Better Understanding Procedural Search Tasks: Perceptions, Behaviors, and Challenges

BOGEUM CHOI, SARAH CASTEEL, JAIME ARGUELLO, ROBERT CAPRA,
School of Information and Library Science University of North Carolina at Chapel Hill

People often search for information to acquire procedural knowledge–"how to" knowledge about step-by-step procedures, methods, algorithms, techniques, heuristics, and skills. A procedural search task might involve implementing a solution to a problem, evaluating different approaches to a problem, and brainstorming on the types of problems that can be solved with a specific resource. We report on a study ($N = 36$) that aimed to better understand how people search for procedural knowledge. Much research has investigated how search task characteristics impact people's perceptions and behaviors. Along these lines, we manipulated procedural search tasks along two orthogonal dimensions: product and goal. The product dimension relates to the main outcome of the task and the goal dimension relates to task's success criteria. We manipulated tasks across three product categories and two goal categories. The study investigated four research questions. First, we examined the effects of the product and goal on participants (RQ1) pre-task perceptions, (RQ2) post-task perceptions, and (RQ3) search behaviors. Second, regardless of the task product and goal, by analyzing participants' think-aloud comments and screen activities we closely examined how people search for procedural knowledge. Specifically, we report on (RQ4) important relevance criteria, types of information sought, and challenges.

Additional Key Words and Phrases: Procedural Search, Qualitative Research, User Studies

ACM Reference Format:
Bogeum Choi, Sarah Casteel, Jaime Arguello, Robert Capra. 2023. Better Understanding Procedural Search Tasks: Perceptions, Behaviors, and Challenges. 1, 1 (October 2023), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

An important question in interactive information retrieval (IIR) is: How do task characteristics influence searchers' perceptions, behaviors, and challenges faced? In this research, we investigate *procedural search tasks*. A procedural search task is one that involves acquiring *procedural knowledge*, which is defined as "how to" knowledge and includes knowledge about step-by-step procedures, algorithms, methods, techniques, heuristics, and skills [2]. In this respect, a procedural search task might involve gathering information to solve a problem, evaluating different approaches to a higher-level task, or understanding the types of problems that can be solved with a specific resource.

Prior research has investigated procedural search tasks from different perspectives. First, studies have investigated how frequently people search for procedural knowledge online [3, 18, 47, 48]. Studies suggest that about 2-3% of all web search queries have "how to" intent. Second, studies have

Authors' addresses: B. Choi, S. Casteel, J. Arguello, and R. Capra. School of Information and Library Science. University of North Carolina at Chapel Hill. Manning Hall, 216 Lenoir Drive, Chapel Hill, North Carolina 27599. Emails: choiboge@unc.edu, scasteel@alumni.unc.edu, jarguello@unc.edu, rcapra@unc.edu.

XXXX-XXXX/2023/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

investigated the *process* through which people acquire procedural knowledge in specific contexts, such as an educational setting [21] or professional setting [13, 22]. Finally, studies have developed novel tools to help people search for procedural knowledge [36, 48, 52].

Our research has two main objectives. First, we aim to better understand how *characteristics* of procedural search tasks may influence searchers' perceptions and behaviors. Second, we aim to closely observe how people search for procedural knowledge online. Specifically, we are interested in better understanding: (1) important criteria used by searchers when judging the relevance of information; (2) the types of information searchers seek; and (3) specific challenges faced by people while searching for procedural knowledge.

We report on a user study ($N = 36$) in which participants completed different types of procedural search tasks. Li and Belkin [29] proposed a taxonomy for classifying search tasks along different orthogonal dimensions. We leveraged this taxonomy to manipulate procedural search tasks along two dimensions: *product* and *goal*. The product dimension relates to the main outcome of the task—What is the searcher trying to produce? Our product manipulation involved three categories: *factual*, *decision*, and *intellectual*. Our factual tasks asked participants to execute a procedure; our decision tasks asked participants to evaluate different approaches to a problem; and our intellectual tasks asked participants to generate new ideas (e.g., brainstorm tasks that could be accomplished with a given resource). The goal dimension relates to whether the goal of the task has measurable success criteria. Our goal manipulation involved two categories: *specific* and *amorphous*. A specific task has a goal that is explicit and objectively measurable. Conversely, an amorphous task has a goal that is less explicit and/or more subjective.

In total, we developed 18 procedural search task that varied along 3 product categories, 2 goal categories, and 3 topical domains (finance, genealogy, and technology). Domain and goal were between-subjects factors. That is, each participant was assigned to one domain-goal combination (e.g., finance-specific). Product was a within-subjects factor. Each participant completed one factual, one decision, and one intellectual task.

Our study investigated four research questions:

- **RQ1:** What is the effect of the task product and goal on participants' pre-task perceptions about the task?
- **RQ2:** What is the effect of the task product and goal on participants' post-task perceptions about the task?
- **RQ3:** What is the effect of the task product and goal on participants' search behaviors?
- **RQ4:** Irrespective of the task, what criteria did participants use to judge relevance, what types of information did participants seek, and what types of challenges did participants face?

To address RQ1-RQ2, participants completed questionnaires before and after each task. To address RQ3, participants interacted with a custom-build web search system that logged all SERP-level interactions (e.g., queries, clicks, mouse, and scroll events). To address RQ4, the study used a think-aloud protocol. Participants' think-aloud comments and screen activities were recorded and analyzed using qualitative techniques to gain insights about relevance criteria, types of information sought, and challenges.

In general, our results for RQ1-RQ3 found that our manipulation of the task product had a stronger effect than our manipulation of the task goal. Our results for RQ4 found that participants used a wide range of criteria when judging relevance, they sought many different types of information, and they experienced many different types of challenges. Based on our results, we propose different tools and interface features to support users with procedural search tasks.

2 RELATED WORK

We build upon and extend prior research on: (1) defining procedural knowledge; (2) understanding the frequency with which users search procedural knowledge online; (3) understanding the *process* through which people acquire procedural knowledge in specific contexts; and (4) developing tools to support procedural search and procedural knowledge acquisition.

2.1 Defining Procedural Knowledge

Prior work has aimed to define procedural knowledge and distinguish it from other types of knowledge. Georgeff and Lansky [24] argue that procedural knowledge involves sequences of actions for achieving particular goals. Procedural knowledge has been distinguished from declarative knowledge as knowing *how* versus knowing *that* [43]. Procedural knowledge has also been distinguished from conceptual knowledge as being knowledge that assists in solving concrete problems rather than knowledge that facilitates the understanding of abstract principles [39].

In the field of education, the Anderson and Krathwohl (A&K) taxonomy [2] was developed to help educators define learning objectives for students. The A&K taxonomy situates learning objectives at the interaction of two orthogonal dimensions: *knowledge type* and *cognitive process*. The taxonomy distinguishes between four knowledge types: factual, conceptual, procedural, and metacognitive knowledge. Procedural knowledge is defined as "how to" knowledge about performing a specific task and includes knowledge about step-by-step procedures, algorithms, techniques, heuristics, methods, and skills.

The A&K taxonomy defines six cognitive processes: remember, understand, apply, analyze, evaluate, and create. The cognitive process dimension is a useful framework for thinking about the complexity of a search task involving a specific type of procedural knowledge. Depending on the searcher's objective, a procedural search task may range from simple to complex: (1) remember—memorizing the steps of procedure X; (2) understand—understanding the steps of X; (3) apply—executing X; (4) analyze—differentiating X from other procedures; (5) evaluate—critiquing X; and (6) create—modifying X to fit a specific situation. In Section 3.2, we argue that our factual, decision, and intellectual tasks can be viewed as apply-, evaluate-, and create-level tasks.

2.2 Procedural Search Online

Understanding procedural search tasks is important because people already use web search engines to acquire procedural knowledge. Völske et al. [47] analyzed about one billion natural language queries (NLQs) issued to Yandex during a one-year period. NLQs accounted for 4% of all queries, and a substantial portion (not specified) were queries of the form "how to [verb]". Common verbs included 'make', 'cook', 'install', 'build', 'calculate' and 'clean'. Interestingly, many 'how to [verb]' queries ended with terms associated with user-specific constraints (e.g., 'at home'). Eickhoff et al. [18] analyzed queries issued to Bing over a one-month period. By analyzing clicks on specific websites (e.g., Wikipedia and eHow), the authors estimated that 3% of all search sessions had *knowledge acquisition intent* involving either declarative or procedural knowledge. For procedural knowledge queries, characteristic n-grams included "how do", "how to", and "can I", which suggests that procedural searches often involve uncertainty about *feasibility*. Bailey and Jiang [3] analyzed three-month's worth of Bing search sessions and developed a taxonomy of web search tasks, which included 26 high-level categories. "Learn how to perform a task" was the 12th most frequent category, accounting for 2% of all sessions. Additionally, procedural search sessions were the 3rd longest (13 queries on average). Weber et al. [48] analyzed a sample of 3,000 queries issued to Yahoo! and found that 2% of all queries had "how-to" intent.

Collectively, these studies suggest that procedural search tasks account for 2-3% of all web search queries and that procedural searches are complex (i.e., involve lots of queries). While 2-3% may seem small, it represents a substantial number of web search queries.

2.3 Procedural Knowledge Acquisition

Studies have also investigated how people acquire procedural knowledge in specific contexts. Ertl [21] studied the effects of prior knowledge and collaborative (vs. individual) learning on procedural knowledge acquisition. Results found two important trends. First, prior knowledge of concepts and definitions resulted in better learning outcomes, which suggests that background conceptual knowledge is an important prerequisite for gaining procedural knowledge. Second, collaborative (vs. individual) learning resulted in better learning outcomes, which suggests that procedural knowledge learners can benefit from being exposed to other people's thought processes and perspectives.

Freund et al. [22] studied the information-seeking practices of software engineers at work. Results found several important trends. First, participants frequently engaged in search tasks involving procedural knowledge (e.g., learning how to do something, troubleshooting a problem, and finding the right tool/resource to solve a problem). Second, participants reported experiencing challenges related to: (1) information overload; (2) too many independent sources of information; (3) inaccurate or obsolete information; and (4) lack of system support for narrowing the search results. Finally, for highly complex tasks, participants preferred information from people with firsthand experience (e.g., asking a colleague or searching on a forum). Byström and Järvelin [6] also found that highly complex tasks (procedural or otherwise) involve greater use of people as information sources.

Pardi et al. [34] studied search behaviors during different types of procedural search tasks: a cognitive task and a physical task. Results found that participants favored visual content, especially during the physical task. Frummet et al. [23] conducted an in-situ study of people's information needs while cooking. The authors developed a hierarchical taxonomy of information needs that included high-level categories related to specific steps, techniques, and recipes with specific inclusion and exclusion criteria. Urgo et al. [45] conducted a study that compared participants' behaviors during search tasks involving procedural versus factual or conceptual knowledge. During procedural search tasks, participants were more likely to engage in creative processes (e.g., modifying and combining procedures to fit their individual preferences). Choi et al. [13] conducted a survey of U.S. intelligence analysts who routinely use an internal system to search for procedural knowledge. Based on their findings, the authors proposed novel features to alleviate the challenges reported by participants.

2.4 Tools to Support Procedural Search

Prior work has also developed tools to support procedural search. Pothirattanachaikul et al. [36] leveraged community Q&A data to develop an algorithm to predict alternative ways to solve a problem. For instance, "taking a sleeping pill", "doing evening exercises", and "drinking chamomile tea" are alternative procedures for "ways to improve sleep". Weber et al. [48] developed algorithms to automatically identify "how-to" queries and propose relevant tips mined from Yahoo! Answers. Tips were defined as nuggets of advice that are short, self-contained, actionable, and not obvious. Yang and Nyberg [52] developed algorithms to suggest queries about the steps of an input procedural query.

In recent years, natural language processing (NLP) researchers have proposed a wide range of methods for automatically populating procedural knowledge bases from unstructured or semi-structured documents [1, 15, 33, 35, 40, 53]. Most of these approaches harness data from websites such as WikiHow, which focus on procedural knowledge, have a consistent document layout,

and include rich metadata. Within these knowledge bases, procedures are typically modeled as sequences of steps, which involve inputs, actions, and outputs. Some of these procedural knowledge bases also model relations between procedures, such as one procedure being a sub-step in another or multiple procedures being alternative ways to perform the same task [15, 35]. Research in this area argues that procedural knowledge bases can support different types of applications, including query-suggestion, question-answering, and conversational search.

