The Effects of Task Complexity on the Use of Different Types of Information in a Search Assistance Tool

BOGEUM CHOI, AUSTIN WARD, YUAN LI, JAIME ARGUELLO, and ROBERT CAPRA,
University of North Carolina at Chapel Hill

In interactive information retrieval, an important research question is: How do task characteristics influence users' needs and behaviors? We report on a laboratory study (N = 32) that investigated the effects of task complexity on the types of information used by participants while searching. Participants completed tasks of four complexity levels and had access to four different types of information provided through a search-assistance tool referred to as the InfoBoxes (IB). The IB tool presented the following types of task-related information (*info-types*) on different tabs: (1) facts, (2) concepts, (3) opinions, and (4) insights. Facts (and opinions) were defined as objective (and subjective) statements relevant to the task. Concepts were defined as important ideas, principles, or entities related to the task. Insights were defined as tips or advice about the task. The study investigated six research questions that considered the effects of task complexity on: (RQ1) participants' pre-/post-task perceptions about useful info-types; (RQ2) use of different info-types during the task; (RQ3) motivations for engaging with the IB; (RQ4) gains from using it; (RQ5) the search stage participants were in while engaging with the IB; and (RQ6) motivations for sometimes avoiding the IB. Our results suggest that task complexity influenced all six types of outcomes. We discuss implications of our results for designing search assistance tools and systems that favor certain types of content based on task characteristics.

CCS Concepts: • **Information systems → Users and interactive retrieval**; **Search interfaces**;

Additional Key Words and Phrases: Cognitive task complexity, search assistance, search behaviors

ACM Reference format:
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 (December 2019), 28 pages.
https://doi.org/10.1145/3371707

1 INTRODUCTION

Current search engines are effective in helping users complete simple search tasks such as fact-finding and information lookup tasks. However, they are less effective in helping users with *complex* search tasks that may involve understanding novel concepts, making sense of conflicting information, and tasks that involve creative problem-solving [16]. One approach to help users with

All authors contributed equally to this article.
This work was supported by NSF grant IIS-1718295. Any opinions, findings, conclusions, and recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the sponsor.
Authors' addresses: B. Choi, A. Ward, Y. Li, 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, austinrw, yuan_li, jarguello, rcapra}@unc.edu.

1046-8188/2019/12-ART9 $15.00
https://doi.org/10.1145/3371707

complex search tasks has been to develop *search assistance* tools that are *complementary* to the main search interface. For example, prior research has designed and evaluated tools to help users formulate better queries [30, 43], avoid spelling errors [19], discover advanced query operators and search functions [39], organize information and take notes [18], and learn from the search "paths" followed by previous users who completed a related search task [12, 45].

Designing effective search assistance tools requires understanding how specific task characteristics influence the types of information users need and the challenges they face. In this article, we focus on task *complexity*, defined as an objective task characteristic (independent of the task doer) [33]. Task complexity has been viewed and manipulated from different perspectives (e.g., References [10, 11, 28]). Wildemuth et al. [48] conducted an extensive review and concluded that prior studies have viewed complex tasks as involving: (1) multiple steps; (2) multiple concepts and/or concept-types; and (3) greater uncertainty about the task's requirements, form of the solution, and processes involved.

In this work, we viewed task complexity from the perspective of *cognitive* complexity, which relates to the types (and variety) of mental processes required by the task. The idea behind cognitive complexity originated from educational research. Anderson and Krathwohl [3] proposed a two-dimensional taxonomy for characterizing learning objectives. Jansen et al. [26] (and later Kelly et al. [28]) adopted the cognitive dimension of this taxonomy to develop and study information-seeking tasks of varying complexity. To illustrate, a simple task might require verifying a specific piece of information, a moderately complex task might require understanding the trade-offs between different alternatives, and a highly complex task might require creating a novel solution to a problem. Several studies have investigated how a task's cognitive complexity may affect searchers' perceptions and behaviors. Specifically, studies have found that cognitively complex tasks are perceived to be more difficult [4, 7, 12, 23, 28, 49] and require more search activity [4, 12, 23, 28, 49].

We report on a laboratory study ($N = 32$) that investigated the following general question: How does the cognitive complexity of a task influence the *types of information* used by searchers to complete the task? Specifically, we investigate this question within the context of search assistance. Participants in the study completed four search tasks of varying levels of cognitive complexity (a within-subjects design). Each task asked participants to search for information and create a written response. To gather information, participants were given access to a search system that included a standard web search engine and a *peripheral* search assistance tool referred to as the InfoBoxes (or IB). The InfoBoxes tool was positioned to the left of the search results and displayed *four* different types of task-related information (or *info-types*) in different tabs: facts, concepts, opinions, and insights. Facts were defined as objective and verifiable statements; concepts were defined as noun-phrases representing important ideas, principles, or entities related to the task domain; opinions were defined as subjective statements representing points of view, judgments, and experiences; and insights were defined as helpful advice or tips regarding the task and the domain. As described in Section 3.3, our motivation for focusing on facts, concepts, opinions, and insights originated from one of our previous studies in which participants were assigned search tasks and asked to take hand-written notes that would help them resume the task later or help someone else [15]. A qualitative analysis of participants' notes found that participants had a tendency to include: factual statements, opinionated statements, important concepts/definitions, and task-specific advice and tips.

In this study, we were interested in investigating a "best-case" scenario in which the information in the IB was of high quality. Thus, for each task, the information in the IB was manually curated using data from a preliminary study. To gain insights about how participants used the info-types in the IB, the main study used a think-aloud protocol with a stimulated recall interview [17]. Participants were asked to think aloud while they searched. Then, after completing all four tasks,

participants observed video recordings of specific times when they engaged with the IB during each session and answered questions about their motivations for engaging with the IB and gains obtained.

Our study investigated six research questions:

- **RQ1:** What is the effect of task complexity on participants' *pre-task* and *post-task perceptions* about needing specific types of information to complete the task?
- **RQ2:** What is the effect of task complexity on participants' *use* of different info-types in the InfoBox search assistance tool during the task? RQ2 was studied from two perspectives: (1) based on logged interaction data and (2) based on observations of participants' use of different info-types in the IB.
- **RQ3:** What is the effect of task complexity on participants' *motivations for engaging* with the IB?
- **RQ4:** What is the effect of task complexity on participants' *gains obtained* from engaging with the IB?
- **RQ5:** What is the effect of task complexity on the *search stages* participants were in when engaging with the IB?
- **RQ6:** What is the effect of task complexity on the participants' *motivations for purposely avoiding* the IB?

This article makes the following three important contributions: First, our results for RQ1–RQ4 suggest that task complexity influences the types of information useful for a task. These results have implications for designing search systems—including peripheral support tools such as the InfoBoxes (IB) tool—that favor specific types of information based on task characteristics. Second, our results for RQ5 suggest that task complexity influences the stages in which users may engage with a support tool such as the IB. These results have implications for designing dynamic help tools that must decide *when* and *how* to intervene to support users. Finally, our results for RQ6 suggest that task complexity impacts the reasons why users may choose to avoid help tools such as the IB. These results have implications for reliably interpreting user interactions (or lack of engagement) with help tools. Additionally, our RQ6 results point to future directions for designing support tools that aggregate different types of information based on task (and sub-task) characteristics.

2 RELATED WORK

Our research builds on several areas of prior work. First, participants in the study had access to different types of task-related information in an auxiliary tool (i.e., the InfoBoxes) that was complementary to the search results. Thus, our work builds on prior research on help systems in IR. Furthermore, we extend prior research on the effects of task complexity on search behaviors, outcomes, and the types of information required to complete the task.

2.1 Search Assistance in IR

Designing IR help systems requires understanding the different types of challenges faced by searchers and their motivations for ignoring help systems. In a large-scale study with digital library searchers, Xie and Cool [50] found that searchers experience difficulty with seven general processes: (1) getting started, (2) identifying relevant resources, (3) navigating a resource, (4) constructing queries, (5) refining queries, (6) monitoring the search process, and (7) evaluating results. Additionally, a wide range of factors was found to contribute to these challenges, including characteristics of the user (e.g., domain knowledge, search experience), task (e.g., task complexity), and system (e.g., topical coverage of the collection).

While search assistance tools have the potential to support users, there are also challenges in designing effective help systems. Prior work suggests that users avoid help systems for several reasons, including the fear of unproductive help-seeking, the cost of disengaging with the primary task, and the refusal to admit defeat [20]. Jansen and McNeese [26] evaluated a system that provided dynamic assistance in the form of query suggestions (i.e., related queries, synonyms, and suggested Boolean operators). Participants avoided the assistance 71% of the time it was offered. However, when participants *noticed* the assistance, they used it. This result suggests that users may not notice search assistance tools when deeply engaged with the task (i.e., cognitive blindness). Finally, Huang and Xie [24] found that participants' learning styles impacted their use of help features in a digital library system (e.g., FAQs, widgets for filtering/re-sorting search results, live chat with a librarian). Reflective learners made less use of help features than active learners.

2.2 Task Complexity

People search for information to complete a specific task. Vakkari [44] defined *task* as "an activity performed to accomplish a goal" (p. 416). Byström and Hansen [9] proposed that information-intensive tasks can be defined at three levels of granularity: a *search task* is a subtask of an *information-seeking task*, and an information-seeking task is a subtask of a *work task*—the higher-level task that triggered the need for information. A large body of research has characterized search tasks along different dimensions. Li and Belkin [33] conducted an extensive review and proposed a unifying framework that characterizes search tasks along two dimensions: (1) generic facets (e.g., self-motivated vs. assigned) and (2) common attributes, including subjective attributes (e.g., perceived difficulty) and objective attributes (e.g., complexity).

Task complexity is an objective task attribute that has been defined and manipulated from different perspectives [48]. Campbell [11] proposed that complex tasks involve: (1) multiple outcomes, (2) multiple paths to the outcomes, (3) high uncertainty about paths, and (4) high interdependence between paths. Prior work has also studied task complexity from the perspective of *a priori* determinability—the degree of uncertainty regarding the task requirements, processes involved, and the form of the solution [7, 8, 10]. Byström and Järvelin [10] used the principle of *a priori* determinability to categorize tasks into five complexity levels. To illustrate, *automatic information processing tasks* (the simplest category) are completely determinable with respect to their requirements, processes involved, and the form of the solution. Conversely, *genuine decision tasks* (the most complex) are completely unexpected, new, and unstructured—neither the requirements, processes, and form of the solution can be characterized in advance. Bell and Ruthven [7] used *a priori* determinability to design tasks of three complexity levels to use in a laboratory study. The simplest tasks were designed to be completely determinable, while the most complex were designed to have more uncertainty regarding three processes: (1) determining the required information, (2) finding relevant sources, and (3) recognizing relevant information.

More closely related to our work, task complexity has also been studied from the perspective of *cognitive* complexity [12, 23, 25, 28, 49], which relates to the types (and variety) of mental processes required by the task. Prior studies have used Anderson and Krathwohl's taxonomy of learning objectives [3] to design search tasks of varying complexity levels [12, 23, 25, 28, 49].

