# The Effects of Manipulating Task Determinability on Search Behaviors and Outcomes

Robert Capra[1], Jaime Arguello[1], Heather O'Brien[2], Yuan Li[1], Bogeum Choi[1]
[1]University of North Carolina at Chapel Hill
[2]University of British Columbia
{rcapra,jarguello}@unc.edu,h.obrien@ubc.ca,{yuanli,choiboge}@live.unc.edu

## ABSTRACT

An important area of IR research involves understanding how task characteristics influence search behaviors and outcomes. Task complexity is one characteristic that has received considerable attention. One view of task complexity is through the lens of *a priori* determinability—the level of *uncertainty* about task outcomes and processes experienced by the searcher. In this work, we manipulated the determinability of comparative tasks. Our task manipulation involved modifying the *scope* of the task by specifying exact items and/or exact (objective or subjective) dimensions to consider as part of the task. This paper reports on a within-subject study ($N$ = 144) where we investigated how our task manipulation influenced participants' perceptions, levels of engagement, search effort, and choice of search strategies. Our results suggest a complex relationship between task scope, determinability, and different outcome measures. Our most open-ended tasks were perceived to have low determinability (high uncertainty), but were the least challenging for participants due to satisficing. Furthermore, narrowing the scope of tasks by specifying items had a *different* effect than by specifying dimensions. Specifying items increased the task determinability (lower uncertainty) and made the task easier, while specifying dimensions did not increase the task determinability and made the task more challenging. A qualitative analysis of participants' queries suggests that searching for dimensions is more challenging than for items. Finally, we observed subtle differences between objective and subjective dimensions. We discuss implications for the design of IIR studies and tools to support users.

## KEYWORDS

Task determinability; user engagement; search behavior

## 1 INTRODUCTION

Search tasks are a central component of interactive information retrieval (IIR). As noted by Toms [21], search tasks play two important roles in IIR research. First, they serve as a vehicle for research. In IIR studies, experimenters must assign search tasks to study participants in order to observe their behaviors and evaluate systems. Second, search tasks are also often used as the object of study (i.e., as independent variables). From this perspective, the study of search tasks helps us understand how task characteristics translate to specific challenges faced by searchers, and informs the design of novel tools to support users.

A large body of research has focused on understanding how search tasks vary along different dimensions, including the search task's main activity (e.g., searching vs. browsing), goal (e.g., well-defined vs. amorphous), and structure (e.g., task complexity) [16]. Search task complexity is one characteristic that has received considerable attention in recent work, and has been found to influence search behaviors and outcomes [3, 5, 7, 8, 11, 13, 25]. Task complexity is itself a complicated construct that has been studied from different perspectives [24]. One influential approach initially proposed by Byström and Järvelin [5] is to view task complexity through the lens of *a priori* determinability. The *a priori* determinability of a task is defined by the degree of uncertainty about the task's required outcomes and the processes involved in gathering the information needed to complete the task. A task with low determinability (i.e., high complexity) is one with high uncertainty about the form of the solution and the processes involved in solving the task.

In this work, we aimed to manipulate the determinability of search tasks indirectly, by manipulating the *scope* of the task (i.e., open-ended versus narrowly focused). In order to control for other task characteristics, we focused on comparative search tasks. For example, one of our tasks asked participants to compare different fertilizers for a home garden. Comparative tasks involve two important activities: (1) identifying different *items* or alternatives for the given category (e.g., organic, synthetic, liquid fertilizers) and (2) understanding how the items differ along different *dimensions* or attributes (e.g., cost, nutrient content, health concerns).

We manipulated the task scope by including or excluding specific items and/or dimensions for participants to consider as part of the comparative task. Our most open-ended tasks did not mention specific items or dimensions to consider. In contrast, our most narrowly focused tasks instructed participants to consider two specific items and one dimension. Additionally, we studied two types of dimensions: objective and subjective. We expected that addressing a *subjective* dimension would involve greater uncertainty (i.e., lower determinability). For example, a subjective dimension might require gathering information from different perspectives and evaluating the credibility of information.

We report on a crowdsourced study ($N$ = 144) that investigated the effects of our task manipulation on participants' perceptions about the task, search behaviors and strategies, and level of engagement during the task. We developed 12 task topics and 6 task versions per topic. Task version was our main independent variable and varied based on the specification of items and/or (objective


*SIGIR '18, July 8–12, 2018, Ann Arbor, Michigan, U.S.A.*

ACM ISBN 978-1-4503-4925-3/18/03…$15.00
https://doi.org/10.1145/3176349.3176380

or subjective) dimensions that should be considered during the comparative search task. We used a within-subject design; each participant completed six search tasks (one per task version). Our study investigates the following five research questions:

**RQ1 & RQ2:** What is the effect of task version on participants' pre-(RQ1) and post-task (RQ2) perceptions about the task? We focus on perceptions related to determinability, subjectivity, prior knowledge/knowledge increase, interest/interest increase, and expected/experienced difficulty.