3 METHODS

To investigate RQ1-RQ4, we conducted a study with 36 participants ($M = 5$, $F = 31$).[1][2] Participants were recruited via an opt-in mailing list of employees at our university. Participants were 21-55 years old and included 30 non-student employees and 6 student employees. The study was conducted during the COVID-19 pandemic and was therefore conducted remotely over the Zoom videoconferencing platform. The study was approved by our university's Institutional Review Board (IRB).

During the study, participants completed three procedural search tasks. Our objective in RQ1-RQ3 was to investigate participants' perceptions and behaviors during different types of procedural search tasks. To this end, we manipulated procedural search tasks along two dimensions: product and goal (Section 3.2). Additionally, we wanted to expose participants to different topical domains. In total, we developed 18 search tasks varying along 3 product categories (factual, decision, intellectual), 2 goal categories (specific, amorphous), and 3 topical domains (finance, technology, genealogy). We chose the domains of finance, technology, and genealogy because they involve procedures such as computing evaluation metrics (finance), connecting hardware (technology), and combining information sources to answer specific questions (genealogy). Topical domain and goal were between-subjects factors. That is, each participant was exposed to only one topical domain (e.g., finance) and one goal category (e.g., specific). Conversely, product was a within-subjects factor. That is, each participant experienced all three product categories (i.e., factual, decision, and intellectual). Six (out of 36) participants were assigned to each domain-goal combination (e.g., finance-specific). Three product categories can be ordered six different ways. Therefore, the six participants assigned to the each domain-goal combination were exposed to our three product categories in a different order.

To investigate RQ4, we employed a combination of observations and think-aloud protocols. Participants were instructed to verbalize their thoughts while gathering information, and we recorded, transcribed, and analyzed their think-aloud comments alongside their screen activities using qualitative techniques (Section 3.6).

In the study, product was a within-subjects factor, while domain and goal were between-subjects factors. This was done for several reasons. First, to limit the study session to one hour, we did not want participants completing more than three search tasks. Therefore, product seemed like an appropriate between-subjects factor. Each participant was exposed to all three product categories. Second, we thought that having participants switch between multiple domains would be too cognitively demanding. Therefore, participants completed three tasks within the same domain. Other studies that have manipulated tasks across multiple domains have also treated domain as a between-subjects factor (possibly for the same reason) [9, 26, 45, 46]. To counteract any learning

[1]Participants completed a short demographics questionnaire at the beginning of the study session. Participants indicated their gender by entering it in a textbox (i.e., we did not provide a set of options). Participants could leave this item blank. All participants self-identified as either male or female.

[2]We had many more female than male participants. We discuss possible implications of this imbalance in Section 7.

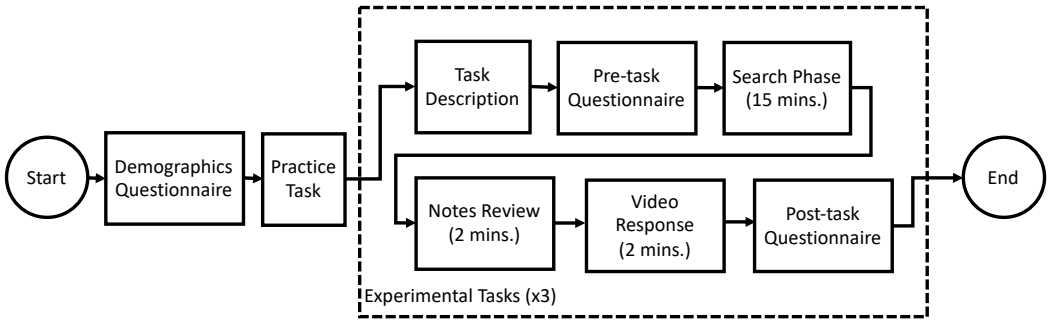


Fig. 1. Study Protocol

effects, participants were exposed to product categories in different order.[3] Finally, we decided to treat goal as a between-subjects factor to mitigate the risk of a "spillover effect" from one search task to another. For example, we did not want the uncertainty of an amorphous task (e.g., Is the task feasible?) to "spillover" onto a subsequent specific task.

3.1 Study Protocol

The study protocol is illustrated in Figure 1. At the start of the study session, participants completed an informed consent form and a demographics questionnaire. Then, participants completed an example search task to practice thinking aloud. Next, participants completed three experimental tasks that followed the same sequence of steps.

First, participants were asked to read the next search task description aloud. Next, participants completed a pre-task questionnaire about their perceptions of the task and expectations (Section 3.3). Then, participants completed the search phase of the task. Participants were given 15 minutes to complete this phase. During the search phase, participants were asked to use a custom-built web search engine to gather information and a Google Doc to take notes. The web search engine was developed using the Bing Web Search API. The search interface included a button to display the task description in case participants wanted to refresh their memory. As previously mentioned, participants were instructed to think aloud while searching. To help participants feel more comfortable thinking aloud, the study moderator muted themselves and turned off their camera during the search phase. The moderator only unmuted themselves to prompt participants to remember thinking aloud. After the search phase, participants completed the video response phase of the task. Participants were given two minutes to review their notes and two minutes to produce a video response for the task. During the video response, participants were instructed to explain to the moderator what they found, what they learned, and how far they got on the task. The video response phase was meant to encourage participants to take the task seriously and avoid satisficing. Finally, after the video response phase, participants completed a post-task questionnaire about their experience during the task (Section 3.4). With their consent, the entire study session was recorded. Participants shared their screen and had their audio and video turned on. To address RQ4, participants' recorded think-aloud comments and screen activities were analyzed using qualitative techniques (Section 3.6). Participants received a US$40 Amazon gift card for participating in the study.

[3]Three product categories can be ordered six different ways. Of the 12 participants assigned to each domain, 2 were exposed to product categories in the same order.

3.2 Task Manipulations

Li and Belkin [29] developed a classification scheme to characterize search tasks along different dimensions. In this research, we manipulated procedural search tasks along two dimensions from Li and Belkin [29]: product and goal. The *product* dimension relates to the main outcome of the task—What is the searcher trying to produce as the main outcome of the task? For example, is the searcher trying to implement a solution to a problem, decide between alternative approaches to a problem, or determine the usefulness of a specific resource for solving problems? The *goal* dimension relates to the extent to which the task outcome has measurable success criteria—How easy or difficult will it be for the searcher to measure success? For example, a search task might have success criteria that are highly concrete and objective (easy to measure) or highly abstract and subjective (difficult to measure).

Our product manipulation involved three categories from Li and Belkin [29]: factual, decision, and intellectual tasks. A *factual* task involves solving a problem by gathering facts, data, or similar types of (mostly objective) information. A *decision* task involves considering alternatives and making a decision. An *intellectual* task involves generating new ideas. In the context of procedural search, a factual task might involve figuring out how to execute a procedure; a decision task might involve comparing alternative procedures to solve a problem and selecting the best one based on personal preferences or constraints; and an intellectual task might involve brainstorming different types of problems that one could solve with a given procedure or resource.

Our goal manipulation involved two categories from Li and Belkin [29]: specific and amorphous. A *specific* task has a concrete objective. That is, it has success criteria that are relatively easy to measure objectively. An *amorphous* task has an objective that is more abstract and subjective. That is, it has success criteria that are more difficult to measure. To illustrate, let us consider two *decision* tasks, one specific and one amorphous. A specific decision task might involve comparing different procedures and selecting an appropriate one based on a single, objective criterion (e.g., the procedure can be executed using available equipment and materials). Conversely, a more amorphous decision task might involve comparing different procedures and the selecting the *best* one based on multiple, subjective criteria (e.g., effort, risk, and durability of the solution). The second decision task is more amorphous because it involves greater uncertainty. The searcher will need to weigh different criteria based on personal preferences and each of the criteria is highly subjective (i.e., there may be differences of opinion).

In total, we developed 18 procedural search tasks that varied across 3 topical domains (finance, genealogy, technology); 3 product categories (factual, intellectual, decision); and 2 goal categories (specific, amorphous).[4] The full version of our tasks is available online.[5] To conserve space, Table 1 only shows our six specific tasks across the domains of finance, genealogy, and technology. To better explain our manipulation of the task product and goal, Tables 2-4 show the template associated with each task across domains.

We chose to develop tasks across the domains of finance, genealogy, technology for several reasons. First, we wanted the study to involve *multiple* domains to help ensure that our results generalize across domains. Many IIR studies have involved task manipulations across multiple domains. For example, Kelly et al. [26] manipulated the cognitive complexity of tasks across the domains of health, commerce, entertainment, and science & technology. Second, we wanted a diverse set of domains to match our diverse pool of participants—student and non-student university

[4]Tasks were developed by three of the authors. First, we defined our three product categories and two goal categories. Then, each author was assigned to a domain and developed all six tasks for that domain. We met several times to discuss our tasks and provide each other feedback. Finally, several tasks were pilot tested with student volunteers.
[5]https://ils.unc.edu/~jarguell/pk_lab_study/pk_lab_study_tasks.pdf

Table 1. Specific tasks across the domains of finance, genealogy, and technology (slightly shortened).

	Factual	Decision	Intellectual
Finance	You want to invest in the stock market and want to compare different publicly traded companies based on their price-to-earnings (P/E) ratio. How do you compute P/E ratio? What information do you need and where is this information typically found?	You want to invest in the stock market. Publicly traded companies can be evaluated based on different metrics. In your opinion, what are the five best metrics? Focus on metrics that are easy to compute using information freely available online. For each metric, explain: (1) how to compute it, (2) what it tells you, (3) what information you need to compute it, and (4) how the required information can be easily found online.	You want to invest in the stock market and are interested in stocks that pay dividends. In your opinion, what are the pros and cons of "average dividend yield" as a metric to evaluate stocks that pay dividends?
Genealogy	Your mother suspects that a distant relative obtained land in Florida from the U.S. Federal Government. How can you find a document confirming which land was transferred to a person named A. F. Williams from the U.S. Federal Government?	One of your relatives named Wladislaw Kowalski arrived in the U.S. from Poland in the 1890s. In your opinion, what are the best genealogy research tools that would help you learn about this relative? Focus on tools that are affordable. For each tool, explain: (1) what it can tell you and (2) why you think it is affordable and complete.	You are helping your parents with genealogy research. You recently discovered a set of documents that might be useful: (1) a land grant from North Carolina Country to one of your relatives from 1792, (2) a survey warrant from 1847, (3) a relative's will from 1858, and (4) letters between one relative and their children from 1895-1901. How might these documents help someone learn about buying and selling land in the 1800s? Explain your reasoning.
Technology	You want to use a Pluggable UD-ULTCDL docking station with a Lenovo X1 Tablet (2nd Gen) computer? You would like the docking station to support both power and a display connection via the USB-C connection. Is it possible to use this docking station with this computer? If so, explain how you know this. If not, explain why not.	A friend gave you an old Dell 2007FP monitor to use as an extra screen with your Microsoft Surface Pro 3 computer. What are different options for connecting your computer to this monitor. Which option would give you the best picture quality? Why?	You are helping a friend with refurbishing old laptop computers, installing Linux on them, and donating them school districts that need them. He recently obtained a large batch of IBM/Lenovo ThinkPads from the 2008 to 2012. You recently discovered the following website: www.thinkwiki.org. How can the information on this website help your friend refurbish these computers?

employees 21-55 years of age. Third, we believe that we chose domains where procedural tasks are fairly common. Finance-related tasks often involve performing calculations, genealogy-related tasks often involve using tools to answer questions, and technology-related tasks often involve solving problems.

Product Manipulation: As described in Tables 2-4, our factual tasks involved: (1) executing a procedure (finance); (2) finding a procedure to solve a problem (genealogy); and (3) understanding the compatibility between two tools (technology). Our decision tasks involved: (1-2) selecting the best procedure to solve a problem (finance, genealogy) and (3) choosing the best way to combine tools to solve a problem (technology). Our intellectual tasks involved: (1) brainstorming the benefits

Table 2. Task templates for finance domain.