Past studies have also investigated how task complexity impacts search behaviors and outcomes. Results have found that complex tasks are associated with greater levels of expected (pre-task) difficulty [12, 23, 28, 49], greater levels of experienced (post-task) difficulty [4, 7, 12, 23, 28, 49], and greater levels of search activity [4, 12, 23, 28, 49]. Finally, Capra et al. [12] investigated the effects of task complexity on participants' use of a search assistance tool that displayed search paths from other searchers who completed the same task. Results found two important trends. First, complex tasks had more use of the search assistance tool. Second, task complexity impacted

participants' *motivations* for engaging with the tool—for simple tasks, participants used the tool to confirm information found on their own, and for complex tasks, participants used the tool to change search strategies.

2.3 Task Complexity and Users' Information Needs

A large body of research has focused on understanding the many different criteria that influence *relevance* from a user's perspective (see Saracevic [42] for an extensive review). Information, of course, can be characterized along different dimensions (e.g., general vs. specific, objective vs. subjective, declarative vs. procedural, nontechnical vs. technical). Thus, research on *relevance criteria* has aimed at understanding how contextual factors (e.g., attributes of the user, situation, or task) influence the *characteristics* of relevant versus non-relevant information. Here, we focus on research related to *task complexity* and the types of information users seek.

Prior studies by Byström and colleagues [8, 10] have investigated the effects of task complexity on the types of information people seek and how they use it. These studies analyzed genuine tasks completed by workers in a professional setting and primarily focused on the *functional role* of information during task completion. In other words, these studies focused on how the information was used or for what purpose. Byström and Järvelin [10] considered the effects of task complexity, viewed through the lens of *a priori* determinability, on the use of three types of information: (1) *problem information* (PI)—information that describes the structure, properties, and requirements of the task; (2) *problem-solving information* (PSI)—information on how to approach/solve the task; and (3) *domain information*—facts, concepts, laws, and theories in the task domain. To further explain, PI helps the task doer understand the task and PSI helps the task doer solve the task (i.e., determine what PI and DI to use and how). Results found differences in the functional role of information used during simple versus complex tasks. Simple tasks required mostly PI and fact-oriented sources to support rule-based processing. Conversely, complex tasks required all three types of information. Particularly, complex tasks required more PSI to support problem-solving. Additionally, complex tasks involved greater use of human experts as sources of information.

In a follow-up (and similar) study, Byström [8] studied the effects of task complexity (i.e., *a priori* determinability) on the use of three different types of information: (1) *task information* (TI)—information that helps the task doer define the task; (2) *task-solving information* (TSI)—information that helps the task doer solve the task; and (3) *domain information* (DI)—general purpose information about the task domain. TI was considered to be mostly factual, while TSI and DI were considered to be either factual or interpretative/subjective. Consistent with Byström and Järvelin [10], simple tasks required mostly TI, while complex tasks required all three types of information (particularly more TSI). Similarly, complex tasks had greater use of human experts as sources of information.

In the above studies, information types were characterized based on how the information was used during the task. Conversely, in our study, information types were defined more strongly based on inherent characteristics (e.g., factual versus opinionated statements). A few studies have also considered the effects of task complexity on inherent characteristics of information used during task completion. Saastamoinen et al. [41] investigated the effects of task complexity on participants' expected use of three types of information: (1) *isolated facts*; (2) *information aggregates*—a group of synthesized facts; and (3) *known-items*—an information object such as a book or report. Participants expected isolated facts to be more useful during simple tasks and information aggregates to be more useful during complex tasks. In a lab study, Zhang [51] had participants complete simple versus complex tasks in the health domain, and analyzed the types of information sources *recalled* by participants after the task. For simple tasks, participants had a tendency to recall sources containing factual and shallow information (e.g., fact sheets, FAQs, summaries). Conversely, for

complex tasks, participants had a tendency to recall information sources containing debatable, subjective, and in-depth information (e.g., scholarly articles and clinical trial reports).

3 METHODS

To investigate our six research questions, we conducted a laboratory study with 32 participants (41% female). Participants were recruited using a campus-wide opt-in mailing list of students and employees at our university. Participants consisted of 15 undergraduate students, 5 graduate students, and 12 non-student employees, and their age ranged from 18 to 58 (M = 30).

The study used a concurrent think-aloud protocol with a retrospective stimulated recall interview. Each participant completed four search tasks of varying levels of complexity: remember, understand, analyze, and create (Section 3.2). The task ordering was rotated among participants using a balanced Latin square. Each task required the participant to search for information and create a written response in a task-specific answer sheet. Participants had access to two tools to find information. First, they had access to a standard search interface that used the Bing API to retrieve web results. Second, the InfoBoxes tool (IB) was displayed to the right of the SERP (Section 3.3). The IB displayed information and links in four different tabs: facts, concepts, opinions, and insights.

3.1 Protocol

After reviewing and signing a consent form, participants first completed a demographics questionnaire. Next, participants were shown a video introducing the study, the search system, and the InfoBoxes (IB) tool. Our goal in this study was to investigate a "best-case" scenario in which the information presented in the IB was high quality. To this end, in the video, participants were told that the IB displayed different types of information (facts, concepts, opinions, insights) found by other people for the same task.

Our study involved asking participants to think aloud while they worked on each task. To familiarize participants with the system and the think-aloud protocol, we asked participants to do a practice task. During the practice task, participants were encouraged to try the IB features but were told they were not required to use it.

Next, participants completed the four main tasks. We used a wide-screen monitor to display the search system and answer sheet side-by-side. The search system (search engine and IB) was displayed on the left side of the screen and the (initially blank) answer sheet was displayed as a Google Doc on the right side of the screen. Both the search system and answer sheet included the task description. Participants had 18 minutes per search task and were notified by the moderator when they had 5 minutes remaining. During pilot testing, we found that 18 minutes was sufficient time to complete each task. Participants in the study took an average of 11.58 minutes per search task (SD = 5.62).

Each task followed the same procedure. Participants were first given the task description and asked to read it aloud. Then, participants completed a pre-task questionnaire (Section 3.5). Next, participants were directed to start the first task using the search interface. During each search, the moderator prompted participants to think aloud if they fell silent. While think-aloud protocols have the potential to alter participants' cognitive processes, they are commonly used in HCI and UX studies [14]. Additionally, to avoid influencing participants, our prompts were simple reminders for participants to keep thinking aloud. After completing each task, participants completed a post-task questionnaire (Section 3.5).

We recorded each search session and participants' verbal think-aloud comments using Morae screen recording software. During each task, the moderator used Morae Observer to carefully monitor the participant's search and marked all points where the participant engaged with the IB based on their actions (e.g., clicks and mouse hovers) and/or think-aloud comments (i.e., verbal

mentions of engaging with the IB). These points of IB use were later used in the stimulated recall interview (Section 3.6).

After completing all four tasks, we asked if the participant needed a break and then began the retrospective stimulated recall interview. During the interview, the moderator revisited each search task, replayed the screen and audio recordings of the participant's *first* and *last* IB use during the task (if any), and asked a set of scripted questions about the participant's use of the IB (e.g., motivations, gains, search stage). In this way, participants' concurrent think-aloud comments were *not* used as a primary data source. Instead, they were used during the retrospective interview to *help participants remember* details about particular instances when they used the IB. Using this stimulated recall approach, we were able to ask specific questions during the stimulated recall interview (after all four tasks) while reducing the risk of influencing participants during the current task or subsequent tasks. Additionally, for each task, the moderator asked whether the participant at any point *purposely avoided* using the IB, and if so, why? Participants' responses to the stimulated recall questions were later analyzed using qualitative techniques to address research questions RQ3–RQ6 (Section 3.7). To address RQ2, during the qualitative coding process, we also noted which info-types participants engaged with during their first/last IB use of each task. The study session lasted about 1.5 hours, and participants were given US$40 for participating in the study.

3.2 Task Complexity Manipulation

All participants completed the same four search tasks, which had different levels of *cognitive* complexity. We used search tasks adapted from Kelly et al. [28] and Capra et al. [13] that were associated with four levels of cognitive complexity: remember, understand, analyze, and create. These complexity levels were inspired by Anderson and Krathwohl's taxonomy of learning objectives [3], originally developed to help educators define learning outcomes, instructional materials, and learning assessment strategies. A *remember* task requires remembering, identifying, or verifying a specific piece of information. An *understand* task requires constructing meaning through summarizing and explaining. An *analyze* task requires decomposing material into constituent parts and determining how the parts relate to each other. A *create* task requires developing a new solution to a problem. Our four search tasks were as follows:

- **Remember:** You recently watched a documentary about people living with HIV in the U.S. You thought the disease was nearly eradicated and are curious to know more about the prevalence of HIV. Specifically, how many people in the U.S. are living with HIV?
- **Understand:** You recently became acquainted with one of the farmers at the local farmers' market. One day, over lunch, he was on a rant about how people are ruining the soil. He was clearly upset, so you're interested in finding out more. What are some human activities that degrade soil fertility?
- **Analyze:** You are planning an extended hiking trip and will not be able to carry all the water you will need. There will be streams near where you are hiking, but a friend said that you might get sick from drinking water directly from the streams. What are some different methods to purify stream water for drinking during long hiking trips and how do they differ?
- **Create:** After the NASCAR season opened this year, your niece became interested in soapbox derby racing. Since her parents are both really busy, you've agreed to help her. The first step is to figure out how to build a soapbox derby car. Identify some basic designs that you might use and create a basic plan for constructing the car.

During each task, participants were asked to search for information and create a written response. Each task had its own structured answer sheet (i.e., Google Doc). The answer sheets were designed

to be an integral part of the assigned tasks. Our goal for using these answer sheets was to reduce variability in how participants interpreted the goals of each task. The remember task (HIV) asked participants to find a specific piece of information. Thus, the answer sheet for the remember task included three boxes for participants to: (1) enter the most confident answer, (2) provide a justification, and (3) enter any alternative answers found. The understand task (soil degradation) required participants to identify and summarize a list of items. Thus, the answer sheet for the understand task included a two-column table for participants to enter up to 12 soil degradation activities (first column) and a brief explanation for each (second column). The analyze task (water purification) required participants to identify and compare/contrast different items (or alternatives) along different dimensions (or attributes). Thus, the answer sheet for the analyze task contained a 7×6 table for participants to populate based on identified water purification methods (rows) and attributes (columns). Finally, the create task (soapbox derby) required participants to propose a plan/design. Thus, the answer sheet for the create task included one large box for participants to: (1) discuss the different plans/designs considered, (2) describe the chosen plan/design, and (3) provide a justification.

