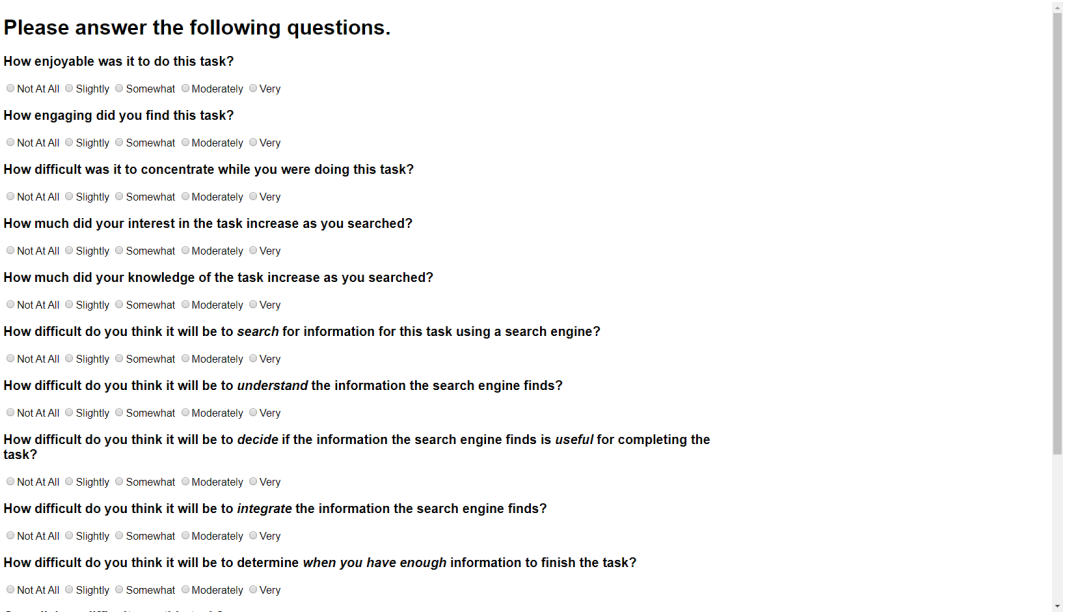


Figure A.5: Post-Task Questionnaire Page Part I

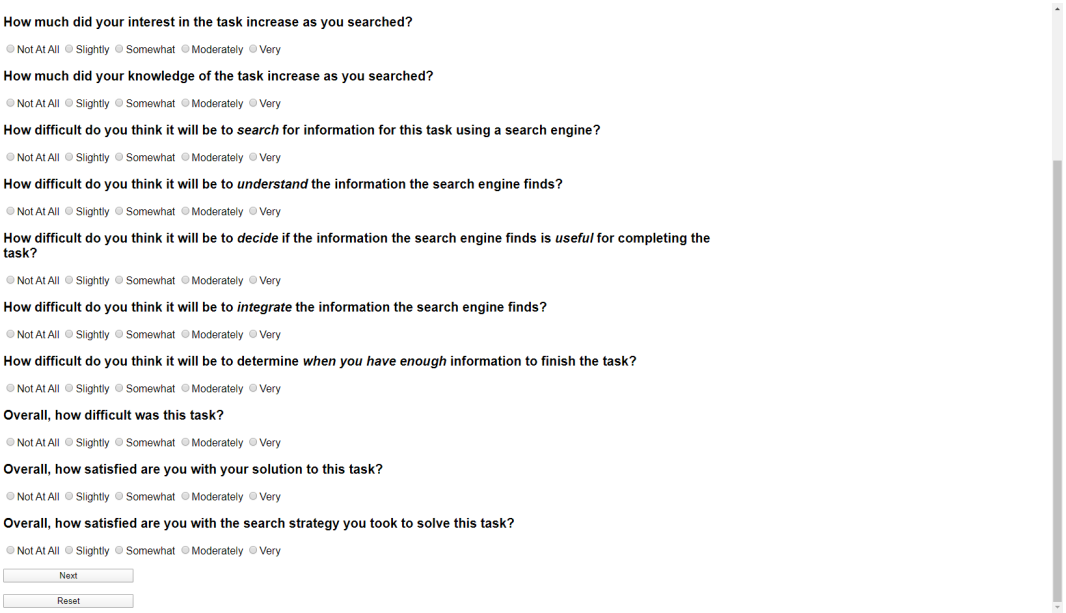


Figure A.6: Post-Task Questionnaire Page Part II

Please enter the following Confirmation Code in the Amazon MTurk page to receive your payment. If the Confirmation Code does not appear, please refresh the page.

Confirmation Code: 5069998

Figure A.7: Confirmation Page