	Factual	Decision	Intellectual
Specific	How do you execute procedure X?	What are the top-N procedures to accomplish goal X based on success criterion Y?	What are the benefits and drawbacks of procedure X?
Amorphous	How do you execute procedure X in scenario Y? Is it possible? If so, how? If not, why not?	What are the top-N procedures to accomplish goal X based on the success criteria that are most important to you?	What are the benefits and drawbacks of procedure X? Are there ways to alleviate the drawbacks? If so, how?
Specific vs. Amorphous Justification	The amorphous task required participants to determine whether X is executable in scenario Y.	The amorphous task required participants to determine which success criteria are most important to them.	The amorphous task required participants to determine whether there are ways to alleviate the drawbacks of X.

Table 3. Task templates for genealogy domain.

	Factual	Decision	Intellectual
Specific	Find a procedure to solve problem X.	What are the top-N procedures to accomplish goal X based on success criterion Y?	You have resources A, B, C, D. Given these resources, can you solve problem X?
Amorphous	Find a procedure to solve problem X under constraint Y. Is it possible? If so, how? If not, why not?	What are the top-N procedures to accomplish goal X based on the success criteria that are most important to you?	If you have resources A and B. What types of problems can you solve by combining these resources?
Specific vs. Amorphous Justification	The amorphous task required participants to determine whether X is can be solved under constraint Y.	The amorphous task required participants to determine which success criteria are most important to them.	The amorphous task is more open-ended, requiring participants to enumerate all the types of problems that can be solved by combining resources.

and drawbacks of a procedure (finance) and (2-3) brainstorming ways in which resources can be used/combined to solve problems (genealogy, technology).

Goal Manipulation: Our goal manipulation was more subtle. It involved adding more complexity and uncertainty to the specific task version within the same domain and product category. In Tables 2-4, the last row provides explanations for why the amorphous task version involves greater uncertainty than the specific task version within the *same* column. Let us consider three examples. First, as described in Table 2 (finance), the factual-amorphous task required participants to execute a procedure *in a scenario where it might not be possible*. Second, as described in Table 3 (genealogy), the decision-amorphous task required participants to determine which success criteria are most important to them. Finally, as described in Table 4 (technology), the intellectual-amorphous task

Table 4. Task templates for technology domain.

	Factual	Decision	Intellectual
Specific	Can tool X be combined with tool Y to achieve goal Z?	How would you best combine tool X with tool Y to achieve goal Z?	How can resource X help you achieve goal Y?
Amorphous	Can tool X be combined with tool Y to achieve **any** of the goals that tool X was designed to achieve?	How would you best combine tool X with tool Y? Why is this the best choice?	How can resource X help you achieve goal Y1, Y2, and Y3.
Specific vs. Amorphous Justification	The amorphous task requires participants to determine which goals can be achieved by combining X with Y.	The amorphous task requires participants to determine which success criteria are most important to them.	The amorphous task requires participants to determine how to use X to achieve **multiple** goals.

required participants to enumerate *all the different problems* that could be solved by combining different resources.

Justification of Product and Goal as Important Dimensions: Based on Li and Belkin's classification scheme [29], search tasks can vary along many different dimensions. We decided to focus on product and goal for three main reasons.

First, past studies have also manipulated search tasks along these two dimensions [16, 30, 31], which suggests that they are important dimensions that tend to impact needs and behaviors. However, these studies did not focus on procedural search tasks.

Second, prior research suggests that these are important dimensions along which *procedural search tasks* can vary. Choi et al., [12] conducted on a survey ($N = 128$) that aimed to understand the characteristics of procedural search tasks conducted "in the wild". Participants were asked to recall a recent procedural search task and were asked questions about the task itself (e.g., goals,) their unique situation (e.g., constraints), and the search process (e.g., relevance criteria). Results suggest that real-world procedural search tasks often vary along both dimensions. In terms of product, some tasks simply required executing a procedure, such as fixing a problem or performing an upgrade (i.e., factual). For some tasks, participants often reported evaluating different alternative based on their unique preferences and constraints (i.e., decision). Finally, some tasks involved generating *new* ideas, such as doing a creative project (i.e., intellectual). In terms of goal, some tasks had a specific end-point that can be measured objectively (i.e., specific) and other tasks were more open-ended and/or had success criteria that were inherently subjective (i.e., amorphous).

Finally, both task dimensions are related to other dimensions studied in prior work. The product dimension is related to the cognitive complexity dimension from A&K's taxonomy [2], described in Section 2.1. Factual tasks can be viewed as apply-level tasks, decision tasks can be viewed as evaluate-level tasks, and intellectual tasks can be viewed as create-level tasks. Several studies have leveraged A&K's cognitive complexity dimension to manipulate search tasks and study their effects [9, 14, 26, 45, 46]. Similarly, the goal dimension is related to the concept of *a priori* determinability—the level of uncertainty about aspects of the task, such its requirements, processes involved, and the form of the solution [6]. Our amorphous tasks were designed to involve greater uncertainty (e.g., Is the task feasible?).

Table 5. PCA component loadings for pre-task questionnaire items about the task requiring specific cognitive activities and types of information.

	Background	SubjectiveDecision	StepByStep
definitions	0.88	0.02	0.10
background	0.80	0.12	-0.19
learn domain	0.63	0.27	0.42
facts	0.59	0.16	0.38
experiential	0.02	0.89	0.03
opinions	0.22	0.86	-0.18
decide alternatives	0.14	0.74	0.24
step-by-step	0.06	-0.01	0.89

3.3 Pre-task Questionnaire

To investigate RQ1 (pre-task perceptions), participants completed a 25-item questionnaire before each task. For all items, participants responded to agreement statements on a 7-point scale ranging from "strongly disagree" to "strongly agree". The pre-task questionnaire asked about perceptions of: (1) prior knowledge and search experience (2 items); (2) expected interest and enjoyment (2 items); (3) expected difficulty (5 items); (4) *a priori* determinability (4 items)[6]; and (5) perception of the task having a clearly defined objective (4 items). Based on Cronbach's α, all measures had high internal consistency (i.e., prior knowledge = 0.88, interest = 0.94, difficulty = 0.90, determinability = 0.84, clear objective = 0.83). Therefore, we averaged responses to these items to form five distinct measures.

Additionally, the pre-task questionnaire included 8 items that asked about perceptions of the task involving specific cognitive activities and requiring specific types of information. In terms of cognitive activities, participants were asked whether they expected the task to require them to: (1) learn about the task domain; (2) gain factual knowledge; and (3) decide between alternatives. In terms of types of information, participants were asked whether they expected the task to require them to gather: (1) definitions; (2) background information; (3) opinions; (4) experiential (or firsthand) knowledge; and (5) step-by-step instructions. We anticipated these items to measure different underlying perceptions. Therefore, we used principal component analysis (PCA) to reduce these 8 measures into a smaller number of components. A PCA with varimax rotation found a three-component solution explaining 71% of the variance. Table 5 shows the component loadings.

As shown in Table 5, each component is labeled based on our interpretation of the underlying construct being captured. The first component (Background) relates to perceptions of the task requiring prerequisite domain knowledge. The second component (SubjectiveDecision) relates to perceptions of the task requiring subjective perspectives to aid in decision making. Finally, the third component (StepByStep) relates to the task requiring step-by-step instructions. Given this PCA output, we decided to average responses to form these three underlying perceptions. As described in Section 3.4, our post-task questionnaire included 8 analogous questions about participants' *actual* experiences during the task. A PCA found the same three underlying components.

[6] *A priori* determinability relates to the level of uncertainty regarding the task requirements, processes involved, and the form of the solution [6]. Low determinability (i.e., high uncertainty) indicates that these aspects of the task are not known in advance and must be determined *as part of the task*

Table 6. PCA component loadings for post-task questionnaire items about the task requiring specific cognitive activities and types of information.

	Background	SubjectiveDecision	StepByStep
definitions	0.87	0.04	0.03
learn domain	0.86	0.17	0.14
facts	0.76	0.01	0.08
background	0.69	0.36	0.07
experiential	0.01	0.87	-0.02
opinions	0.32	0.80	-0.24
decide alternatives	0.10	0.61	0.47
step-by-step	0.15	-0.10	0.91

3.4 Post-task Questionnaire

To investigate RQ2 (post-task perceptions), participants completed a 20-item questionnaire after each task. Again, across all items, participants responded to agreement statements on a 7-point scale ranging from "strongly disagree" to "strongly agree". The post-task questionnaire asked about perceptions of: (1) learning (1 item); (2) interest and enjoyment (2 items); (3) difficulty (5 items); and (4) the task having a clearly defined objective (4 items).

Based on Cronbach's α, the items for measures 2-4 had high internal consistency (i.e., interest = 0.91, difficulty = 0.89, clear objective = 0.81). Therefore, we average responses to these items to form three distinct measures. The item for "learning" was analyzed individually.

Additionally, as in the pre-task questionnaire, the post-task questionnaire included 8 items aimed at capturing participants' perceptions of the task involving specific cognitive activities and requiring specific types of information. Again, a PCA with varimax rotation found a three-component solution explaining 72% of the variance. Table 6 shows the component loadings. Interestingly, the components in Table 6 (post-task perceptions) are analogous to those in Table 5 (pre-task perceptions). Therefore, we adopt the same component labels and definitions from Section 3.3.

3.5 Search Behaviors

To investigate RQ3 (search behaviors), we logged participants' interactions with the search system and computed the following behavioral measures:

(1) **Queries:** number of queries issued.
(2) **Abandoned Queries:** number of queries without a click.
(3) **Clicks:** number of results clicked.
(4) **Avg. Click Rank:** average rank of results clicked.
(5) **Scroll Distance:** total scroll distance in the session, measured in SERP-heights.
(6) **Time to First Click:** time (in secs.) to the first click in the session.
(7) **Avg. Time to First Click:** average time (in secs.) between each query and its first click (if any).
(8) **Pct. Unique Queries:** percentage of queries not issued by another participant.
(9) **Pct. Unique Query Terms:** percentage of query terms not used by another participant.
(10) **Pct. Unique Clicks:** percentage of URLs not clicked by another participant.

Measures 8-10 captured the extent to which participants engaged in search behaviors that diverged from other participants for the same task. Prior work has found that more difficult tasks are associated with more divergent search behaviors [26, 46]. In the context of our study, participants might exhibit more divergent behaviors during certain task types (e.g., tasks with amorphous versus specific goals).

3.6 Qualitative Analysis of Search Sessions

To investigate RQ4, we conducted a qualitative analysis of participants' think-aloud comments and recorded screen activities. Our analysis focused on the following aspects: (RQ4.1) relevance criteria used by participants when determining the usefulness of information encountered during the search; (RQ4.2) the types of information participants engaged with during the search; and (RQ4.3) the challenges faced by participants during the search. We analyzed search sessions from 18 out of 36 participants mainly for two reasons. First, we reached a point of saturation after analyzing search sessions from 12 participants. That is, no new codes were introduced after analyzing the first 12 participants. We coded an additional 6 participants to ensure that the search sessions included in the analysis were balanced across our three manipulated variables (i.e., domain, product, and goal). Additionally, this type of analysis requires significant time and effort. Therefore, the decision to limit our analysis to 18 participants was also a practical one.

Qualitative Data Preparation: The first step in analyzing participants' search sessions was to represent each session as a sequence of *observation units*. Each observation unit was associated with a specific *action* taken by the participant either on the search interface, on a document, or on the Google Doc used to take notes. SERP-level actions included instances in which the participant issued a new query and decided to click or not click a specific search result. Document-level actions included instances in which the participant decided to read or explicitly ignore a specific piece of information in a document. Note-level actions included instances in which the participant copy/pasted information into the Google Doc or made a new note. Using Microsoft Excel, each observation unit was entered as a new row. For each observation unit, we marked the timestamp on the recorded video, wrote a textual description of the action performed by the participant, and, if applicable, transcribed any think-aloud comments made by the participant around the time the action was performed.

To validate our data preparation approach, two of the authors processed 9 sessions from 3 participants. The level of agreement was 96% based on the Jaccard Coefficient.[7] Given this high level of agreement, both authors processed the remaining participants separately. Ultimately, at the end of the data preparation phase, each search session was represented as a sequence of observation units that could be traced back to the recorded video using the timestamp. Our analysis included 888 observation units associated with 54 search sessions from 18 participants.

Qualitative Coding: The search sessions were analyzed using an inductive approach [20, 27], whereby the coding scheme was developed by examination of the observed behaviors and verbal comments in a bottom-up fashion. The coding process involved watching the video recordings of each session, generating codes, and assigning codes to observation units that were pertinent to: (1) relevance criteria, (2) types of information sought/used, and (3) challenges faced.