**RQ3:** What is the effect of task version on participants' level of engagement during the task? We measured aspects of engagement using O'Brien's User Engagement Scale [18].

**RQ4:** What is the effect of task version on participants' search behaviors? We examined measures associated with search effort and the extent to which participants diverged from each other in their choice of queries and clicks.

**RQ5:** What is the effect of task version on participants' search strategies? Through a qualitative analysis of participants' queries, we investigated the differences in querying strategies observed for different task versions.

We build on our previous work to investigate the relationships between task scope, determinability, and searchers' perceptions and behaviors [8]. In this new study, we investigate differences between specifying objective and subjective dimensions in the task description, explore the effects of our task manipulation on user engagement, and present an analysis of participants' queries to explain *why* or *how* certain task versions affected task performance.

## 2 RELATED WORK

Our research sought to investigate the impacts of search task characteristics on search behaviors and outcomes, and to understand factors that influenced aspects of user engagement.

**Tasks and Task Characteristics:** People engage in information seeking and searching to complete a specific task. Byström and Hansen [4] argued that tasks can be defined at three levels of granularity: work tasks, information-seeking tasks, and information search tasks. An information search task is done within the context of an information-seeking task, and both are done within the context of a work task. In this paper, we manipulated tasks at the *search task* level (the most granular).

A large body of prior work has focused on characterizing search tasks along different dimensions. Li and Belkin [16] proposed a framework for characterizing search tasks along two dimensions: (1) generic facets (e.g., self-motivated vs. assigned) and (2) common attributes, which include objective attributes (e.g., task complexity) and subjective attributes (e.g., a user's domain knowledge).

**Task complexity:** Search task complexity is an important characteristic that has been found to influence search behaviors. Task complexity has been defined as an inherent property of the task (independent of the task doer), and has been manipulated from different perspectives [24]. Early work by Campbell characterized task complexity as a function of four different "sources": (1) the number of required outcomes, (2) the number of paths to the outcomes, (3) the level of uncertainty about the paths, and (4) the degree of interdependence between the paths [6]. In this respect, a highly complex task may involve many outcomes and paths, as well as high levels of uncertainty about the paths (e.g., some paths may be more effective) and interdependence between the paths (e.g., some paths may require progress on other paths).

Task complexity has also been studied from the perspective of *cognitive* complexity. Cognitive complexity is related to the amount of mental effort and/or learning required to complete a task. Jansen *et al.* [11] (and later Kelly *et al.* [13, 25]) used Anderson and Krathwohl's taxonomy of learning outcomes from educational theory [1] to create tasks with different levels of cognitive complexity. For example, the simplest tasks (called *remember* tasks) required verifying or searching for a specific fact, moderately complex tasks (called *analyze* tasks) required identifying items associated with a specific category and understanding their differences, and the most complex tasks (called *create tasks*) required finding a new solution to a problem. In this paper, we focus on comparative tasks, which fall under *analyze* tasks in the cognitive complexity framework.

More closely related to our work, Byström and Järvelin [5] (and later Bell and Ruthven [3]) reduced task complexity to *a priori* determinability. *A priori* determinability is concerned with the level of *uncertainty* associated with the task—a highly determinable task is one with low uncertainty. Byström and Järvelin [5] defined *a priori* determinability as the extent to which a searcher is able to internalize the task and deduce: (1) the required outcomes, (2) the information needed to produce the outcomes, and (3) the steps required to gather the needed information. Later, Bell and Ruthven [3] manipulated the *a priori* determinability of tasks in an laboratory study. Tasks were designed to influence the *a priori* determinability of: (1) the information needed to complete the task, (2) the strategy to search for and identify the needed information, and (3) the need to gather and synthesize information from different sources.

Wildemuth *et al.* [24] conducted an extensive literature review of ways in which search task complexity has been defined and/or manipulated. Based on this review, the authors proposed that task complexity involves three main components: (1) multiplicity of steps (i.e., complex tasks require more steps), (2) multiplicity of concepts (i.e., complex tasks involve more concepts and/or types of concepts), and (3) determinability (i.e., complex tasks have more uncertainty about the task goals and processes).

In a position paper, Toms [22] advocated for studying complexity from the *work task* perspective (i.e., the higher-level task motivating the need to search). From this perspective, complex tasks require processes beyond performing individual searches, including information extraction, analysis, comparison, prediction, modification, and manipulation. Toms presented a typology of tools to support users with some of these processes.