3.3 Search System, InfoBoxes, and Info-types

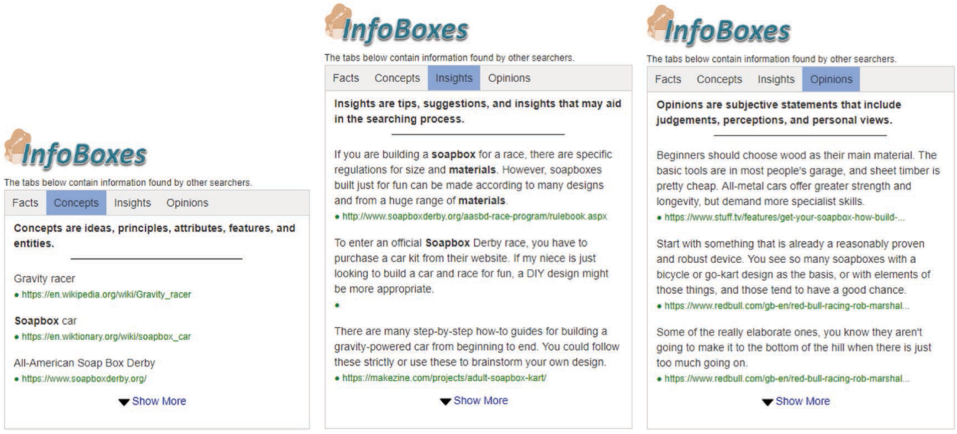
To complete the tasks, participants interacted with a search system that provided: (1) a standard search interface on the left and (2) the InfoBoxes (IB) tool on the right (Figure 1). The standard search interface used the Bing API to return web results, had pagination controls at the bottom, and displayed 10 results per page. The system logged all user interactions (clicks, mouse events, scrolls, etc.). As shown in Figure 1, the IB displayed four different types of information (info-types) related to the task: facts, concepts, opinions, and insights. In Section 3.4, we describe how the IB was populated with task-specific information gathered by human annotators.

The idea of focusing on facts, concepts, opinions, and insights originated from a previous study [15] that investigated the types of information people use to complete tasks. Qualitative methods were used to identify different types of information that participants ($N = 24$) included in notes they were asked to take during two search tasks. Among the types of information included were: (1) factual details about the task domain (i.e., analogous to our facts), (2) important concepts and definitions (i.e., analogous to our concepts), and (3) tips and advice about the task domain and approaches for solving the task (i.e., analogous to our opinions and insights).

To further explain, we selected info-types to investigate in the current study based on several desiderata. First, we selected info-types based on inherent characteristics of the information (e.g., facts, opinions) rather than the functional role the information might play during the task (e.g., problem-solving information, as in Byström et al. [8, 10]). The reason for this was that we wanted to explore info-types that could be potentially identified and extracted from documents based on current IR techniques (e.g., classifying objective vs. subjective statements, extracting insights from Q&A forums). Second, we selected info-types that: (1) could play different functional roles (e.g., be used as domain information, problem-solving information) and (2) that would be relevant to tasks of different levels of cognitive complexity (e.g., remember, understand, evaluate, create). Finally, we wanted to explore info-types that were grounded in prior work. As previously mentioned, in our previous study [15], we identified factual details, important concepts, and tips and advice as important information types.

In the current study, the IB was displayed to participants after issuing the first query of the search session and remained visible on the SERP throughout the rest of the session. In the IB, the items for each info-type were displayed on different tabs and the left-most tab was open by default (as shown in Figure 1(a)). The order of tabs was rotated using a balanced Latin square, and each order (out of four) was repeated eight times across our 32 participants (i.e., each participant

(a) Search Interface and InfoBoxes with Facts Tab Open



(b) Concepts Tab. (c) Insights Tab. (d) Opinions Tab.

Fig. 1. Search interface and InfoBoxes.

saw the same order across all four tasks). Each info-type tab (Figures 1(a)–1(d)) displayed eight items vertically. The top-three items were displayed by default and all eight items were displayed if the participant clicked a "show more" link. As described in Section 3.4, facts, concepts, and opinions were originally extracted verbatim from pages by human annotators. In contrast, insights were written by annotators and did not necessarily originate from one specific page. Thus, all facts, concepts, and opinions displayed in the IB had clickable links that participants could use to navigate to the page from which the item was extracted. In contrast, some insights had clickable links and others did not.

3.4 Preliminary Data Collection

For each task, the IB was populated using data from an initial data collection effort. Three Ph.D. students in information science (referred to as annotators) were recruited for this purpose. The annotators were not authors on this article, nor were informed of our research questions.

The preliminary data collection proceeded in two phases. During the first phase, each annotator was given all four tasks and asked to search the web to gather at least 10 items for each info-type. Annotators were asked to identify items that would help someone else complete the same task. Facts were defined as "objective and verifiable statements (rather than opinions) that may include background information about the task, statistics, numbers, and descriptions." Concepts were defined as "noun-phrases representing important ideas, principles, attributes, features, or entities (e.g., people, places) related to the task domain." Opinions were defined as "subjective statements (rather than facts) that may include judgments, perceptions, or personal views." Finally, insights were defined as "tips, suggestions, or recommendations that you think would be useful to someone working on this task." Annotators were instructed that facts, concepts, and opinions should be extracted verbatim from web pages and were asked to record the page's URL. In contrast, annotators were instructed that the insights could either be extracted from pages (and if so, to record the URL) or written by themselves.

During the second phase, the goal was to determine *which* items to display in the IB for each task/info-type pair and their order. To this end, for each task/info-type pair, we first pooled together the items identified by the three annotators and removed duplicates. Then, we asked each annotator to rank the pooled items based on their relevance to the task. Finally, for each task/info-type, we merged the annotators' three rankings based on each item's average rank and selected the top-eight items to display in the IB. Table 1 in the Appendix shows examples of the top-eight facts, concepts, opinions, and insights related to the *create* task (soapbox derby).

With respect to this initial data collection, an important question is: To what extent did annotators agree on the quality of items? Treating "top-eight" and "other" as two categories, the Fleiss' kappa agreement between assessors was $\kappa_f = 0.287$, which is considered *fair* agreement [32]. Fleiss' kappa is the range $[-1, +1]$, with $\kappa_f = -1$ indicating perfect disagreement, $\kappa_f = +1$ indicating perfect agreement, and $\kappa_f = 0$ indicating random agreement (i.e., considering each assessor's own distribution). We believe that agreement was not higher, because annotators at this point were asked to rank items that were *already* of high quality. That is, annotators were asked to rank the union of top-10 items found by each annotator for a given task/info-type pair. Therefore, all items were presumably at least *topically* relevant. Interestingly, agreement was *higher* for simpler tasks (remember = 0.337, understand = 0.311, analyze = 0.278, and create = 0.178) and for more objective info-types (facts = 0.351, concepts = 0.322, opinions = 0.177, and insights = 0.254).[1]

3.5 Questionnaires

Participants completed a pre- and post-task questionnaire before/after each task. On both questionnaires, participants indicated their level of agreement with statements using a 7-point scale, from "strongly disagree" (1) to "strongly agree" (7). Both questionnaires are included in Tables 2 and 3 in the Appendix.

Pre-task: The pre-task questionnaire (Table 2) included 13 items to measure participants' perceptions and expectations about the task: (1) expected difficulty (4 items), (2) *a priori* determinability (3 items), (3) level of interest (1 item), and (4) level of prior domain knowledge (1 item). We created aggregate measures using the items associated with expected difficulty (Cronbach's α = .884) and determinability (α = .846). Additionally, we included 4 items to measure participants'

[1]In an extensive literature review on studies of relevance, Saracevic [42] states that assessor agreement on relevance judgments tends to hover around 30%. Importantly, this 30% value involves percent agreement, which is not corrected for random change agreement, and involves judgments made on documents/items that are relevant and non-relevant. One should expect agreement to be even lower when comparing items of similar quality. Consistent with this hypothesis, Arguello et al. [5] also found lower inter-annotator agreement on pairwise preference judgments between pairs of items of similar (versus dissimilar) quality.

expectations about needing specific info-types to address the task: (1) factual information, (2) conceptual information (e.g., terminology, ideas, principles), (3) subjective information (e.g., opinions, recommendations), and (4) insights (e.g., task-related advice, tips).

Post-task: The post-task questionnaire (Table 3) included 14 items to measure participants' perceptions of their experience completing the task: (1) experienced difficulty (4 items), (2) satisfaction (4 items), (3) interest increase (1 item), and (4) knowledge increase (1 item). Again, we created aggregate measures using the items for experienced difficulty (Cronbach's $\alpha = .889$) and satisfaction ($\alpha = .894$). Additionally, similar to the pre-task questionnaire, we included 4 items to measure participants' perceptions about needing specific types of information during the task: (1) factual information, (2) conceptual information, (3) subjective information, and (4) insights.

3.6 Stimulated Recall Interview

The stimulated recall interview was conducted after participants completed all four tasks. During the interview, the moderator revisited each of the four search sessions completed by the participant. For each task, the moderator played the recording (±10 seconds) of the participant's *first* and *last* IB use during the session (if any). To help participants recall details (e.g., motivations, gains) of each IB use, during each playback, participants could see the screen recording and hear their own think-aloud comments.

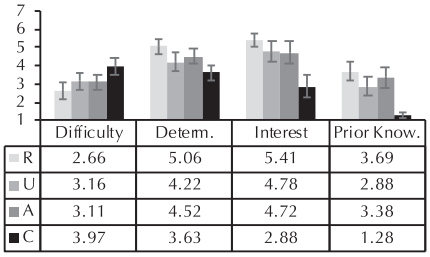
After each playback, participants were asked a set of questions about the specific IB use: (1) motivations for engaging with the IB, (2) gains obtained from the IB (if any), and (3) the search stage associated with the IB use. Additionally, for each task (including those without an observed IB use), participants were asked if there were times during the task where they purposely avoided the IB and if so, why? All questions were open-ended, except for the question about search stage. For stage, participants were asked "what they were doing at the time" and given the following choices: (1) initiation—getting an initial understanding of the task, (2) planning—strategizing about the information needed to complete the task, (3) pursuing—searching for a specific piece of information for the task, and (4) verifying—evaluating the accuracy, credibility, or completeness of information already found. We also included "other" as an option, but very few participants indicated other stages.

We chose to ask about only the *first* and *last* IB use for several reasons. First, we wanted to limit the experimental sessions to 1.5 hours and reduce the effects of fatigue during the retrospective interviews. To accomplish this, in retrospective interviews, it is common to use a sampling approach in which the moderator asks the participant about particular cases of interest. Second, we wanted to use a sampling approach that would ask each participant about a *similar* set of IB uses for each task. In this respect, the first and last instance of IB use were consistent, easy to identify, and helped eliminate moderator bias in selecting which IB uses to ask about. Finally, we expected the first and last IB uses for each task to be interesting "endpoints" that would help us gain insights about our research questions (i.e., motivations, gains, stage).

3.7 Qualitative Coding

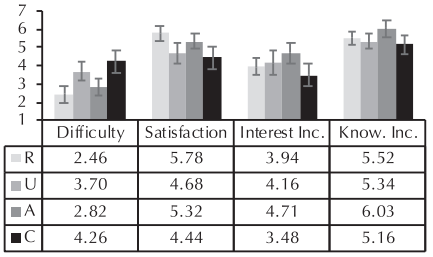
Participants' responses during the stimulated recall interviews were recorded and analyzed using qualitative techniques. Open-ended responses were coded along three dimensions (each with their own disjoint set of codes): (1) motivations for IB use, (2) gains obtained from the IB, and (3) motivations for avoiding the IB.