The coding process proceeded in two phases. During the first phase, two of the authors independently coded all sessions from the same three participants. Then, the authors met to develop an initial set of codes and reconcile their codes. A third author reviewed the codes and participated in the discussion. At this stage, we focused on: (1) clarifying the boundaries between the three aspects of RQ4 (i.e., relevance criteria, information types, and challenges); (2) clarifying the definition of codes within each aspect; and (3) distinguishing codes from each other to ensure that codes were semantically distinct. During the second phase, three of the authors coded the remaining data. Following a common practice recommended in team-based open coding, the authors shared a *living codebook* [10, 37] where they recorded new codes, definitions, and examples. The authors met

[7]The Jaccard Coefficient measures the similarity between two sets. It corresponds to the intersection divided by the union of the two sets.

periodically and used the document as a shared reference to discuss new codes and apply these to search sessions they had already coded. The final codebook is available online.[8]

4 RESULTS

Next, we report on results with respect to our four research questions (RQ1-RQ4). In RQ1-RQ3, we investigate the effects of the task's product and goal on participants pre-task perceptions, post-task perceptions, and search behaviors. To investigate these effects, we used mixed-effects ANOVAs with the task's product as a within-subjects factor and the task's goal as a between subject's factor. Since the task's product has three levels (factual, decision, intellectual), we used Bonferroni-corrected paired t-tests to test for differences between all pairs of conditions.[9]

4.1 RQ1: Effects on Pre-task Perceptions

In RQ1, we investigate the effects of the task's product and goal on participants' pre-task perceptions. Specifically, we focus on pre-task perceptions of: (1) prior knowledge and search experience in the task domain; (2) expected interest and enjoyment; (3) expected difficulty; (4) *a priori* determinability; and (5) perceptions of the task having a clear objective. Additionally, we focus on the extent to which participants expected the task to require: (6) learning about the domain by gathering facts, definitions, and background knowledge; (7) gathering subjective information to inform decision-making; and (8) gathering step-by-step instructions. Figures 2 and 3 show the effects of the task's product and goal on these perceptions. As in all figures related to RQ1-RQ3, we show means and 95% confidence intervals for different product and goal categories.

Effects of Product: As shown in Figure 2, the task's product had a significant effect on four types of perceptions.

First, the task's product had a significant effect on participants' prior knowledge and search experience ($F(2, 68) = 5.00$, $p < .01$). Participants reported significantly lower perceptions for intellectual versus factual tasks ($p < .01$) and decision tasks ($p < .05$).

Second, the task's product had a significant effect on participants' perceptions of the task having a clear objective ($F(2, 68) = 10.34$, $p < .001$). Again, participants reported significantly lower perceptions for intellectual versus factual tasks ($p < .005$) and decision tasks ($p < .005$).

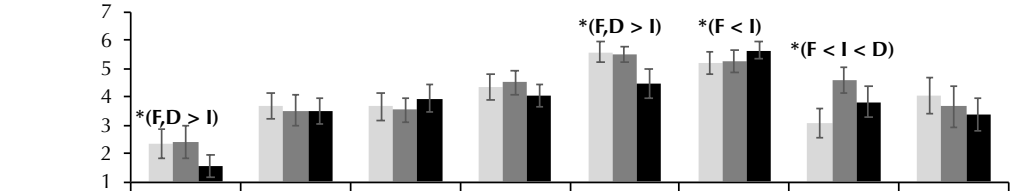
Third, the task's product had a significant effect on the extent to which participants expected the task to involve learning about the domain by gathering facts, definitions, and background information ($F(2, 68) = 5.00$, $p < .01$). Participants reported significantly higher perceptions for intellectual versus factual tasks ($p < .05$).

Finally, the task's product had a significant effect on the extent to which participants expected the task to require subjective information to make decisions ($F(2, 68) = 13.78$, $p < .001$). Here, we found significant differences between all pairs of product categories: (1) decision versus intellectual ($p < .05$); (2) decision versus factual ($p < .001$); and (3) intellectual versus factual ($p < .05$). Perceptions were highest for decision tasks and lowest for factual tasks.

Effects of Goal: The task's goal had a much weaker effect on participants' pre-task perceptions. As shown in Figure 3, the task's goal only had a significant effect on participants' perceptions of the task requiring step-by-step information ($F(1, 34) = 4.65$, $p < .05$). Participants reported higher perceptions for specific versus amorphous tasks.

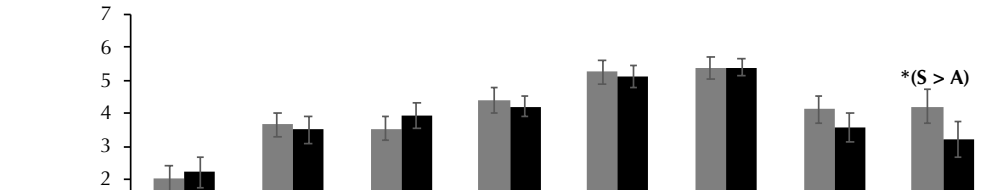
[8] https://ils.unc.edu/~jarguell/pk_lab_study/codebook.pdf

[9] We report on the main effects of product and goal. We also considered their interaction effects on all outcome variables and found no significant effects.



	prior know. search	interest enjoyment	difficulty	determin.	clear objective	req. background know.	req. subj. decide	req. step-by-step
factual	2.35	3.68	3.67	4.33	5.60	5.19	3.10	4.06
decision	2.42	3.53	3.56	4.51	5.49	5.28	4.61	3.67
intellectual	1.60	3.53	3.96	4.06	4.47	5.65	3.84	3.39

Fig. 2. Effects of the task product on pre-task perceptions.



	prior know. search	interest enjoyment	difficulty	determin.	clear objective	req. background know.	req. subj. decide	req. step-by-step
specific	2.03	3.66	3.54	4.40	5.26	5.36	4.12	4.20
amorphous	2.21	3.50	3.92	4.20	5.11	5.39	3.58	3.20

Fig. 3. Effects of the task goal on pre-task perceptions.

4.2 RQ2: Effects on Post-task Perceptions

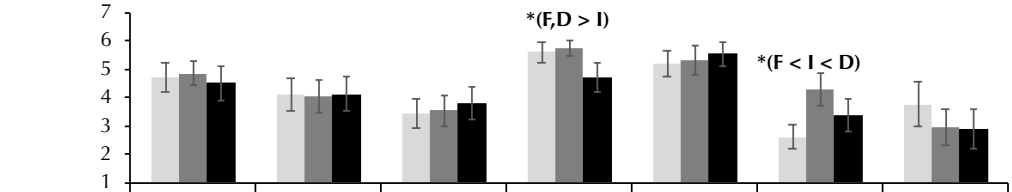
In RQ2, we investigate the effects of the task's product and goal on participants' post-task perceptions. Specifically, we focus on post-task perceptions of: (1) learning; (2) interest and enjoyment; (3) difficulty; and (4) perceptions of the task having a clear objective. Additionally, we focus on the extent to which participants perceived the task to require: (5) learning about the domain by gathering facts, definitions, and background knowledge; (6) gathering subjective information to inform decision-making; and (7) gathering step-by-step instructions.

Effects of Product: As shown in Figure 4, the task's product had a significant effect on two types of perceptions.

First, the task's product had a significant effect on participants' perceptions of the task having a clear objective ($F(2, 68) = 10.27$, $p < .001$). Participants reported significantly lower perceptions for intellectual versus factual tasks ($p < .005$) and decision tasks ($p < .005$).

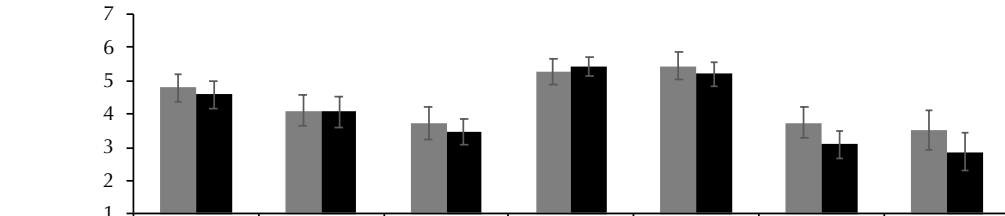
Second, the task's product had a significant effect on the extent to which participants expected the task to require subjective information to make decisions ($F(2, 68) = 15.51$, $p < .001$). Here, we found significant differences between all pairs of product categories: (1) decision versus intellectual ($p < .05$); (2) decision versus factual ($p < .001$); and (3) intellectual versus factual ($p < .05$). Perceptions were highest for decision tasks and lowest for factual tasks.

Effects of Goal: As shown in Figure 5, the task's goal did not have a significant effect of participants' post-task perceptions.



	know. gain	interest enjoyment	difficulty	clear objective	req. background know.	req. subj. decide	req. step-by-step
factual	4.72	4.11	3.45	5.59	5.17	2.60	3.75
decision	4.83	4.01	3.53	5.72	5.30	4.29	2.94
intellectual	4.50	4.13	3.79	4.69	5.52	3.36	2.86

Fig. 4. Effects of the task product on post-task perceptions.



	know. gain	interest enjoyment	difficulty	clear objective	req. background know.	req. subj. decide	req. step-by-step
specific	4.80	4.10	3.72	5.25	5.44	3.75	3.50
amorphous	4.57	4.07	3.47	5.42	5.21	3.09	2.87

Fig. 5. Effects of the task goal on post-task perceptions.

4.3 RQ3: Effects on Search Behaviors

In RQ3, we investigate the effects of the task's product and goal on participants' search behavior. Specifically, we focus on behavioral measures described in Section 3.5. To conserve space, we only include figures for measures with significant differences based on the task's product and goal.

Effects of Product: As shown in Figure 6, the task's product had a significant effect on four behavioral measures.

First, the task's product had a significant effect on the number of queries issued by participants ($F(2, 68) = 4.14$, $p < .05$). Participants issued more queries during factual versus intellectual tasks ($p < .05$).

Second, the task's product had a significant effect on the number of search results clicked during the session ($F(2, 68) = 4.72$, $p < .001$). Participants clicked on more search results during decision versus intellectual tasks ($p < .05$).

Finally, the task's product had a significant effect on the percentage of unique queries ($F(2, 68) = 8.13$, $p < .001$) and unique URLs clicked ($F(2, 68) = 4.14$, $p < .05$). As described in Section 3.5, these

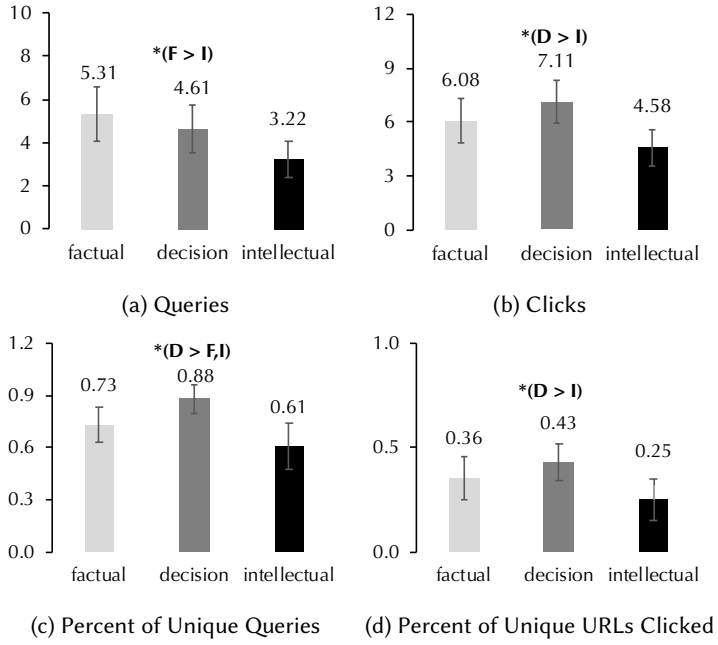


Fig. 6. Effects of the task product on search behaviors.

measures refer to the percentage of queries not issued by another participant and the percentage of URLs not clicked by another participant. Participants issued more unique queries during decision tasks versus factual tasks ($p < .05$) and intellectual tasks ($p < .005$). Participants clicked on more unique URLs during decision versus intellectual tasks ($p < .05$).

Effects of Goal: The task's goal did not have a significant effect on any behavioral measure.

4.4 Summary or RQ1-RQ3 Results

Effects of the Task Product (Factual vs. Decision vs. Intellectual): Our intellectual tasks were the most open-ended. They asked participants to consider the pros and cons of a procedure or determine the types of problems that can be solved with one or more resources. In terms of pre- and post-task perceptions (RQ1-RQ2), participants reported the lowest levels of pre-task prior knowledge for intellectual tasks. Additionally, intellectual tasks were perceived to lack a clearly defined objective (pre- and post-task) and require more background information (pre-task). In terms of search behaviors (RQ3), intellectual tasks that the fewest number of queries and clicks. One possible explanation is that intellectual tasks involved more reading and less searching.

Our decision tasks asked participants to evaluate alternative approaches to a task or problem. In terms of pre- and post-task perceptions (RQ1-RQ2), participants perceived decision tasks to require more subjective information (e.g., people's opinions) to aid in decision-making (pre- and post-task). In terms of search behaviors (RQ3), decision tasks had the most queries not issued by other participants and the most URLs not clicked by other participants. Our decision tasks asked participants to evaluate alternatives but did not specify the exact alternatives that participants should consider. One possible explanation is that different participants considered different alternatives, resulting in more divergent behaviors (i.e., more unique queries issued and URLs clicked).

Table 7. RQ4 summary of results.

Relevance Criteria	Information Types	Challenges
• Task-relevant verbs & nouns • Source familiarity/trust • Source type • Level of specificity • Situational similarity • Constraints • Form of presentation • Intended audience	• Definitions • Background information • How-to information • Applied information • Examples • Experiential information • Tips & advice • Rules of thumb • Visuals	• Terminology issues • Knowing where to start • Insufficient details • From general to specific & vice-versa • Dealing with uncertainty • Conceptual knowledge gaps • Finding relevant verbs • Fixation effects

Effects of the Task Goal (Specific vs. Amorphous): Broadly speaking, as described in Tables 2-4, our amorphous tasks involved greater uncertainty. Participants reported requiring less step-by-step information for amorphous tasks. One possibility is that participants required other types of information during amorphous tasks (e.g., information about task feasibility).

4.5 RQ4: Qualitative Analysis Results

In RQ4, we investigate how people search for procedural knowledge irrespective of the task's product and goal. To this end, we conducted a qualitative analysis of think-aloud comments and screen activities from all search sessions from 18 (out of 36) participants. Our qualitative analysis focused on three main aspects: (RQ4.1) relevance criteria, (RQ4.2) types of information sought, and (RQ4.3) challenges. Table 7 summarizes our RQ4 results with respect to these three main aspects.

4.5.1 RQ4.1: Relevance Criteria. In our qualitative analysis of participants' think-aloud comments, we identified eight categories of relevance criteria used by participants in their searches for procedural knowledge: (1) the presence of task-relevant verbs and nouns; (2) familiarity/trust in the source; (3) the type of source (e.g. discussion forum vs. manufacturer webpage); (4) the level of specificity of the information; (5) the similarity of the described situation to the participants' situation; (6) task constraints; (7) the form of the presentation of the information (e.g. table, list, image); and (8) the perceived intended audience of the information.

Task-relevant Verbs and Nouns: Procedural tasks involve doing something to one or more things. In this respect, procedural tasks involve verbs and nouns. For example, our tasks involved computing a P/E ratio, finding a land record, and connecting a computer to a monitor. Not surprisingly, participants described looking for task-relevant verbs and nouns as indicators of a document's relevance. Participants looked for verbs that referred to specific actions involved in the process (e.g., "calculate", "search", "troubleshoot"). Similarly, participants looked for nouns mentioned in the task description (e.g., "P/E ratio", "land grant", "Dell FP2007") as well as broader concepts related to the task domain (e.g., "stock investment", "land record", "port").

Participants sometimes faced challenges when using verbs and nouns to judge relevance. Some participants had to learn about task-relevant verbs not included in the task description. For example, one of our tasks asked participants to help a friend refurbish a set of old Lenovo laptops. Participants sometimes looked for documents containing the verb "refurbish" but missed documents containing relevant verbs such as "clean", "(re)install", "upgrade", and "test". P26 learned that refurbishing can involve repairing and then used the word "troubleshoot" to judge relevance.

With respect to task-relevant nouns, participants sometimes struggled with name variants. For example, one of our technology tasks asked participants to determine if a Lenovo X1 second

generation tablet computer could be used with a Pluggable UD-ULTCDL docking station. During this task, participants struggled to determine if documents referring to "Lenovo X1 tablet", "Lenovo Thinkpad X1", and "Lenovo X1 Yoga Gen 2" were relevant to the task. For example, are different name variants referring to the same computer model? What are the essential parts of "Lenovo X1 second generation tablet computer"? Does the "second generation" aspect make a difference? What about documents mentioning to "X1" but not "tablet"? P25 noted: "So the problem is I don't see specifically a Gen 2 listed... I see X1 tablet but it doesn't say 2nd gen... and then I see a Lenovo X1 or Lenovo ThinkPad X1 Yoga Gen 2 but that doesn't match the description."

Source Familiarity/Trust: Participants described being familiar with and trusting specific sources. For example, for our genealogy tasks, participants described being familiar with websites such as Archives.gov and Ancestry.com. Similarly, during our technology tasks, participants mentioned that they trusted manufacturer websites (e.g., Microsoft, Lenovo) for authoritative information about specific products. Participants also described situations where a lack of familiarity/trust caused them to ignore specific websites. For example, P13 noted: "It says stateofflorida.com... I'm guessing that is not really the State of Florida [website]".

Source Associated with Information Type: Participants associated specific sources with certain types of information, which helped them decide whether a website might contain relevant information. For example, P1 associated Wikipedia with basic information such as definitions: "Since I don't know what price-to-earnings ratio is, I'm going to start by looking at Wikipedia." In another example, P25 said: "I'm going to try one more thing and just see if the monitor has any reviews". Then P25 went to the website CNET to read reviews from "a trusted source".

Level of Specificity of the Information: Sometimes participants needed specific information to help them with their procedural task. Other times they needed general information to help them understand aspects of the task domain and context.

Sometimes general information prompted participants to look for more specific information. For example, for the task of connecting a Microsoft Surface Pro 3 computer to a Dell 2007FP monitor, P25 found a page with general information titled "Connect Surface to a Monitor". This page instructed readers to "look at the video ports on your TV, monitor, or projector [to] determine what adapters and cables you need." Based on this general instruction, P25 realized they needed more specific information: "So what I want to do is... go to a new search... to see the video port specs for the Dell 2007FP... so I'm clicking the Dell page."

Other times participants moved in the opposite direction (from specific to general information). For example, while working on a genealogy task, P15 found information about land records and noted: "So this is specifically about [land] records... I don't feel like I'm ready to go straight to that... I want to know about the Homestead Act... so I'm just going to go to the encyclopedia [clicked on Britannica.com]".

Participants also described finding information that was at just the right point between general and specific. For example, during a genealogy task, P15 found a page on the U.S. National Archives website with a section titled "How Land Records Can Help You". P15 commented: "So this is talking about land records in general... I think this is good background information for what land records could help me find out... so even if I don't get a lot of specific information from the documents, they would be helpful in a greater search for finding out more about my ancestors."

Task/Situational Similarity: Procedural tasks are often situated in unique contexts. For example, searchers may have certain types of prior experience, constraints, and limitations that may affect the feasibility of different solutions/approaches. Participants described how the similarity of the situation described in a webpage influenced their view of its relevance to their own situation. For example, while deciding which search results to click on, P1 noted: "My scenario is that I'm going to try to compare P/E ratios... I'm trying to understand what the P/E ratio means for someone

who is using it to compare companies." In another example, while working on a genealogy task, P14 noted that the hyperlink "property search" on a webpage did not match their need for *historical* information: "Hmm... property search... probably not historical though... I don't think this will be super helpful."

Task Constraints: As is typical of procedural tasks, our tasks involved different constraints. Participants often questioned whether information was relevant to their constraints. Interestingly, participants sometimes struggled with deciding which constraints were essential and which could be relaxed or completely ignored. For example, for the task of helping a friend refurbish Lenovo laptops from 2008-2012, P33 struggled with deciding whether they needed different information for each year.

Form of Presentation: Participants also noted that the form in which information was presented on a webpage influenced their relevance assessments. For example, P7 described how a web page that included a formula helped clarify the concept of dividend yield: "[The] formula for dividend yield is as follows... annual dividends per share divided by price per share... it's kind of nice to see it in a formula." Similarly, P1 described how seeing information presented in list form influenced their relevance judgements: "This website said pros and cons but then it didn't really have a nice pros and cons list... So I'm going to try the next one [which did have a list]." During a technology task, P33 wanted to know what a particular monitor connector looked like and commented on wanting a labelled image. Finally, in several cases, participants commented on wanting information in table form during tasks that involved comparing alternatives and assessing the compatibility between products.

Intended Audience: A document's intended audience (as perceived by participants) was also a factor in determining relevance. For example, during a technology task, P31 noted: "Okay... so we're gonna go with WikiHow... they say things in a way that people like me can understand." Similarly, during a finance task that involved computing P/E ratio, P7 noted that a webpage assumed pre-requisite knowledge that they did not have: "Okay, so here's the formula. I mean, it's just like what it says... it's the market value versus the earnings... I don't know where you find this information." Finally, during another finance task, P3 noted: "that [webpage] kind of assumed I knew what I was doing and that I was ready to start investing... I'm going to go back and read something else."

4.5.2 RQ4.2: Information Types: Participants searched for nine different types of information: (1) definitions; (2) background information; (3) how-to information; (4) applied information; (5) examples; (6) experiential information; (7) tips and advice; (8) rules of thumb; and (9) visuals.

Definition: Participants searched for definitions of important concepts related to the task. Some tasks required more conceptual understanding than others. For instance, in order to complete our finance tasks, participants needed to gain a basic understanding of different stock valuation metrics. Such metrics are intended to be used by investors/analysts and are therefore highly technical. Many participants started our finance tasks by searching for definitions of those concepts (e.g., P/E ratio, dividend yield, etc.). To many participants, a basic definition was insufficient; they searched further for an extended definition, which often included background information or applied information.

Background Information: Participants searched for background information about the task domain. Background information included historic information and information about assumptions associated with a solution/approach. For instance, one of the genealogy tasks asked participants to explain how they could use information in a U.S. Homestead land entry file to learn about their family's immigration history. Participants often begun this task by searching for information about the Homestead Act of 1862. Similarly, some of our finance tasks asked participants to compare different stock valuation metrics. Some participants begun the task by searching for information

about the value of investing in stocks. Background information may not directly support the execution of a procedural task. However, it helped participants define and understand the task in a broader context.

How-to Information: As expected, participants searched for step-by-step instructions on how to complete the task or a sub-step of the task. For example, while working on one of our technology tasks, participants wanted step-by-step instructions on how to install Linux Mint on a ThinkPad 10 computer. Participants also sought step-by-step instructions for tasks that did not involve tangible objects. For example, for one of our finance tasks, participants sought step-by-step instructions on how to compute a P/E ratio. Similarly, for one of our genealogy tasks, participants sought step-by-step instructions on how to request land records.

Applied Information: Participants searched for both conceptual and procedural information placed in a *specific, real-world* context. Examples included: (1) explanations of why a concept or procedure is important; (2) explanations of how a concept can be used in practice (i.e., a use-case scenario); and (3) explanations of how the outcome of a procedure should be interpreted. For example, for tasks that involve computing P/E ratio, participants found the following information useful: "Investors use P/E ratio to determine the relative value of a company's shares in an apples-to-apples comparison"; "P/E ratios are useful in comparing a company to similar companies or their own historical records"; and "A high P/E ratio could mean a company's stock is overvalued or that investors expect high growth in the future."

Examples: Participants found examples useful across task topics. For example, for tasks that involve computing P/E ratio, participants were biased toward search results whose title contained the term "example". Such results demonstrated the steps to compute a P/E ratio using synthetic or real-world data. Sometimes, participants were able to use an example to answer their questions. For example, when trying to find the kinds of information in a U.S. Homestead land entry file, P13 found an example and learned about the fields in the file.