The coding scheme used in this study was *adapted* from one developed in a previous study. In Capra et al. [12], we investigated participants' use of a search assistance tool that displayed *search trails* from other participants who completed the same task. As in this study, we investigated



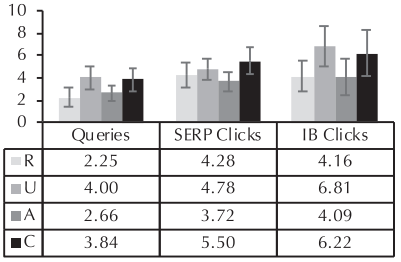
	Difficulty	Determ.	Interest	Prior Know.
R	2.66	5.06	5.41	3.69
U	3.16	4.22	4.78	2.88
A	3.11	4.52	4.72	3.38
C	3.97	3.63	2.88	1.28

(a) Pre-task Perceptions.



	Difficulty	Satisfaction	Interest Inc.	Know. Inc.
R	2.46	5.78	3.94	5.52
U	3.70	4.68	4.16	5.34
A	2.82	5.32	4.71	6.03
C	4.26	4.44	3.48	5.16

(b) Post-task Perceptions.



	Queries	SERP Clicks	IB Clicks
R	2.25	4.28	4.16
U	4.00	4.78	6.81
A	2.66	3.72	4.09
C	3.84	5.50	6.22

(c) Logged Interaction Measures.

Fig. 2. Effects of task complexity on participants' pre-/post-task perceptions and behaviors.

participants' motivations for engaging with the search trails, gains obtained, and motivations for avoiding them.

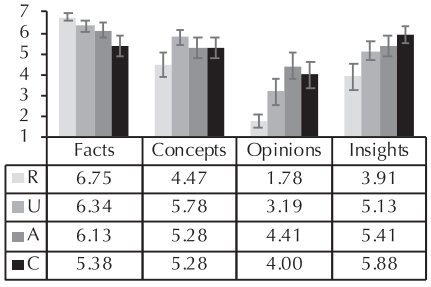
In the present study, the coding of participants' open-ended responses proceeded in three phases. During the first phase, the goal was to adapt the coding scheme from Capra et al. [12] for the current study. To this end, four of the authors independently coded the interview responses from two participants (eight sessions). Each author modified the coding scheme based on newly observed phenomena (i.e., added, dropped, and/or merged codes). Next, the four authors met to discuss their coding schemes and defined a preliminary coding scheme. During the second round of coding, the same four authors coded the interview responses from one new participant (four sessions). Again, the four authors met to discuss the coding scheme, make changes, and arrive at a final coding scheme. Finally, two of the authors (re-)coded all interview responses using the final coding scheme. During this final stage, each author coded responses from 16 participants (half and half). Then, the authors reviewed each other's annotations and resolved any inconsistencies. Ultimately, each dimension (motivations for use, non-use, and gains) had six to eight codes.

4 RESULTS

4.1 RQ0: Task Complexity Manipulation Check

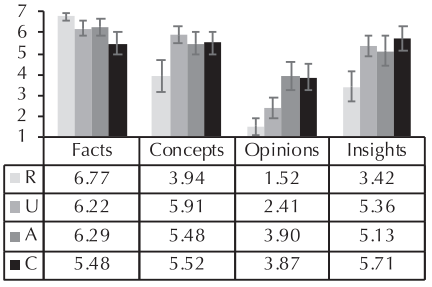
Before presenting our RQ1-RQ6 results, we report on the effects of task complexity on participants' pre-/post-task perceptions and search behaviors. We performed this analysis to verify whether complex tasks were more difficult and required more interaction.

Figures 2(a) and 2(b) show the means and 95% confidence intervals of participants' pre- and post-task responses according to task complexity. Participants' perceptions were measured using four pre-task factors (expected difficulty, determinability, prior knowledge, interest) and four post-task factors (experienced difficulty, satisfaction, knowledge increase, interest increase). To analyze the effects of task complexity on these measures, we conducted repeated-measures ANOVAs. In all our analyses, we used the *modified* Bonferroni-correction method in Keppel [29] (p. 169) for



	Facts	Concepts	Opinions	Insights
R	6.75	4.47	1.78	3.91
U	6.34	5.78	3.19	5.13
A	6.13	5.28	4.41	5.41
C	5.38	5.28	4.00	5.88

(a) Pre-task expectations.



	Facts	Concepts	Opinions	Insights
R	6.77	3.94	1.52	3.42
U	6.22	5.91	2.41	5.36
A	6.29	5.48	3.90	5.13
C	5.48	5.52	3.87	5.71

(b) Post-task perceptions.

Fig. 3. RQ1: Task complexity effects on participants' pre-/post-task perceptions of info-type need.

post hoc comparisons. Slight variations in the F-statistic's degrees of freedom are due to missing questionnaire responses.

Pre-task Measures: Task complexity had a significant effect on all pre-task measures: difficulty, ($F(3, 93) = 8.16$, $p < .001$), determinability ($F(3, 93) = 14.58$, $p < .001$), interest ($F(3, 93) = 21.80$, $p < .001$), and prior knowledge ($F(3, 93) = 23.23$, $p < .001$). Post hoc comparisons found the following differences: difficulty (R, U, A < C), determinability (R > U, A > C), interest (R, U, A > C), and prior knowledge (R, U, A > C).

Post-task Measures: Task complexity had a significant effect on all post-task measures: difficulty, ($F(3, 90) = 12.06$, $p < .001$), satisfaction ($F(3, 90) = 8.35$, $p < .001$), interest increase ($F(3, 90) = 3.56$, $p < .05$), and knowledge increase ($F(3, 90) = 3.37$, $p < .001$). Post hoc comparisons found the following differences: difficulty (R, A < U, C), satisfaction (R, A > U, C), interest increase (A > C), and knowledge increase (A > C).

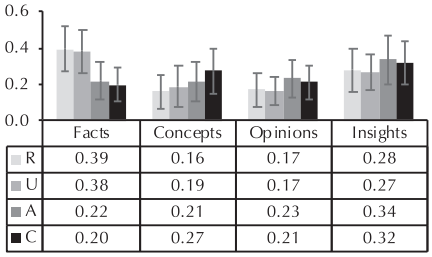
Logged Interaction Measures: Figure 2(c) shows the differences in three interaction measures according to task complexity: (1) number of queries issued, (2) number of SERP clicks, and (3) number of IB clicks. Task complexity had an effect on all measures: queries ($F(3, 93) = 5.45$, $p < .01$), SERP clicks ($F(3, 93) = 3.71$, $p < .05$), and IB clicks ($F(3, 93) = 3.581$, $p < .05$). Post hoc comparisons found the following differences: queries (R < U, C; U > A), SERP clicks (R, A < C), and IB clicks (R < U; U > A).

Summary: These results suggest that our manipulation of task complexity worked. Participants perceived complex tasks (particularly C vs. R) to be more difficult and less determinable, and reported lower levels of interest, prior knowledge, and satisfaction for complex tasks. Similarly, complex tasks required more search activity. We also note that task version U (soil degradation) turned out to be more complex than task version A (water purification). We elaborate on this unexpected outcome in Section 5.

4.2 RQ1: Perceptions of Info-type Need

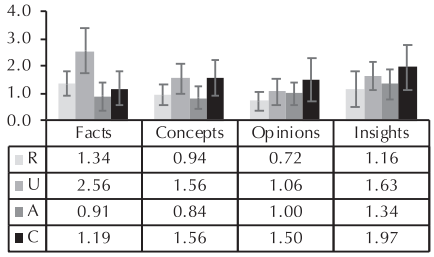
In RQ1, we investigate the effects of task complexity on participants' pre- and post-task perceptions about needing specific info-types to complete the task. To address this question, we analyzed participants' responses to the four items in the pre-/post-task questionnaires about whether the task required facts, concepts, opinions, and insights. Figures 3(a) and 3(b) show the differences in these measures according to task complexity.

Pre-task: Task complexity had a significant effect on all pre-task measures: facts ($F(3, 93) = 19.70$, $p < .001$), concepts ($F(3, 93) = 7.164$, $p < .001$), opinions ($F(3, 93) = 22.93$, $p < .001$), and insights ($F(3, 93) = 14.98$, $p < .001$). Post hoc comparisons found the following differences: facts (R > U, A > C), concepts (R < U; U > C), opinions (R < U, A, C; U < A), and insights (R < U, A, C).



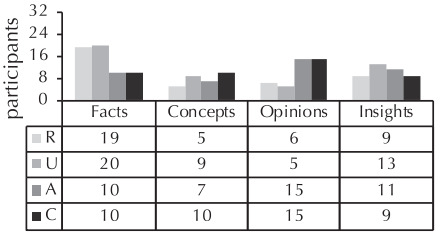
	Facts	Concepts	Opinions	Insights
R	0.39	0.16	0.17	0.28
U	0.38	0.19	0.17	0.27
A	0.22	0.21	0.23	0.34
C	0.20	0.27	0.21	0.32

(a) Info-type exposure time.



	Facts	Concepts	Opinions	Insights
R	1.34	0.94	0.72	1.16
U	2.56	1.56	1.06	1.63
A	0.91	0.84	1.00	1.34
C	1.19	1.56	1.50	1.97

(b) Info-type clicks.



	Facts	Concepts	Opinions	Insights
R	19	5	6	9
U	20	9	5	13
A	10	7	15	11
C	10	10	15	9

(c) Observed info-type use.

Fig. 4. RQ2: Task complexity effects on info-type use.

Post-task: Task complexity had a significant effect on all post-task measures: facts ($F(3, 93) = 13.17$, $p < .001$), concepts ($F(3, 93) = 11.94$, $p < .001$), opinions ($F(3, 93) = 20.31$, $p < .001$), and insights ($F(3, 93) = 12.65$, $p < .001$). Post hoc comparisons found the following differences: facts (R > U, A > C), concepts (R < U, A, C), opinions (R < U < A, C), and insights (R < U, A, C).

Summary: Our RQ1 results suggest three main trends about participants' pre-/post-task perceptions. First, participants perceived facts to be more useful during simple versus complex tasks. Second, participants perceived concepts to be more useful for complex versus simple tasks. Third, participants perceived opinions and insights, which tend to focus on subjective information (viewpoints, tips, advice), to be more useful for complex versus simple tasks.

4.3 RQ2: Info-type Use

In RQ2, we consider the effects of task complexity on participants' use of different info-types in the IB while completing tasks. We explore RQ2 from two perspectives: (1) using logged interaction data and (2) using observations from the stimulated recall interviews about participants' first and last IB use (if any) during each session.

Logged Interaction Measures: To analyze the extent of participants' use of specific info-types, we computed the following interaction measures: (1) the info-types' normalized exposure time and (2) the number of total clicks on an info-type tab. The normalized exposure time was measured as the percentage of time the info-type was visible (out of the total time the IB was visible). Figures 4(a)–4(b) show these measures according to task complexity.