Experiential Information: Participants searched for information from people with firsthand experience with the same task or similar tasks. First-person accounts illustrate how someone else approached the task, what they did, what they avoided, and whether their approach worked (e.g., trial-and-error). Participants sought experiential information in online reviews and forums. For example, during one of the genealogy tasks, P15 noted: "Learning how other people have used these [land entry files] to find out more about their immigration history would be useful." First-person accounts sometimes served as examples demonstrating the procedure in a real-world context.

Tips/Advice: Participants searched for expert tips and advice about the task. For example, one of our genealogy tasks involved researching genealogy tools, selecting the best one, and providing a justification. During this task, participants often looked for recommendations provided by seemingly reputable sources. Tips often provided "plan B" alternatives and highlighted the importance of certain steps. For example, during one of our technology tasks, P32 found the following information valuable: "If you cannot find your ThinkPad model listed, you can find the required power information on a label on the bottom of your ThinkPad." Similarly, P26 found the 'tips' section on a WikiHow page (i.e., how to refurbish a computer) valuable.

Rules of Thumb: Interestingly, participants found rules of thumb particularly useful. Rules of thumb are statements that are generally true or recommendations from experts based on experience. Different from tips/advice, rules of thumb might include insights, guides, principles, and heuristics that are *broadly* applicable to many situations. A tip might be: "When making this recipe, add one tablespoon of salt if you are using unsalted chicken broth." Conversely, a rule of thumb might be: "When seasoning a dish, start with a small amount of salt and add more as needed."

For example, during our technology tasks, participants found the following rules of thumb useful: "Ports you usually see include HDMI, DisplayPort, and VGA"; "Adapters and cables are often sold

separately"; and "In most cases, DVI outperforms VGA in every way." As another example, during a finance task, P2 found the following statement helpful: "Trailing P/E ratios are the most commonly used ratio."

Visuals: Participants found visual information (i.e., images, videos) useful. We observed three ways in which visual information helped participants.

First, participants used images to learn about unknown technical terms needed to address their needs. This was especially true during tasks involving multiple physical components, such as our technology tasks. For example, P33 needed a specific adapter that they could not remember the name of. After searching for "cable with the two screws", they found an image of the adapter. This led them to a webpage that included the adapter's name. P33 noted: "Yes, I am looking for those two... I need pictures... I don't even know what they are called."

Second, visual information helped participants identify the different parts of a physical product. For example, participants found labeled diagrams of a product useful during our technology tasks. Such diagrams helped participants understand the different physical objects involved in the task.

Third, sometimes, visual information gave participants a definite answer to their questions. For instance, for the task of connecting a specific monitor to a specific computer, seeing an image of an adapter helped participants determine that the task was feasible.

4.5.3 RQ4.3: Challenges. Participants experienced nine types of challenges during their procedural search tasks: (1) terminology issues; (2) not knowing where to start; (3) insufficient details; (4) transitioning from general-to-specific or specific-to-general information; (5) dealing with uncertainty; (6) lacking conceptual knowledge; (7) finding relevant verbs; (8) fixation effects; and (9) other issues.

Terminology Issues: Participants faced challenges associated with terminology related to the task domain. These included: (1) not knowing the definition of a term or abbreviation; (2) encountering too many unknown terms and not knowing which ones to prioritize; (3) making incorrect inferences about a term's meaning from its context; (4) not knowing which terms uniquely identify a specific entity; (5) not knowing whether different terms are synonymous; and (6) not knowing how to reconcile different definitions of the same term.

To illustrate, with respect to #4, some participants had difficulty determining whether "Thinkpad" is the same as "Lenovo Thinkpad". In this case, "Thinkpad" is the unique identifier and "Lenovo" can be ignored (i.e., all Thinkpads are made by Lenovo). With respect to #5, some participants had difficulty determining whether "Dell 2007FP" is the same as "Dell Ultrasharp". In such cases, participants had difficulty constructing effective queries and judging relevance. Finally, with respect to #6, participants had difficulty reconciling different definitions of the same concept. For example, when gathering information about "P/E ratio", participants encountered different definitions, such as "the ratio between a company's stock price and its earnings-per-share" and "how much investors pay per dollar of annual company earnings". In such cases, participants struggled to determine whether different definitions were equivalent or whether one of them was incorrect or incomplete.

The above challenges are associated with terms encountered during the search. Other challenges were associated with important terms that were unknown to the participant and *not* encountered during the search. For example, one of our genealogy tasks asked participants to determine a way to verify whether land was granted to a specific individual by the U.S. federal government. This task was much easier if participants learned about the term "land patent".[10] Participants who did not learn about this term had greater difficulty.

Not Knowing Where to Start: In some cases, participants faced challenges in deciding how to start the task. These challenges mostly stemmed from participants' unfamiliarity with the task

[10]A land patent is a legal document that verifies that a specific piece of land has been granted to an individual.

domain. For example, for the genealogy task described above, participants did not know whether they should start by searching for the individual's name or an online database that might help with completing the task.

Insufficient Details: Participants also struggled when they encountered information that lacked important details. For example, when reading step-by-step instructions, participants sometimes wanted additional information about: (1) how to execute a step and (2) the rationale behind a specific step (i.e., Why is it important or necessary?). Such details were sometimes missing from procedural documents encountered by participants.

To illustrate, with respect to #1, participants were frustrated when they found information on how to compute "P/E ratio" but not information on how to gather the necessary values to compute the "P/E ratio" for a specific company in a specific year. P7 located information that explained the components necessary to calculate the P/E ratio. However, the page did not provide information on where to source the values needed. In response, P7 said: "Now I don't know where to find this information... one must divide the current stock price by the earnings per share... but where is the earnings per share?" Similarly, with respect to #2, participants struggled with computing "P/E ratio" when they did not understand how it can be used to determine if a stock is overvalued or undervalued.

From General to Specific and Vice-Versa: Procedural tasks often involve personal preferences and situational constraints. Personal preferences can be influenced by an individual's prior knowledge, known skills, and personality traits. Situational constraints can be influenced by resources, tools, and materials that are readily available. This means that procedural documents can have different levels of specificity. To illustrate, consider a document that describes how to execute a procedure with an example. An example can be highly general—describing a common or generic implementation of the procedure—or highly specific—heavily influenced by specific preferences and constraints of the author.

Our results found that participants faced challenges with highly general and highly specific information. When participants encountered highly general information, they did not know how to tailor the procedure to fit their more constrained scenario. Conversely, when participants encountered highly specific information, they did not know how to generalize the procedure to a less constrained scenario.

To illustrate, one of our tasks asked participants to learn how to connect a Microsoft Surface Pro 3 tablet to a Dell 2007FP monitor. One participant struggled to make use of step-by-step instructions on "How to Connect a Microsoft Surface Tablet to a Monitor". In this case, the participant struggled to decide whether these general instructions were relevant to the exact models specified in the task description (i.e., going from general to specific).

In other cases, participants encountered highly specific information when they needed more general information. For the task of refurbishing old ThinkPads, P33 searched "how to install linux onto thinkpad", looking for general information about the installation process. However, the first result retrieved was "Installing Linux Mint 13 Xfce on a ThinkPad X200". P33 noted: "This is referring to the process for a specific model, the X200... they might be radically different... I'm assuming they're similar processes... [but] that's just an assumption."

Dealing with Uncertainty: Participants struggled with three different types of uncertainty. First, procedural tasks can involve objectives that may not be attainable. Indeed, some of our amorphous tasks explicitly indicated that the objective may not be attainable. Participants had to decide whether the task's objective was attainable given available resources (e.g., free online information). Some participants indicated that this made the task daunting.

Second, participants struggled with not knowing what the outcome of a task should look like. For example, one of our finance tasks asked participants to compute the "P/E ratio" for Apple Inc.

Participants struggled with not knowing whether a specific outcome is *typical*. Is a typical "P/E ratio" in the tens, hundreds, or thousands? Not knowing whether a specific outcome is typical prevented participants from knowing if they had implemented a procedure correctly.

Third, participants struggled with not knowing if there is a *typical* way to execute a procedure. In such cases, participants wanted to find "rules of thumb", "best practices", and "common heuristics" associated with a procedure. Participants had difficulty finding this type of information. For example, while learning to compute "P/E ratio", some participants learned about different variants (e.g., "backward P/E ratio" and "forward P/E ratio"). In such cases, participants struggled with deciding which variant is more commonly used.

Conceptual Understanding: Procedural search tasks often involve important concepts. For example, our finance tasks involved different metrics that are used to evaluate companies from different perspectives. These tasks involved concepts such as "stock price", "earnings-per-share", and "dividends". Participants had difficulty determining whether they understood such concepts well enough to complete the task. For example, one of our finance task asked participants to judge average dividend yield as a metric to evaluate stocks. P7 struggled to understand the concept of "dividend yield". This prevented them from comprehending almost all the subsequent information encountered during the search session. For example, they struggled to understand what a high or low "dividend yield" means. P7 noted: "The dividend yield shows how much a company pays out in dividends each year relative to its stock price. Oh, I don't know what that means... the higher the dividend yield, the greater the possibility that a stock price is undervalued... I am fading on this one."

This was often the case for tasks that asked participants to *implement* a procedure. Based on Anderson and Krathwohl's taxonomy [2] (Section 2.1), implementing a procedure is an *apply*-level task. It may be that such tasks require a deeper understanding of relevant concepts than simpler tasks, such as memorizing the definition of a concept (i.e., a *remember*-level task) or recognizing an example of a concept (i.e., an *understand*-level task).

Finding the Relevant Verb: Procedural tasks often involve *both* verbs and nouns (e.g., "refurbishing laptops", "computing price-to-earnings ratio", "connecting laptop to docking station"). Participants often issued queries that retrieved documents that were relevant to the noun but *not* the verb of the task. For instance, for the task of refurbishing old ThinkPads, P27 initially searched "refurbish IBM Lenovo" but found no results with the term "refurbish" in their titles. We see two possible explanations for this trend. First, some of our tasks involved proper nouns (e.g., "Lenovo X1 Tablet"), which have relatively few name variants. Second, it may be that search systems are more strongly influenced by query-terms that are nouns versus verbs. For example, it may be that verbs (e.g., install, compute, connect) are more frequent and therefore less influential in the ranking of documents (i.e., have lower IDF weights).

Fixation Effects: Participants often had expectations about the task that turned out to be problematic. These included expectations about requirements of the task, how it should be approached, and the form of the solution. In some cases, participants had difficulty abandoning or adjusting these expectations during the search. Participants' fixation on preconceived notions about the task sometimes prevented them from solving a problem or recognizing relevant information. For example, for one of our genealogy tasks, two participants expected to find a searchable database of land records. Both participants failed to explore other approaches and could not complete the task.

Other Issues: Finally, participants faced challenges that are not specific to procedural search tasks. These included: (1) encountering documents with too much text; (2) not knowing the trustworthiness of a specific resource; and (3) seeking information that was not readily available (e.g., the price of a specific online service).

5 DISCUSSION

We conducted a user study ($N = 36$) to investigate the effects of two dimensions of procedural search tasks (i.e., product and goal) on participants' perceptions (RQ1-RQ2) and search behaviors (RQ3). Additionally, we conducted a qualitative analysis of 18 participants' think-aloud comments and screen activities to identify important relevance criteria, information types, and challenges (RQ4). In Sections 5.1-5.4, we discuss our findings in relation to prior work. In Section 6, we discuss tools and interface features that may better support users with procedural search tasks.

5.1 RQ1-RQ3: Effects of Task Product and Goal

Effects of Task Product: Participants completed procedural search tasks that varied along three product categories: *factual* tasks asked participants to execute a procedure; *decision* tasks asked participants to evaluate alternatives; and *intellectual* tasks asked participants to generate new ideas.

Before and after the task, factual and decision tasks were perceived to have clear objectives. However, factual tasks were perceived to require the least amount of subjective information (e.g., opinions) and decision tasks were perceived to require the greatest amount of subjective information. Intellectual tasks were perceived to have unclear objectives and require *some* amount of subjective information (i.e., more than factual but less than decision tasks). Additionally, for intellectual tasks, participants reported the lowest levels of pre-task prior knowledge and anticipated needing the greatest amount of background information.