In terms of exposure time, task complexity had a significant effect on the exposure time of facts ($F(3, 93) = 8.05, p < .001$). Post hoc comparisons found that facts had more exposure time for task versions R, U versus A, C. We also observed a trend that subjective info-types (opinions, insights) had more exposure time for complex tasks (A, C), although this trend did not reach significance.

In terms of clicks, task complexity had a significant effect on the number of clicks for facts ($F(3, 93) = 8.41, p < .001$). Post hoc comparisons found that facts had more clicks for task U vs. R, A, C.

	Facts	Concepts	Opinions	General	Resource	Confirm	Overview	Interest
R	13	1	0	0	3	13	0	7
U	5	7	0	4	4	6	5	8
A	3	9	4	3	3	3	1	8
C	5	2	6	5	3	2	11	8

Fig. 5. RQ3: Task complexity effects on motivations for IB use.

Observed use: We also analyzed info-type use based on analysis of the stimulated recall interviews. As part of the qualitative analysis, we noted which info-types were involved in each first/last IB use (based on participants' comments and screen actions). We then aggregated these at the participant/task level as shown in Figure 4(c) (e.g., 19 participants were observed using IB facts during task R).

Task complexity had a significant effect on the use of facts (Cochran's Q test, $Q(3) = 13.44$, $p < .01$) and opinions ($Q(3) = 15.78$, $p < .01$). Post hoc comparisons found the following differences: facts (R, U > A, C) and opinions (R, U < A, C).

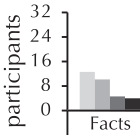
Summary: These results suggest two main trends. First, participants used facts more during the two simplest tasks (R, U) versus the two most complex tasks (A, C). Specifically, facts had significantly more exposure time and instances of observed use for tasks R and U versus A and C. Between the two simplest tasks (R, U), facts had more logged clicks for task U than task R (and also A, C). One possible interpretation is that both tasks R and U required facts, but U required more information than R (i.e., a list versus a single piece of information). The second trend is that opinions had more instances of observed use for complex tasks (A, C) versus simple tasks (R, U).

4.4 RQ3: Motivations for IB Use

Based on our qualitative coding of participants' responses during the stimulated recall interviews (Section 3.7), we identified four main motivations for using the IB: (1) to find a specific type of information (further sub-divided as described below); (2) to confirm the accuracy or completeness of previously found information (*confirm*); (3) to get an overview of the topic (*overview*); and (4) motivations based on interest and curiosity (*interest*). The *interest* motivation often focused on using the IB as a tool for exploration rather than to find information for a specific goal. We further sub-divided participants' motivations to find specific information into five categories based on their stated information goal: (to find) (a) *facts*, (b) *concepts*, (c) *opinions*, (d) *general* knowledge about the topic, and (e) useful *resources* such as an authoritative website.

Figure 5 shows the effects of task complexity on participants' self-reported motivations for using the IB, aggregated at the participant/task level (e.g., 13 participants reported looking for specific *facts* for the remember (R) task). Task complexity had a significant effect on the motivations to find *facts* ($Q(3) = 12.64, p < .01$), to find *concepts* ($Q(3) = 11.93, p < .01$), to find *opinions* ($Q(3) = 12.46$, $p < .01$), to *confirm* previously found information ($Q(3) = 14.80$, $p < .01$), and to get an *overview* ($Q(3) = 19.93$, $p < .001$). Post hoc comparisons found the following differences: facts (R > U, A, C), concepts (R < A; A > C), opinions (R, U < C), confirm (R > A, C), and overview (R, A < C).

Summary: These results show four main trends. First, task complexity had an effect on participants' motivations to use the IB to find specific types of information. Specifically, more participants were motivated to find *facts* for task R (simplest), *concepts* for task A (moderately complex), and *opinions* for task C (most complex). Second, for task R, more participants were motivated to use the IB to *confirm* previously found information. Third, for task C, more participants were motivated



	Facts	Concepts	Opinions	General	Resource	Reassure	Strategy	No Gain
R	13	0	0	0	2	16	1	10
U	10	5	0	3	5	5	4	12
A	5	9	1	1	7	5	2	7
C	4	2	8	10	7	2	4	11

Fig. 6. RQ4: Task complexity effects on gains from IB use.

to use the IB to get an *overview* of the task topic. Finally, across all tasks, participants were equally motivated to use the IB to find a useful *resource* or because of their *interest* (i.e., non-goal-oriented, exploratory use).

4.5 RQ4: Gains from IB Use

Based on the stimulated recall interviews, we identified four main categories of gains that participants reported from using the IB: (1) gained specific information (further sub-divided as described below); (2) gained reassurance about the completeness or accuracy of previously found information (*reassurance*); (3) gained ideas about things to search for (*strategy*); and no gain, meaning that the participant used the IB and did not report any gains (*no gain*). We further sub-divided participants' specific information gains into five categories based on the information gained: (a) a *fact* or answer, (b) *concepts*, (c) *opinions*, (d) *general* knowledge about the topic, and (e) useful *resources* such as an authoritative website.

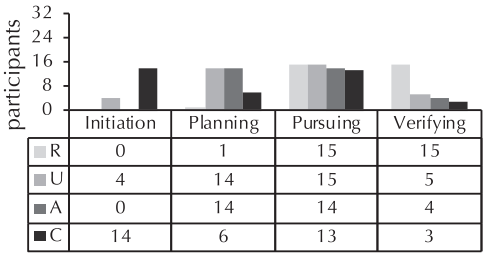
Figure 6 shows the effects of task complexity on participants' self-reported gains from using the IB, aggregated at the participant/task level. Task complexity had a significant effect on reported gains of *facts* ($Q(3) = 10.13$, $p < .05$), *concepts* ($Q(3) = 14.53$, $p < .01$), *opinions* ($Q(3) = 21.48$, $p < .001$), *general knowledge* ($Q(3) = 18.30$, $p < .001$), and *reassurance* ($Q(3) = 201.12$, $p < .001$). Post hoc comparisons found the following differences: facts (R > A, C), concepts (R, C < A), opinions (R, U, A < C), general knowledge (R, U, A < C), and reassurance (R > U, A, C).

Summary: These results show three main trends. First, task complexity affected the specific types of information participants gained from the IB. Specifically, more participants gained *facts* for task R (simplest), *concepts* for task A (moderately complex), and *opinions* and *general knowledge* for task C (most complex). Second, participants mainly reported gaining *reassurance* for task R. Finally, few participants reported gaining a new *strategy* from the IB, which suggests that participants mostly used the IB to gain information rather than ideas about things to search for on their own.

4.6 RQ5: Stage of IB Use

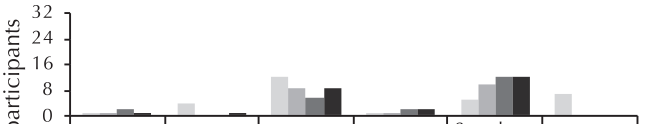
Figure 7 shows the effects of task complexity on participants' self-reported stages of IB use. Task complexity had a significant effect on IB use during: initiation ($Q(3) = 32.75$, $p < .001$), planning ($Q(3) = 21.99$, $p < .001$), and verifying ($Q(3) = 16.61$, $p < .01$). Post hoc comparisons found the following differences: initiation (R, U, A < C), planning (R, C < U, A), and verifying (R > U, A, C).

Summary: These results show four main trends. First, more participants used the IB for verifying during the simplest task (R). Second, more participants used the IB for planning during the moderately complex tasks (U, A). Third, more participants used the IB for initiation for the most complex task (C). Finally, across all tasks, participants equally used the IB for pursuing (i.e., while implementing a chosen search strategy).



	Initiation	Planning	Pursuing	Verifying
R	0	1	15	15
U	4	14	15	5
A	0	14	14	4
C	14	6	13	3

Fig. 7. RQ5: Task complexity effects on stage of IB use.



	Effort	Hindsight	Mismatch	Novelty	Search on Own	Easy Task
R	1	4	12	1	5	7
U	1	0	9	1	10	0
A	2	0	6	2	12	0
C	1	1	9	2	12	0

Fig. 8. RQ6: Task complexity effects on IB non-use motivations.

4.7 RQ6: Motivations for Non-use

Based on the stimulated recall interviews, we identified six main reasons participants *purposely avoided* using the IB: (1) it required too much effort (*effort*); (2) in hindsight, it would have been helpful to use it (*hindsight*); (3) there was a mismatch between the information in the IB and my needs (*mismatch*); (4) unfamiliarity with the IB (*novelty*); (5) I preferred to search on my own (*search on own*); and (6) the task was easy/straightforward (*easy task*).

Figure 8 shows the effects of task complexity on participants' self-reported reasons for avoiding the IB. Task complexity had a significant effect on *hindsight* ($Q(3) = 8.60$, $p < .05$) and *easy task* ($Q(3) = 21.00$, $p < .001$). Post hoc comparisons found the following differences: hindsight (R > U, A) and easy task (R > U, A, C).

Summary: These results suggest three main points. First, many participants started the tasks by searching on their own. Second, a tool such as the IB may not be perceived as necessary for very simple tasks such as task R. Finally, across all task types, participants reported various types of mismatches between their needs and the information presented in the IB. In this study, the IB contained information relevant to the task, but was static throughout the search session. One possible explanation is that participants who reported "mismatches" wanted information relevant to the current *sub-task* (i.e., relevant to the most recent query and search activity).

5 DISCUSSION

Next, we summarize our results, compare them to results from prior work, and discuss their implications.

5.1 Summary of Results

Task complexity affected task difficulty and search behaviors (RQ0)—Consistent with prior work [4, 12, 23, 28, 49], our RQ0 results suggest that our manipulation of task complexity had an

effect on participants' perceptions and behaviors. More cognitively complex tasks were perceived to be more difficult, less *a priori* determinable, and had lower levels of post-task satisfaction. Similarly, complex tasks required more search interaction (e.g., more queries and SERP/IB clicks).

Our RQ0 analysis also found an unexpected result—task U turned out to be more complex than task A. To reiterate, task U required generating a list (i.e., human activities that degrade soil fertility) and task A required comparing alternatives along different dimensions (i.e., water purification methods). One possible explanation is that task U had nuances that increased its complexity. First, not all phenomena that degrade soil fertility are caused by human activities (i.e., some are natural phenomena). Thus, task U required participants to learn about this distinction during the task. Second, task U involved more subjectivity. For example, there is disagreement about the extent to which certain human activities (e.g., grazing livestock) degrades soil fertility. Thus, task U required participants to assess the credibility of information found and synthesize opinions. Interestingly, our results suggest that these nuances of task U were not obvious from the task description. Hence, tasks U and A had similar levels of *pre-task* difficulty. However, compared to task A, task U had significantly greater levels of *post-task* difficulty, more queries issued, and more IB clicks.

Task complexity affected participants' perceptions about requiring specific types of information (RQ1)—The general trend in Figures 3(a)–3(b) is that facts were perceived to be more useful during the simplest task (R), while concepts, opinions, and insights were perceived to be more useful during more complex tasks (U, A, C). The same trend was observed in terms of participants' pre-task perceptions (Figure 3(a)) and post-task perceptions (Figure 3(b)), suggesting that participants were able to *anticipate* the types of information required by the task.