Our results suggest that our tasks varied by complexity, with factual tasks being the simplest and intellectual tasks being the most complex. Search task complexity has been characterized from different perspectives [49]. From the perspective of cognitive complexity [2] (see Section 2.1), our factual tasks can be viewed as *apply*-level tasks (moderately complex), our decision tasks as *evaluate*-level tasks (more complex), and our intellectual tasks as *create*-level tasks (most complex). Additionally, our intellectual tasks were the most open-ended. Researchers have argued that open-ended tasks are complex because they have many possible outcomes [7] and greater uncertainty about the outcomes [6].

Our results resonate with prior work. Choi et al. [14] also found that participants perceived opinions and insights to be more useful during complex versus simple tasks. Similarly, Byström and Järvelin [6] and Freund et al. [22] found that searchers prefer information from other people with firsthand experience during complex tasks. Our results extend these prior findings by showing that complex *procedural* search tasks also involve greater use of subjective information.

In terms of search behaviors, our intellectual tasks had the fewest queries and clicks. Additionally, they had the fewest queries and clicked URLs not issued/clicked by other participants. We attribute this trend to the fact that our intellectual tasks involved generating new ideas on how a given resource (e.g., a specific website) might be useful in completing procedural tasks. Therefore, intellectual tasks required less searching and more of other activities (e.g., browsing and brainstorming). Prior studies have found that cognitively complex tasks involve more search activity [9, 14, 26]. However, our results suggest that this is not always the case.

Effects of Task Goal: Our manipulation of the task goal did not have strong effects. As one exception, before the task, participants anticipated amorphous (vs. specific) tasks to require less step-by-step information. Our amorphous tasks were designed to involve greater uncertainty (e.g., task feasibility). One possibility is that participants perceived amorphous tasks to require *different* types of information beyond step-by-step information (e.g., information about the feasibility of the task). Prior work has also found that search tasks with greater uncertainty require more diverse types of information [11, 14].

5.2 RQ4.1: Relevance Criteria

Participants used eight categories of relevance criteria to decide whether information was useful to them: (1) task-relevant verbs and nouns; (2) source familiarity/trust; (3) source type (e.g., discussion forum, review website); (4) level of specificity; (5) situational similarity; (6) task constraints; (7) form of presentation; and (8) target audience. Many of these categories are not new. For example, topicality, reputation, depth, scope, and understandability of the information are important relevance criteria that have been documented in prior work [4, 5, 17, 38, 42, 51].

Our results on relevance criteria make two main contributions. First, they validate prior findings by showing that common relevance criteria also apply to procedural search tasks. Second, some relevance criteria seem uniquely important during procedural search tasks. To support this claim, we compare against findings from two of our previous studies: (1) Choi et al. [13] considered relevance criteria used by intelligence analysts using a procedural knowledge base and (2) Choi et al. [12] considered relevance criteria used by people during everyday procedural search tasks. Next, we discuss themes encountered across our studies.

Source Familiarity/Trust: Source familiarity and trust are important relevance criteria during procedural search tasks. In Choi et al. [13], participants often used author reputation to judge relevance. In Choi et al. [12], participants noted that credibility and popularity were important relevance criteria. In our study, participants trusted different types of sources depending on the task domain. For example, they trusted encyclopedic articles during our finance tasks, government websites during our genealogy tasks, and manufacturer websites during our technology tasks.

Level of Specificity: Prior work has noted that "specificity" is a common relevance criterion [5, 38]. Our prior studies also found that the "level of detail" on a page is important [12, 13]. In all three studies, participants complained that procedural documents often include step-by-step instructions but lack details on how to execute specific steps. However, our results in this study also suggest that specific information is not always better—participants sometimes wanted *general* information that is broadly applicable to tasks of the same type.

Task Constraints and Situational Similarity: Procedural tasks often involve constraints such as available tools, materials, and resources, as well as preferred methods and techniques. Results from all three studies suggest that task constraints are important in judging relevance. In Choi et al. [13], participants complained that documents do not always explicitly list the tools and resources needed to execute a procedure. In Choi et al. [12], participants commented on a wide range of constraints that influenced their decisions about information being useful (e.g., inputs, time, money, external help, etc.). A novel finding in our study is that participants, in the process of judging relevance, sometimes struggled in deciding whether specific constraints could be relaxed.

Presentation Format: Presentation format is an important criterion during procedural search tasks. Choi et al. [12] also found that participants often preferred documents with visuals, lists, and easily recognizable blocks of text such as inputs, steps, tips, alternatives, and pros & cons.

Target Audience: Procedural documents often have a target audience (e.g., novices vs. experts). In all three studies, participants mentioned wanting procedural documents that matched their level of knowledge in the task domain.

5.3 RQ4.2: Information Types Used

We observed that participants sought and used different types of information: (1) definitions; (2) background information; (3) how-to information; (4) applied information; (5) examples; (6) experiential information; (7) tips; (8) rules of thumb; and (9) visuals. Some of these deserve further discussion.

Declarative and Procedural Knowledge: It is established that one needs *both* declarative and procedural knowledge to be able to perform real-world tasks [41, 44]. In our data, participants sought declarative information such as definitions and background information, as well as procedural information such as step-by-step instructions.

Applied Information: Applied information was somewhat unexpected. Applied information describes a *real-world context* in which a procedure can be used. Applied information allowed participants to imagine use cases for a procedure, and to understand the relevance of a procedure in a larger context. Applied information made the procedural task more relatable.

Experiential Information: The importance of experiential information during procedural search tasks resonates with findings from our two prior studies. Choi et al. [13] found that intelligence analysts often search an internal procedural knowledge base to connect with domain experts. Similarly, Choi et al. [12] found that participants often gained "tips and warnings" from people with firsthand experience with the task. Additionally, survey respondents were asked to describe websites visited during their procedural task. About 10% of websites were social forums.

Examples: Examples serve an important function during procedural search tasks. Choi et al. [12] also found that participants valued seeing examples during their procedural search tasks. Compared to abstract instructions, seeing an *instantiation* of a procedure being executed provided participants with more concrete knowledge about what to do, how, and what to expect. We observed the same trend in the current study. Other studies have also found that people often prefer examples versus abstract instructions due to their specificity and concreteness [19, 28].

Visuals: Participants relied heavily on visual materials (i.e., images or videos). Choi et al. [12] also found that visual content helped participants understand the environment of the task and specific actions/movements that are difficult to describe verbally. Other research has also found that people rely heavily on video materials during procedural search tasks [32, 34, 50].

Diversity of Types: Lastly, participants leveraged different types of information during their searches. As previously noted, in terms of cognitive complexity [2], our tasks ranged from moderately to highly complex. Thus, this trend resonates with prior work, which found that complex tasks require different types of information [6, 11, 14].

5.4 RQ4.3: Challenges Faced

Participants encountered a wide range of challenges, including: (1) terminology issues; (2) not knowing where to start; (3) insufficient details, (4) issues with documents being too general or specific; (5) dealing with uncertainty; (6) understanding important concepts; (7) finding relevant verbs; and (8) fixation effects. Below, we highlight several of these to provide insights into how challenges manifest during procedural search tasks. We also discuss how our findings relate to prior work.

Insufficient Details: Procedural tasks involve steps that must be followed and specific inputs/tools that must be used. Participants were frustrated when documents lacked detailed information about how to execute specific steps and their rationales. Understanding the rationale behind a step can reduce uncertainty and enable searchers to *modify* a procedure to fit their unique situation (e.g., preferences and constraints). Choi et al. [12] also found that participants needed implementation details about steps and their rationales.

General vs. Specific Information: Participants had difficulty with documents that were too general or too specific. When documents were too general, participants did not know how to apply the procedure to their more specific situation. When documents were too specific, participants did not know which components of the procedure were generalizable (applicable to variations of the task) and which were not generalizable (applicable only to the specific scenario described in the document). Choi et al. [12] found a similar trend.

Dealing with Uncertainty: Procedural tasks involve a strong *execution* component. People must *perform* each step as opposed to only understand it. This execution component made participants feel uncertain about: (1) the feasibility of the task; (2) whether an approach might apply to their unique situation; and (3) what the outcome of the task should look like. In our two prior studies, participants also reported struggling with uncertainty. Choi et al. [13] found that participants struggled with not knowing if the sought-after information existed. Choi et al. [12] found that participants reported gaining confidence from specific types of information (e.g., comments about the task being easy to perform).

6 IMPLICATIONS FOR SYSTEM DESIGN

Based on our results, the following novel search tools may support users with procedural searches.

Alleviating Terminology Issues: Participants struggled with different types of terminology issues, especially when they lacked domain knowledge. Given a query, systems should suggest important terms and concepts to help searchers learn about the domain. Additionally, systems should suggest terms that are synonymous with certain query terms. For example, given the query "PE ratio", a system could emphasize that "price-to-earnings ratio" and "P/E ratio" also refer to the same concept. Finally, systems should highlight query-terms that are unnecessary to identify a unique entity. For example, given the query "Lenovo Thinkpad", the term "Lenovo" is unnecessary because all Thinkpads are made by Lenovo. This way, systems could help searchers know: (1) which terms could be removed in subsequent queries and (2) which terms should be prioritized when judging relevance.

Connecting Documents: Search systems typically treat documents as independent units. Search systems could support users by linking information from different documents. Our results suggest that procedural search tasks require more than just step-by-step information. For documents with step-by-step instructions, systems could link the steps to other documents that describe: (1) the inputs and outputs of the step; (2) relevant concepts and definitions; (3) the rationale behind the step; (4) details on how to *execute* the step; (5) alternative ways to achieve a similar outcome; (6) tips and advice; (7) rules of thumb; and (8) visuals. Linking documents could alleviate some of the challenges encountered by our participants, such as lacking conceptual knowledge, encountering documents with insufficient details, and dealing with uncertainty (e.g., not knowing what the outcome of a step should look like).

Dealing with Constraints: Systems could help searchers deal with constraints in two ways. First, given a procedural query (e.g., "making lasagna"), systems could suggest queries that are expanded with constraints that are typical of the task (e.g., "making vegetarian lasagna" or "making quick lasagna"). Such a feature might even help searchers become aware of constraints they have not thought of. Second, systems could automatically highlight specific types of information within documents to help searchers judge relevance based on their constraints. For example, systems could enable searchers to highlight parts of a document that mention: (1) required tools and materials; (2) prerequisite skills; (3) price information; and (4) time commitment information.

Query-by-Example with Specificity Adjustment: The following tools involve querying-by-example, which enables searchers to issue documents as queries. Procedural documents have different levels of *specificity*. Our participants sometimes struggled with documents that were too specific for their needs. In some cases, they wanted to gain a more general understanding of a procedure before delving into the details. In other cases, the specific scenario described in the document was different from theirs and they did not know if it was applicable to their scenario. In both cases, searchers might benefit from a more general document. Systems could enable searchers to issue a specific document as a query and request a more general version of the procedure (e.g., going from a recipe for "Pastitsio or Greek lasagna" to "classic lasagna").

Similarly, participants sometimes struggled with documents that were too general for their needs. Here, participants did not know how to apply the information to their specific situation. Systems could also enable searchers to request documents about related procedures that satisfy specific criteria. For example, upon encountering a document on "How to Connect a Laptop to a Monitor", a searcher might request related documents that are relevant to "Microsoft Surface Pro 3" and "Dell 2007FP".

Query-by-Example with Complexity Adjustment: Participants also struggled with procedural documents that were too complex. Another query-by-example tool could enable searchers to issue a complex document to request documents with similar procedures intended for novices.

Integration with Chat-Based AI tools: For many of the implications discussed in this section, we see opportunities for chat-based and generative AI tools to assist users. For example, large-language models (LLMs) could provide back-end support in alleviating terminology issues by helping users identify synonymous and unnecessary terms. LLMs could also help compare and summarize additional documents to help users understand rationales, details, and alternatives associated with a particular procedural step. Generative AI systems could help support dealing with constraints and adjustment of the specificity of results by allowing users to follow-up with a query that requests more specific or general results. In addition, chat-based and generative AI tools show great promise in supporting information searches, especially in helping users with unfamiliar tasks [8]. Procedural search tasks often involve gaining an overview of a domain and then understanding a sequence of steps. Current Generative AI tools (e.g. ChatGPT4) provide summaries of topics and processes that could be helpful to users in understanding a domain. Generative AI tools are also able to present information in sequences/steps and to break-down topics into sub-components. These are all important aspects of procedural search tasks that we observed in our study. We believe that generative AI could play an important role in supporting these user needs.