Our RQ1 results suggest that complex tasks require: (1) a greater variety of information types (i.e., concepts/definitions, opinions, and insights about the task domain and/or approaches to the task); and (2) more subjective information (i.e., opinions/insights versus facts). Our RQ1 results are largely consistent with our RQ2 results on participants' actual use of info-types in the IB. Next, we discuss our RQ2 results and compare our RQ1–RQ2 results to those from prior work.

Task complexity affected participants' use of info-types in the IB (RQ2)—Our RQ2 results (Figures 4(a)–4(c)) are consistent with those from RQ1—facts were used more during simpler tasks (R, U) and other info-types were used more during complex tasks (A, C). Based on observations of participants' first/last IB use of the session (Figure 4(c)), opinions were used more often during complex tasks (R, U < A, C).

Our RQ1–RQ2 results suggest that complex tasks require: (1) a greater variety of information types and (2) more subjective information types. This trend in our results resonates with several veins of prior work. First, our results are consistent with and extend prior work by Byström et al. [8, 10]. In these studies, task complexity was viewed through the lens of *a priori* determinability, and information types were defined based on their *functional role*—based on *how* the information was used during the task. Similar to our results, Byström et al. [8, 10] also observed a greater variety in the functional roles of information during complex versus simple tasks. Specifically, simple tasks required mostly problem information (i.e., to help define the task). Conversely, complex tasks required different types of information, including problem information, problem-solving information (i.e., procedural information on how to approach the task), and domain information (i.e., general information in the task domain). Compared to Byström et al. [8, 10], our info-types (facts, concepts, opinions, and insights) were more strongly defined based on inherent characteristics of the information. Thus, our RQ1–RQ2 results suggest that the findings from Byström et al. [8, 10] generalize to other info-type definitions.

An important question is: Why did complex tasks require a greater variety of our info-types? Considered in conjunction with results from Byström et al. [8, 10], it is possible that the concepts,

opinions, and insights in our study served multiple functional roles. For example, it is possible that concepts had a tendency to serve as domain information, and that opinions/insights had a tendency to serve as domain information and/or problem-solving information.

As previously mentioned, our RQ1–RQ2 results also found that complex tasks required more subjective information (i.e., opinions/insights vs. facts/concepts). This finding is consistent with Zhang [51]. In that study, participants were asked to recall information sources after completing simple vs. complex tasks. After simple tasks, participants had a greater tendency to recall sources containing objective information (e.g., fact sheets, FAQs). Conversely, after complex tasks, participants had a greater tendency to recall sources containing subjective or debatable information (e.g., clinical study reports). Considering Campbell's perspective [11], complex tasks have more paths to a solution, greater uncertainty about paths, and greater interdependence between paths. This suggests a possible explanation of our results—subjective/experiential information is more useful during tasks that require deciding which path(s) to follow and why.

Task complexity affected participants' motivations for engaging with the IB (RQ3)—Our RQ3 results suggest three important trends (Figure 5). First, task complexity had an effect on the types of information being sought in the IB. Participants reported engaging with the IB to find factual information during the simplest task (R), conceptual information during moderately complex tasks (U, A), and opinionated information during the most complex task (C). This trend resonates with our RQ1–RQ2 results and suggests that complex tasks require more subjective information. Second, during the simplest task (R), participants reported engaging with the IB to confirm the veracity and completeness of information found on their own. In other words, during the simplest task, participants were able to do the task on their own, but engaged with the IB to confirm the quality of information found. Finally, during the most complex task (C), participants reported engaging with the IB to gain an overview of the task domain. This trend resonates with results from Byström et al. [8, 10], which found that complex (versus simple tasks) required more domain information (i.e., background information) and problem-solving information (i.e., information on how to approach the task).

Task complexity affected participants' gains obtained from the IB (RQ4)—Participants reported different gains obtained by engaging with the IB (Figure 6). Our RQ4 results are largely consistent with our RQ3 results and suggest four main trends. First, task complexity had an effect on the types of information gained from the IB. Similar to our RQ3 results, participants reported gaining factual information during the simplest task (R), conceptual information during moderately complex task (A), and opinionated information during the most complex task (C). Second, during the simplest task (R), participants were motivated to engage with the IB to confirm information (RQ3) and reported gaining reassurance (RQ4). Third, during the most complex task (C), participants were motivated to engage with the IB to gain an overview of the task domain (RQ3) and reported gaining general information (RQ4). Finally, during complex tasks, more participants reported learning about useful resources to solve the task. While this upward trend did not reach significance, it also resonates with findings from Byström et al. [8, 10], which found that complex tasks required more problem-solving information (i.e., information on how to approach the task).

Task complexity affected the stages participants were in when engaging with the IB (RQ5)—Our RQ5 results suggest four main trends (Figure 7). First, *verifying* was the most frequently cited stage for our simplest task (R). This result is consistent with our RQ3–RQ4 results, which found that participants during task R mostly engaged with the IB to confirm information and that they gained reassurance. Second, *planning* (strategizing about the needed information) was a more frequent stage for moderately complex tasks (U, A). Third, *initiation* (getting an initial understanding of the task) was a more frequent stage for our most complex task (C). This result is also consistent with our RQ3–RQ4 results, which found that participants during task C mostly

engaged with the IB to get an overview of the task and gained general information. Finally, *pursuing* (searching for a specific piece of information) was an equally frequent stage for all tasks.

Capra et al. [12] also investigated the effects of task complexity on participants' use of a peripheral search assistance tool that displayed search trails. An analysis of participants' motivations for engaging with the tool found similar trends as our RQ3–RQ5 results. During simple tasks, participants used the tool mostly to verify information found on their own. Conversely, during complex tasks, participants used the tool to get started with the task and to discover new search strategies.

Task complexity had an effect on participants' motivations for avoiding the IB (RQ6)—Our RQ6 results suggest three main trends (Figure 8). First, as one might expect, two motivations were cited by more participants during our simplest task (R): (1) the task was easy and (2) the IB could have been useful in hindsight.

Second, more participants cited preferring to "search on their own" during complex tasks, although this upward trend did not reach significance. We see two possible explanations for this upward trend. First, prior work has found that users sometimes avoid help systems due to the cost of cognitively disengaging (and subsequently re-engaging) with the main task [20]. One possible explanation is that the cost of disengaging with the main task increases with its complexity. In other words, once someone engages with a complex task, there may be a greater opportunity cost of disengaging with the task. Second, prior work has found that searchers' strategies *diverge* more during complex versus simple tasks [28]. In other words, during a simple task searchers tend to adopt similar strategies. Conversely, during complex tasks searchers tend to adopt unique "paths" that are different from others' paths. Thus, a second explanation is that for complex tasks participants preferred to search on their own because the information in the IB was less compatible with their unique approach to the task. A similar trend was found in Capra et al. [12]. In that study, participants more often cited avoiding a search assistance tool, because they preferred to search on their own during complex (versus simple) tasks. From Campbell's perspective on task complexity [11], complex tasks have more paths to the solution (i.e., are more open-ended). Thus, searchers approach complex tasks in different ways.

Finally, across all levels of task complexity, participants often cited avoiding the IB, because they did not expect it to have the information needed at the time (i.e., a mismatch). One possible explanation is that certain needs of participants were not effectively addressed by our characterization of task-related information as facts, concepts, opinions, and insights. A second explanation stems from the fact that the information in the IB was static throughout the search session. Perhaps some participants needed information specific to the current sub-task, rather than the task as a whole. As discussed below, this issue could be addressed with a dynamic IB tool that could extract info-types from the underlying corpus (or external resources) and display information that is relevant to the user's current sub-task.

5.2 Implications and Opportunities for Future Work

Here, we discuss implications of our results and opportunities for future work.

Task complexity affected info-type use (RQ1–RQ3)—Our results found that task complexity had an effect on which info-types participants perceived to be useful pre- and post-task (RQ1) and on which info-types participants used during the task (RQ2). Additionally, task complexity affected which info-types participants were motivated to seek when engaging with the IB—facts for the simplest task (R), concepts for a moderately complex task (A), and opinions for the most complex task (C).

An important implication of this result is that systems may be able to infer the complexity of a current user's task and favor specific types of information in a search assistance tool such as the IB or in the main search results. Several findings from prior work suggest that this is possible.

First, prior research has found that task complexity influences search behaviors [12, 23, 28, 49]. As one might expect, complex tasks require more interaction (e.g., more queries, clicks, bookmarks) and have more evidence of trial-and-error or backtracking (e.g., more abandoned queries, more clicks without a bookmark, and more repeated queries with the same search intent). Prior studies have found some success in automatically inferring (subjective) task *difficulty*, a construct closely related to (objective) task complexity [4, 36].

Second, while in this article we used high-quality, manually curated information to populate the IB, prior work suggests it may be possible to extract/display task-related facts, concepts, opinions, and insights automatically. In terms of facts vs. opinions, prior NLP research has focused on using machine learning to distinguish between objective vs. subjective statements [46, 47]. In terms of concepts, prior IR research has focused on extracting and displaying query-related *facets* (as used in traditional faceted search) from the top search results [31] or external knowledge bases [27]. Finally, insights (i.e., tips and advice about the task or domain) could be extracted from online sources that focus on experiential information (e.g., forums , Q&A sites, social media).

Task complexity affected the stage of IB use (RQ3–RQ5)—Our results found that task complexity affected participants' motivations for engaging with the IB, the gains they obtained, and the stage(s) during which they used the IB. For our simplest task (R), participants engaged with the IB to confirm information (RQ3), gained reassurance from interacting with it (RQ4), and engaged while in the verifying stage (RQ5). Conversely, for our most complex task (C), participants engaged with the IB to get an overview of the task (RQ3), gained general information from it (RQ4), and engaged during the task's initiation phase (RQ5). In other words, these trends suggest that during simple tasks participants engaged with the IB after searching on their own (i.e., toward the end of the session). However, during complex tasks, participants engaged with the IB while getting started with the task. This result resonates with findings from Byström et al. [8, 10], who found that complex tasks required more problem-solving information (i.e., to understand possible approaches to the task).

These results suggest opportunities and challenges for designing dynamic search assistance tools similar to the IB. In terms of opportunities, inferring task complexity may help a system decide *when* to provide assistance—early in the session for complex tasks, and later in the session for simple tasks. Furthermore, inferring task complexity may help a system make decisions about how to populate and rank items in info-type boxes/tabs. For instance, for simple tasks (for which users mostly want verification), the system might prioritize information types similar to those encountered during the session. Conversely, for complex tasks (for which users mostly want overview information), the system could prioritize general (vs. specific) information.

Results also suggest challenges for designing dynamic support systems. For complex tasks, participants had difficulty getting started with the task—at the start of the session. Paradoxically, at the start of a session, a dynamic system may not have enough information to predict task complexity or to predict the task topic to rank items. Recent studies have aimed to infer task characteristics (e.g., difficulty) *early* in a search session versus at the end [4, 38]. These studies provide a starting point, but more research is needed.