7 CAVEATS AND LIMITATIONS

One possible limitation of our study is that our participant pool was highly skewed toward females ($M = 5$, $F = 31$). It is unclear whether this imbalance influenced our results. For example, as part of RQ2, we considered the effects of the task's product and goal on perceptions of knowledge gained during the search task and found no significant effects. Studies in education, have found that females tend to underestimate their performance on a learning assessment [25]. Therefore, perhaps our task manipulations may have had an effect with a more balanced participant pool. The effects of this imbalance in our participant pool is an open question for future work.

8 CONCLUSION

We reported on a user study ($N = 36$) aimed at investigating how people search for procedural knowledge. Procedural search tasks were manipulated along two dimensions: product and goal. In RQ1-RQ3, we investigated the effects of our task manipulation on participants' pre-task perceptions, post-task perceptions, and search behaviors. In RQ4, we performed a qualitative analysis of 18 participants' search sessions (i.e., think-aloud comments and screen activities). This analysis focused on three main themes: relevance criteria, information types sought and used, and challenges. Our results for RQ1-RQ3 found that our manipulation of the task product had much stronger effects than our manipulation of the task goal. With respect to RQ1-RQ2, our results suggest that decision and intellectual tasks were more complex than factual tasks. Factual tasks were perceived to have clear objectives and *not* require subjective information. Decision tasks were also perceived to have clear objective but require the greatest amount of subjective information to aid in decision making. For intellectual tasks, participants reported the lowest levels of prior knowledge and the greatest need for background information. Intellectual tasks were also perceived to have unclear objectives

and require *some* amount of subjective information (i.e., more than factual and less than intellectual tasks). With respect to RQ3, intellectual tasks (arguably the most complex based on participants' perceptions) required fewer queries and clicks. This result is somewhat at odds with prior work, which has found that complex tasks require more SERP-level activity. A possible explanation is that intellectual tasks required less searching and more brainstorming and navigation. With respect to RQ4, our results found that participants used a wide range of criteria when judging relevance. Additionally, participants sought many different types of information and experienced different types of challenges. Based on our results, we have proposed different tools and interface features that may support users during procedural search tasks.

9 ACKNOWLEDGEMENTS

This material is based upon work supported in whole or in part with funding from the Department of Defense. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

REFERENCES

[1] Eyob N. Alemu and Jianbin Huang. 2020. HealthAid: Extracting domain targeted high precision procedural knowledge from on-line communities. *Information Processing & Management* 57, 6 (2020). https://doi.org/10.1016/j.ipm.2020.102299

[2] Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, complete edition.*

[3] Peter Bailey and Li Jiang. 2012. User task understanding: a web search engine perspective. (2012). https://www.microsoft.com/en-us/research/publication/user-task-understanding-a-web-search-engine-perspective/ Presentation delivered at the NII Shonan: Whole-Session Evaluation of Interactive Information Retrieval Systems workshop. 8-11 October 2012, Shonan, Japan.

[4] Panos Balatsoukas and Ian Ruthven. 2012. An eye-tracking approach to the analysis of relevance judgments on the Web: The case of Google search engine. *Journal of the American Society for Information Science and technology* 63, 9 (2012), 1728–1746.

[5] Carol L Barry. 1994. User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science* 45, 3 (1994), 149–159.

[6] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information Processing and Management* 31, 2 (1995), 191–213.

[7] Donald J. Campbell. 1988. Task Complexity: A Review and Analysis. *The Academy of Management Review* 13, 1 (1988), 40–52.

[8] Robert Capra and Jaime Arguello. 2023. How does AI chat change search behaviors? *arXiv* (2023). https://arxiv.org/abs/2307.03826

[9] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. 2015. Differences in the Use of Search Assistance for Tasks of Varying Complexity. Association for Computing Machinery, New York, NY, USA, 23–32.

[10] M Ariel Cascio, Eunlye Lee, Nicole Vaudrin, and Darcy A Freedman. 2019. A team-based approach to open coding: Considerations for creating intercoder consensus. *Field Methods* 31, 2 (2019), 116–130.

[11] Bogeum Choi and Jaime Arguello. 2020. A Qualitative Analysis of the Effects of Task Complexity on the Functional Role of Information. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 328–332.

[12] Bogeum Choi, Jaime Arguello, and Robert Capra. 2023. Understanding Procedural Search Tasks "in the Wild" *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, to appear.

[13] Bogeum Choi, Sarah Casteel, Robert Capra, and Jaime Arguello. 2022. Procedural Knowledge Search by Intelligence Analysts *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 169–179.

[14] Bogeum Choi, Austin Ward, Yuan Li, Jaime Arguello, and Robert Capra. 2019. The Effects of Task Complexity on the Use of Different Types of Information in a Search Assistance Tool. *ACM Trans. Inf. Syst.* 38, 1, Article 9 (dec 2019), 28 pages. https://doi.org/10.1145/3371707

[15] Cuong Xuan Chu, Niket Tandon, and Gerhard Weikum. 2017. Distilling Task Knowledge from How-To Communities. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, 805–814. https://doi.org/10.1145/3038912.3052715

[16] Michael J. Cole, Jacek Gwizdka, Chang Liu, Ralf Bierig, Nicholas J. Belkin, and Xiangmin Zhang. 2011. Task and user effects on reading patterns in information search. *Interacting with Computers* 23, 4 (05 2011), 346–362. https://doi.org/10.1016/j.intcom.2011.04.007

[17] Abe Crystal and Jane Greenberg. 2006. Relevance criteria identified by health information users during Web searches. *Journal of the American Society for Information Science and Technology* 57, 10 (2006), 1368–1382.

[18] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of within-Session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/2556195.2556217

[19] Elsa Eiriksdottir and Richard Catrambone. 2011. Procedural instructions, principles, and examples: How to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human factors* 53, 6 (2011), 749–770.

[20] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.

[21] Bernhard Ertl. 2009. Conceptual and Procedural Knowledge Construction in Computer Supported Collaborative Learning. In *Proceedings of the 9th International Conference on Computer Supported Collaborative Learning (CSCL'09)*. International Society of the Learning Sciences, 137–141.

[22] Luanne Freund, Elaine G. Toms, and Julie Waterhouse. 2005. Modeling the information behaviour of software engineers using a work - task framework. *Proceedings of the American Society for Information Science and Technology* (2005).

[23] Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. "What Can I Cook with These Ingredients?" - Understanding Cooking-Related Information Needs in Conversational Search. *ACM Transactions of Information Systems* 40, 4, Article 81 (2022).

[24] M.P. Georgeff and A.L. Lansky. 1986. Procedural knowledge. *Proc. IEEE* 74, 10 (1986), 1383–1398. https://doi.org/10.1109/PROC.1986.13639

[25] Sara M González-Betancor, Alicia Bolívar-Cruz, and Domingo Verano-Tacoronte. 2019. Self-assessment accuracy in higher education: The influence of gender and performance of university students. *Active Learning in Higher Education* 20, 2 (July 2019), 101–114. https://doi.org/10.1177/1469787417735604 Publisher: SAGE Publications.

[26] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 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 (ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 101–110.

[27] Helvi Kyngäs. 2020. Inductive content analysis. *The application of content analysis in nursing science research* (2020), 13–21.

[28] Jo-Anne LeFevre and Peter Dixon. 1986. Do written instructions need examples? *Cognition and Instruction* 3, 1 (1986), 1–30.

[29] Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing Management* 44, 6 (2008), 1822–1837.

[30] Jingjing Liu, Michael J. Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J. Belkin, Jun Zhang, and Xiangmin Zhang. 2010. Search Behaviors in Different Task Types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries (JCDL '10)*. Association for Computing Machinery, New York, NY, USA, 69–78.

[31] Jingjing Liu, Chang Liu, and Nicholas Belkin. 2013. Examining the Effects of Task Topic Familiarity on Searchers' Behaviors in Different Task Types. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries (ASIST '13)*. American Society for Information Science, USA.

[32] Irene C Michas and Dianne C Berry. 2000. Learning a procedural task: effectiveness of multimedia presentations. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 14, 6 (2000), 555–575.

[33] Dena Mujtaba and Nihar Mahapatra. 2019. Recent Trends in Natural Language Understanding for Procedural Knowledge. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. 420–424. https://doi.org/10.1109/CSCI49370.2019.00082

[34] Georg Pardi, Yvonne Kammerer, and Peter Gerjets. 2019. Search and Justification Behavior During Multimedia Web Search for Procedural Knowledge. In *Companion Publication of the 10th ACM Conference on Web Science (WebSci '19)*. ACM, New York, NY, USA, 17–20.

[35] Hogun Park and Hamid Reza Motahari Nezhad. 2018. Learning Procedures from Text: Codifying How-to Procedures in Deep Neural Networks. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, 351–358. https://doi.org/10.1145/3184558.3186347

[36] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Sumio Fujita, Akira Tajima, and Katsumi Tanaka. 2017. Mining Alternative Actions from Community Qamp;A Corpus for Task-Oriented Web Search. In *Proceedings of the International Conference on Web Intelligence (WI '17)*. ACM, New York, NY, USA, 607–614. https://doi.org/10.1145/3106426.3106461

[37] Victoria Reyes, Elizabeth Bogumil, and Levin Elias Welch. 2021. The living codebook: Documenting the process of qualitative data analysis. *Sociological Methods & Research* (2021), 0049124120986185.

[38] Reijo Savolainen and Jarkko Kari. 2006. User-defined relevance criteria in web searching. *Journal of Documentation* (2006).

[39] Michael Schneider and Elsbeth Stern. 2010. The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental psychology* 46, 1 (2010), 178.

[40] Pol Schumacher, Mirjam Minor, Kirstin Walter, and Ralph Bergmann. 2012. Extraction of Procedural Knowledge from the Web: A Comparison of Two Workflow Extraction Approaches. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 739–747. https://doi.org/10.1145/2187980.2188194

[41] Eliot R Smith. 1994. Procedural knowledge and processing strategies in social cognition. *Handbook of social cognition* 2 (1994), 99–152.

[42] Arthur Taylor. 2012. User relevance criteria choices and the information search process. *Information Processing & Management* 48, 1 (2012), 136–153.

[43] Timon ten Berge and Rene van Hezewijk. 1999. Procedural and Declarative Knowledge: An Evolutionary Perspective. *Theory & Psychology* 9, 5 (1999), 605–624. https://doi.org/10.1177/0959354399095002

[44] Timon Ten Berge and René Van Hezewijk. 1999. Procedural and declarative knowledge: An evolutionary perspective. *Theory & Psychology* 9, 5 (1999), 605–624.

[45] Kelsey Urgo and Jaime Arguello. 2022. Understanding the "Pathway" Towards a Searcher's Learning Objective. *ACM Transactions of Information Systems* 40, 4 (2022).

[46] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2020. The Effects of Learning Objectives on Searchers' Perceptions and Behaviors *(ICTIR '20)*. ACM, New York, NY, USA, 77–84. https://doi.org/10.1145/3409256.3409815

[47] Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. 2015. What Users Ask a Search Engine: Analyzing One Billion Russian Question Queries. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 1571–1580. https://doi.org/10.1145/2806416.2806457

[48] Ingmar Weber, Antti Ukkonen, and Aris Gionis. 2012. Answers, Not Links: Extracting Tips from Yahoo! Answers to Address How-to Web Queries. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 613–622. https://doi.org/10.1145/2124295.2124369

[49] Barbara Wildemuth, Elaine G. Toms, and Luanne Freund. 2014. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation* 70, 6 (Oct. 2014), 1118–1140.

[50] Monte B Wynder and Peter F Luckett. 1999. The effects of understanding rules and a worked example on the acquisition of procedural knowledge and task performance. *Accounting & finance* 39, 2 (1999), 177–203.

[51] Yunjie Xu and Zhiwei Chen. 2006. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 961–973.

[52] Zi Yang and Eric Nyberg. 2015. Leveraging Procedural Knowledge for Task-Oriented Search *(SIGIR '15)*. ACM, New York, NY, USA, 513–522. https://doi.org/10.1145/2766462.2767744

[53] Ziqi Zhang, Philip Webster, Victoria Uren, Andrea Varga, and Fabio Ciravegna. 2012. Automatically Extracting Procedural Knowledge from Instructional Texts using Natural Language Processing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 520–527.