Task complexity affected motivations for IB avoidance (RQ6)—Our results suggest that task complexity affected participants' motivations for avoiding the IB. These results also have implications for designing support tools similar to the IB.

Effective support tools must be able to distinguish between good and bad abandonment (i.e., non-use). In our case, for our simplest task (R), participants cited avoiding the IB because the task was easy, but acknowledged the IB might have been useful in retrospect. Thus, inferring task complexity may help a system interpret a lack of interaction. For simple tasks, a lack of interaction

may not necessarily mean that the tool provided non-relevant information. Conversely, for complex tasks, a lack of interaction may more reliably signal a negative outcome.

For more complex tasks (A, C), participants cited avoiding the IB because they preferred to "search on their own." Prior research has shown that users often avoid help tools because of the cost of disengaging (and re-engaging) with the main task [20]. Our results suggest that complex tasks may have a higher cost of disengaging, incentivizing users to avoid any assistance. To alleviate this challenge, support tools such as the IB may need to clearly convey or explain *how* they can help. For example, a system such as the IB should display as the default tab the info-type that is the most relevant to the current task (e.g., by inferring task complexity as advocated above).

Irrespective of task complexity, participants mostly avoided the IB because of "mismatch" (RQ6)—Across all complexity levels, participants cited avoiding the IB, because they did not expect it to have the information needed at the time. This result also presents opportunities for future work. In our study, the IB contained high-quality, task-related information, but was static throughout the session. As advocated above, future work should consider a dynamic IB tool that is populated based on a user's activities during the search session. A dynamic IB tool may provide two types of benefits. First, it may provide information that better matches a user's specific approach to the task. Prior work has found that as task complexity increases, users *diverge* in their approaches to the task [28]. Thus, a dynamic IB tool might be especially beneficial during complex tasks, by providing information that matches a user's specific approach to the task. Second, a dynamic IB tool might be able to provide information that is not only relevant at the task-level, but also at the current sub-task level. Finally, in this study, we characterized info-types as facts, concepts, opinions, and insights. Information can be characterized in myriad other ways. Thus, future work is needed to determine whether certain needs of users can be more effectively met with other info-types not considered in our study.

5.3 Caveats and Limitations

Our study has several caveats worth noting. First, we manipulated task complexity using only four tasks, one per complexity level (R, U, A, and C). In this respect, task complexity was potentially confounded with other task characteristics (e.g., task topic/domain). Alternatively, we could have designed and used more than one task per complexity level, allowing us to possibly tease apart the effects of task complexity versus other task characteristics. That being said, our RQ0 results suggest that our task complexity manipulation worked and is mostly consistent with prior work [4, 12, 23, 28, 49]. For example, complex tasks were perceived to be more difficult before and after the task, particularly when comparing between the simplest and most complex tasks (R vs. C). As one exception (and as previously mentioned), task U turned out to be more difficult than expected due to nuances of the task (e.g., it involved more subjectivity). Using multiple tasks per complexity level would have resulted in a more robust analysis. However, due to the amount of effort involved in this type of user study, it is not uncommon for IIR studies to investigate the effects of task characteristics using one task per category (e.g., see References [6, 34, 35]). Future work should further investigate our findings in the context of additional tasks/topics.

Second, we note that think-aloud protocols have the potential to influence participants' behaviors. Prior work [21, 22] has found that procedures that involve concurrent verbalizations at Ericsson and Simon's [1] Level 1 (vocalization of "inner speech") and Level 2 (verbalization of thoughts) typically have minimal (if any) impacts on performance. Verbalizations at Ericsson and Simon's [1] Level 3 (descriptions and explanations of activities) are more likely to impact task performance in various ways. However, think-aloud protocols are a powerful and commonly used tool in human-computer interaction and interactive IR studies to gain insights into users' actions [2, 37, 40]. In our study, we took efforts to minimize the effects and reactiveness of the think-aloud

protocol (i.e., we used the same protocol in all conditions; the moderator did not ask probing questions, but instead reminded participants to "Keep thinking aloud about what you are thinking" if they fell silent for a period of time). However, it is possible that thinking aloud could have altered participants' behaviors.

A third caveat involves our qualitative analysis of participants' use of the IB, which influenced our RQ3–RQ5 results on motivations, gains, and stage. As previously mentioned, to avoid overwhelming participants and to prevent moderator bias during the retrospective interview, we asked targeted questions about participants' first and last instance of IB use rather than *every* instance of IB use. It is reasonable to assume that the first/last instance of IB use occurred towards the beginning/end of the task. It is likely that this influenced the types of behaviors we observed for RQ3–RQ5. For example, in terms of RQ3, participants cited engaging with the IB to get an overview of the task (probably at the beginning) and to confirm information found on their own (probably at the end). Similarly, this bias toward beginning/end IB use likely influenced the gains observed in RQ4 and the stages of IB use observed in RQ5. We choose to focus on the first and last IB uses, because they represent well-defined events and interesting points of contrast. Our RQ3–RQ5 results should be interpreted in the context that we inquired about the first and last instance of IB use. It is possible that some motivations, gains, and stages related to instances of IB use did not emerge from our analysis, because we missed instances of IB use during the middle of the task.

6 CONCLUSION

A core question in IR is: How do task characteristics influence the *types of information* that are required or useful for completing the task? Much research has investigated this question under the umbrella of *relevance criteria*—understanding how contextual factors (i.e., attributes of the user, task, or situation) influence relevance from a user's perspective [42]. Our research in this article adds to this body of knowledge by focusing on one task dimension (i.e., cognitive complexity) and by characterizing task-related information as: factual statements, concepts, opinionated statements, and insights (i.e., helpful advice about approaches to the task or its domain). Compared to prior work [10], our four different *info-types* varied based on inherent characteristics of the information (e.g., objective vs. subjective). We reported on a laboratory study that investigated the effects of task complexity on the info-types used by participants to complete the task. Participants completed four tasks of varying complexity and had access to task-related facts, concepts, opinions, and insights in an *auxiliary* search assistant tool (the InfoBoxes or IB) that was complementary to a standard web search engine.

In terms of our six research questions (RQ1–RQ6), our results found the following trends: First, our results suggest that simple tasks required simple information (i.e., facts) and complex tasks required complex information—concepts/definitions about the domain, opinions, and insights about approaches to the task. This trend was consistent in terms of participants' pre-/post-task perceptions (RQ1), info-types used (RQ2), motivations for engaging with the IB (RQ3), and gains obtained from the IB (RQ4).

Second, our results found an interesting trend in terms of participants' engagement with facts versus opinions. Facts were used more for simple tasks, and opinions were used more for complex tasks. This trend was consistent based on participants' observed use of info-types in the IB (RQ2), motivations for engaging with the IB (RQ3), and gains obtained from the IB (RQ4). Prior research has argued that complex tasks involve greater uncertainty. From Campbell's perspective [11], complex tasks have more paths to the outcome(s) and more uncertainty about paths. Similarly, from the perspective of *a priori* determinability, complex tasks have more uncertainty about the form of the solution, required inputs, and processes involved in addressing the task. One possible explanation

is that *reducing* the uncertainty of complex tasks often requires information that is *subjective* or *experiential* in nature—perspectives, anecdotes, evaluations, recommendations, tips, and advice.

Third, our RQ3–RQ5 results suggest that task complexity had an effect on participants' goals for engaging with the IB. For simple tasks, participants mostly engaged with the IB to *confirm information* found on their own, and for complex tasks, participants mostly engaged with the IB to get an *initial overview* of the task domain. Consistent with this trend, participants engaged with the IB during the *verifying* stage for simple tasks and during the *initiation* stage for complex tasks (RQ5).

Finally, our RQ6 results point to several reasons why users may avoid a support tool such as the IB. Participants often avoided the IB, because they did not expect it to have the information needed at the time. This points to two areas for future work. First, in this study, the IB contained high-quality curated information, but was static throughout the session. Future work may consider a dynamic IB tool that is algorithmically populated with info-types extracted from the underlying collection or an external resource. A dynamic IB may be able to provide information that is more relevant to a user's current needs—information relevant to the current sub-task. Second, in this study, we characterized information nuggets as facts, concepts, opinions, and insights. Information can be characterized in many other ways. Thus, future research should consider whether *other* info-types may help address the various needs of searchers.

The main finding from our study suggests that task complexity influences the types of information that are useful for completing the task. Results from prior work [4, 12, 23, 28, 49] (and our RQ0 results) also suggest that task complexity influences search behaviors that can be captured by a search system. As a next step, future research should focus on *predicting* task complexity and using these inferences to promote specific types of information in the main search results or an auxiliary support tool such as the InfoBoxes.

APPENDIX

A EXAMPLE INFO-TYPES AND PRE-/POST-TASK QUESTIONNAIRES

Table 1. Examples of Facts, Concepts, Opinions, and Insights for Our *Create* Task on "Soapbox Derby," the Most Complex

Rank	Facts	Concepts	Opinions	Insights
1	A car's acceleration is impeded as it rolls down the hill by aerodynamic drag, vibration, friction, moment of inertia, and most important: driving. Minimizing the impact of these primary influences will increase the car's speed.	All-American Soap Box Derby	eBay and Gumtree are useful sources for parts, but it's worth spending some time perusing second-hand bike shops and skips. We got the wheels for our prototype from a kid's bike that had been thrown out.	If you are building a soapbox for a race, there are specific regulations for size and materials. However, soapboxes built just for fun can be made according to many designs and from a huge range of materials.
2	Even at relatively low speeds, aerodynamics are important. Make sure your design interacts as little as possible with airflow.	Billy cart	Building a gravity racer is battle of compromises that limit its performance. But what if you could have your cake and eat it? Here is a proposal that might allow you to do just that.	To enter an official Soapbox Derby race, you have to purchase a car kit from their website. If my niece is just looking to build a car and race for fun, a DIY design might be more appropriate.
3	These instructions demonstrate how to build an inexpensive simple Soap Box Derby Car. You should be able to build this car for about $50.	gravity car	Strong wheels and strong axles are the key.	There are many step-by-step how-to guides for building a gravity-powered car from beginning to end. You could follow these strictly or use these to brainstorm your own design.

(Continued)

Table 1. Continued

Rank	Facts	Concepts	Opinions	Insights
4	Soap Box Cars are homemade vehicles that use only gravity to move—no engine. Make a Soap Box Car with plywood and wagon wheels, or get ambitious—with fiberglass frames and ball bearing hubs.	Soap Box Derby	If something is built on a sensible chassis—something with a robust frame, four wheels and a means of steering it—you can stick anything on top of it.	Consider if you want to buy a soapbox kit or build the car from the scratch. Kits will provide much of what you need to build the car, but will also curtail some of your creative freedom.
5	Selecting the best parts and determining the best car setup for racing involves trade-offs. The difficulty for contestants is working out the trade-offs to make their car as fast as possible on a specific track and ramp combination.	Aero-dynamics	The designs that tend to go the quickest are when the driver is sat as low as possible and between the wheels, rather than up high, as they are prone to topple over.	It could be helpful to know that if you want to compete in official races the parts on your car all need to be approved.
6	While many commercial kits are available for purchase that can lead to a working car being put together in only a few hours, schematics for building a car from scratch are still available to the public. Many of these blueprints can result in a car that is suitable for a derby at a fraction of the cost of a commercial.	AASBD	Some of the really elaborate ones, you know they aren't going to make it to the bottom of the hill when there is just too much going on.	There are countless inventive designs and ideas for soapbox and gravity-powered cars on Pinterest, including images and how-to videos.
7	Box car racing, commonly referred to as soap box car racing, is the building and driving of home-made cars. These small cars do not contain a motor and simply rely on the power of gravity, reaching speeds up to 30 miles per hour.	Soapbox car	Start with something that is already a reasonably proven and robust device. You see so many soapboxes with a bicycle or go-kart design as the basis, or with elements of those things, and those tend to have a good chance.	Soapbox cars don't have to be expensive. They can also be built from second-hand materials.
8	Cars competing in this and related events are unpowered, relying completely upon gravity to move.	Gravity racer	Beginners should choose wood as their main material. The basic tools are in most people's garage, and sheet timber is pretty cheap. All-metal cars offer greater strength and longevity, but demand more specialist skills.	—

Due to human error, we mistakenly included only seven (versus eight) insights for the *create* task.

Table 2. Pre-task Questionnaire Items Designed to Measure Expected Difficulty (4 Items), Determinability (3 Items), Interest, Prior Knowledge, and Expected Need for Facts, Concepts, Opinions, and Insights

Diff1: I think it will be difficult to complete this task.
Diff2: I think it will be difficult to search for information to complete this task.
Diff3: I think it will be difficult to integrate the information I find to complete this task.
Diff4: I think it will be difficult to decide when I have enough information to complete this task.
Det1: Right now, I know what specific things to look for to complete this task.
Det2: Right now, I know what steps I need to take to complete this task.
Det3: Right now, I know what my solution to this task will look like.
I am interested in this topic.
I already know a lot about this topic.
This task will require gathering factual information.
This task will require learning important concepts about the topic.
This task will require gathering information about people's feelings, tastes, and opinions.
This task will require gaining new insights, knowledge, and tips.

Table 3. Post-task Questionnaire Items Designed to Measure Experienced Difficulty (4 Items), Satisfaction (4 Items), Interest Increase, Knowledge Increase, and the Use for Facts, Concepts, Opinions, and Insights During the Task

Diff1: It was difficult to complete this task.
Diff2: It was difficult to search for information to complete this task.
Diff3: It was difficult to integrate the information I found to complete this task.
Diff4: It was difficult to decide when I had enough information to complete this task.
Sat1: I am satisfied with the amount of information I found to complete this task.
Sat2: I am satisfied with the amount of time I spent on this task.
Sat3: I am satisfied with the quality of information I found to complete this task.
Sat4: I am satisfied with the strategy I took to find information for this task.
My interest in this topic has increased.
My knowledge of this topic has increased.
This task required gathering factual information.
This task required learning important concepts about the topic.
This task required gathering information about people's feelings, tastes, and opinions.
This task required gaining new insights, knowledge, and tips.

REFERENCES

[1] K. A. Ericsson and H. A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data (rev. ed.)*. The MIT Press, Cambridge, MA.

[2] Obead Alhadreti and Pam Mayhew. 2018. Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the CHI*. ACM, New York, NY, Article 44, 12 pages.

[3] Lorin W. Anderson and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Pearson.

[4] Jaime Arguello. 2014. Predicting search task difficulty. In *Proceedings of the ECIR*. Springer, 88–99.

[5] Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette. 2011. A methodology for evaluating aggregated search results. In *Proceedings of the ECIR*. Springer-Verlag, 141–152.

[6] Kumaripaba Athukorala, Dorota Glowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *J. Assoc. Inform. Sci. Technol.* 67, 11 (2016), 2635–2651.

[7] David J. Bell and Ian Ruthven. 2004. Searchers' assessments of task complexity for web searching. In *Proceedings of the ECIR*. Springer, 57–71.

[8] Katriina Byström. 2002. Information and information sources in tasks of varying complexity. *J. Assoc. Inform. Sci. Technol.* 53, 7 (2002), 581–591.

[9] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *J. Assoc. Inform. Sci. Technol.* 56, 10 (2005), 1050–1061.

[10] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Inform. Proc. Manag.* 31, 2 (1995), 191–213.

[11] Donald J. Campbell. 1988. Task complexity: A review and analysis. *Acad. Manag. Rev.* 13, 1 (1988), 40–52.

[12] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. 2015. Differences in the use of search assistance for tasks of varying complexity. In *Proceedings of the SIGIR*. ACM, 23–32.

[13] Robert Capra, Jaime Arguello, and Yinglong Zhang. 2017. The effects of search task determinability on search behavior. In *Proceedings of the ECIR*. Springer, 108–121.

[14] Lynne Cooke. 2010. Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Trans. Prof. Commun.* 53, 3 (2010), 202–215.

[15] Anita Crescenzi, Yuan Li, Yinglong Zhang, and Robert Capra. 2019. Towards better support for exploratory search through an investigation of notes-to-self and notes-to-share. In *Proceedings of the SIGIR*. ACM.

[16] Shane Culpepper, Fernando Diaz, and Mark Smucker (Eds.). 2018. Report from the third strategic workshop on information reterival in lorne (SWIRL'18). *SIGIR Forum* 52, 1 (2018), 34–90.

[17] Lynette D. Henderson and Julie Tallman. 2006. *Stimulated Recall and Mental Models: Tools for Teaching and Learning Computer Information Literacy*. Scarecrow Press, 55–91.
[18] Debora Donato, Francesco Bonchi, Tom Chi, and Yoelle Maarek. 2010. Do you want to take notes?: Identifying research missions in Yahoo! Search pad. In *Proceedings of the WWW*. ACM, 321–330.
[19] Huizhong Duan, Yanen Li, ChengXiang Zhai, and Dan Roth. 2012. A discriminative model for query spelling correction with latent structural SVM. In *Proceedings of the EMNLP-CoNLL*. Association for Computational Linguistics, 1511–1521.
[20] Garett Dworman and Stephanie Rosenbaum. 2004. Helping users to use help: Improving interaction with help systems. In *Proceedings of the CHI*. ACM, 1717–1718.
[21] Mark C. Fox, A. Ericsson, and R. Best. 2011. Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychol. Bull.* 137, 2 (2011), 316–344.
[22] Morten Hertzum, Kristin D. Hansen, and Hans H. K. Andersen. 2009. Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behav. Inform. Technol.* 28, 2 (2009), 165–181.
[23] Xiao Hu and Noriko Kando. 2017. Task complexity and difficulty in music information retrieval. *J. Assoc. Inform. Sci. Technol.* 68, 7 (2017), 1711–1723.
[24] Chunsheng Huang and Iris Xie. 2011. Help feature interactions in digital libraries: Influence of learning styles. In *Proceedings of the ASIST*.
[25] Bernard J. Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *Inform. Proc. Manag.* 45, 6 (2009), 643–663.
[26] Bernard J. Jansen and Michael D. McNeese. 2005. Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *J. Assoc. Inform. Sci. Technol.* 56, 14 (2005), 1480–1503.
[27] Zhengbao Jiang, Zhicheng Dou, and Ji-Rong Wen. 2017. Generating query facets using knowledge bases. *IEEE Trans. Knowl. Data Eng.* 29, 2 (2017), 315–329.
[28] 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 ICTIR*. ACM, 101–110.
[29] Geoffrey Keppel and Thomas D. Wickens. 1991. *Design and Analysis: A Researcher's Handbook* (3rd ed.). Prentice Hall.
[30] Youngho Kim and W. Bruce Croft. 2014. Diversifying query suggestions based on query documents. In *Proceedings of the SIGIR*. ACM, 891–894.
[31] Weize Kong and James Allan. 2016. Precision-oriented query facet extraction. In *Proceedings of the CIKM*. ACM, 1433–1442.
[32] J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.
[33] Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Inform. Proc. Manag.* 44, 6 (2008), 1822–1837.
[34] Chang Liu, Jingjing Liu, and Nicholas J. Belkin. 2014. Predicting search task difficulty at different search stages. In *Proceedings of the CIKM*. ACM, 569–578.
[35] 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 JCDL*. ACM, 69–78.
[36] Jingjing Liu, Chang Liu, Michael Cole, Nicholas J. Belkin, and Xiangmin Zhang. 2012. Exploring and predicting search task difficulty. In *Proceedings of the CIKM*. ACM, 1313–1322.
[37] Stephann Makri, Ann Blandford, and Anna L. Cox. 2010. This is what I'm doing and why: Reflections on a think-aloud study of Dl users' information behaviour. In *Proceedings of the JCDL*. ACM, 349–352.
[38] Matthew Mitsui, Jiqun Liu, and Chirag Shah. 2018. How much is too much?: Whole session vs. first query behaviors in task type prediction. In *Proceedings of the SIGIR*. ACM, 1141–1144.
[39] Neema Moraveji, Daniel Russell, Jacob Bien, and David Mease. 2011. Measuring improvement in user search performance resulting from optimal search tips. In *Proceedings of the SIGIR*. ACM, 355–364.
[40] Judith Ramey, Ted Boren, Elisabeth Cuddihy, Joe Dumas, Zhiwei Guan, Maaike J. van den Haak, and Menno D. T. De Jong. 2006. Does think aloud work?: How do we know? In *Proceedings of the CHI EA*. ACM, 45–48.
[41] Miamaria Saastamoinen, Sanna Kumpulainen, Pertti Vakkari, and Kalervo Järvelin. 2013. Task complexity affects information use: A questionnaire study in city administration. *Inform. Res.* 19, 4 (2013).
[42] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *J. Assoc. Inform. Sci. Technol.* 58 (11 2007), 2126–2144.
[43] Milad Shokouhi. 2013. Learning to personalize query auto-completion. In *Proceedings of the SIGIR*. ACM, 103–112.
[44] Pertti Vakkari. 2003. Task-based information searching. *Ann. Rev. Inform. Sci. Technol.* 37, 1 (2003), 413–464.
[45] Ryen W. White and Jeff Huang. 2010. Assessing the scenic route: Measuring the value of search trails in web logs. In *Proceedings of the SIGIR*. ACM, 587–594.

[46] Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the CICLING*. Springer-Verlag, 486–497.
[47] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Comput. Ling.* 30, 3 (2004).
[48] Barbara M. Wildemuth, Luanne Freund, and Eliane G. Toms. 2014. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *J. Document.* 70, 6 (2014), 1118–1140.
[49] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, tanning beds, tattoos, and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the IIiX*. ACM, 254–257.
[50] Iris Xie and Colleen Cool. 2009. Understanding help seeking within the context of searching digital libraries. *J. Assoc. Inform. Sci. Technol.* 60, 3 (2009), 477–494.
[51] Yan Zhang. 2012. The impact of task complexity on people's mental models of medlineplus. *Inform. Proc. Manag.* 48, 1 (2012), 107–119.

Received March 2019; revised October 2019; accepted November 2019