**Search behaviors and outcomes:** Past studies have also considered how search behaviors and outcomes are influenced by task complexity (using a specific definition). Studies have found that complex tasks are associated with greater levels of expected difficulty [3, 7, 10, 13, 25], experienced difficulty [2, 3, 7, 10, 13, 25], and search effort [2, 7, 10, 11, 13, 25]. Prior work also found that complex tasks are associated with a greater variety of query re-formulation types (e.g., adding, deleting, replacing, narrowing, and broadening terms or concepts) [23].

In prior work, we experimented with a similar manipulation of comparative tasks [8]. Results found the following trends. First, specifying items made the task easier in terms of pre-/post-task

perceptions and level of search effort. Second, specifying dimensions had no effect in terms of pre-task perceptions, but made the task more difficult in terms of post-task perceptions and level of search effort.

**User engagement:** Task characteristics may also affect searchers' engagement with the task. User engagement (UE) is characterized by the depth of a user's cognitive, emotional, and temporal investment during a search interaction [17]. Most UE research has focused on understanding how system characteristics (e.g., interface aesthetics, interactive features) and information content (e.g., multimedia, sentiment, interestingness) influence user engagement [19]. Less research has studied the influence of task characteristics on user engagement; our research attempts to address this gap.

Related to our work, the degree of "challenge" presented by a task has been highlighted as an important component of user engagement [20]. Specifically, users experience higher levels of engagement during interactions that carefully leverage their prior knowledge and skills—"easy" tasks may result in boredom, while "hard" tasks can result in frustration and disengagement. Kelly *et al.* [13] found that participants reported higher levels of engagement for tasks at higher levels of cognitive complexity, suggesting that the tasks used in their study did not exceed the threshold that leads to frustration and disengagement. In the current study, we hypothesized that our task manipulation would influence the level of challenge experienced by participants in completing the tasks, and would possibly influence the level of engagement.

## 3 TASK MANIPULATION

Our goal in this study was to manipulate the determinability of *comparative search tasks*. Comparative search tasks are ones in which a user needs to compare a set of *items* across a set of *dimensions*. During a comparative task, users must complete three steps: (1) identify different items belonging to the given category, (2) identify different dimensions along which the items may differ, and (3) understand how the items differ along the dimensions. For example, a comparative task might involve a user trying to compare different types of fertilizer for a home garden. In this case, a user might need to consider different items (i.e., types of fertilizer), such as organic, synthetic, and liquid fertilizers, as well as different dimensions, such as the fertilizer's cost, nutritional content, and health concerns.

Our task manipulation involved narrowing/broadening the *scope* of comparative tasks by including/excluding the exact items and/or dimensions to be considered as part of the task. Additionally, as a novel contribution of this paper, we were interested in comparing two types of dimensions: objective and subjective. We envisioned that an *objective* dimension (e.g., cost) would require gathering factual information, while a *subjective* dimension (e.g., health concerns) would require gathering information about people's feelings, perceptions, opinions, and/or experiences.

We used 12 task topics and 6 task versions per topic, for a total of 72 task descriptions. Each task description consisted of two parts: (1) a motivating background story that was consistent across all tasks with the same topic, and (2) an information request that was manipulated based on our six task versions. Below, we define our six task versions and provide examples. In the examples, the items and dimensions are shown in bold.

- **Unspecified (U):** no items or dimensions specified. "A friend of yours recently decided to quit smoking, and asked for your help in choosing a method. What are different methods to help people quit smoking and how do they differ?"
- **Items (I):** specified two items to compare, but no dimension. "A friend... How do **nicotine gum** and **nicotine patches** differ as methods to quit smoking?"
- **Objective dimension (O):** specified an objective dimension, but no items. "A friend... What are different methods to help people quit smoking and how do they differ in terms of their **average treatment length**?"
- **Subjective dimension (S):** specified a subjective dimension, but no items. "A friend... What are different methods to help people quit smoking and how do they differ in terms of **how difficult it is to stop the treatment**?"
- **Items + Objective dimension (IO):** specified two items to compare and one objective dimension. "A friend... How do **nicotine gum** and **nicotine patches** differ as methods to quit smoking in terms of their **average treatment length**?"
- **Items + Subjective dimension (IS):** specified two items to compare and one subjective dimension. "A friend... How do **nicotine gum** and **nicotine patches** differ as methods to quit smoking in terms of **how difficult it is to stop treatment**?"

Figure 1 illustrates our six task versions as a matrix to compare items (rows) and dimensions (columns). The unspecified (U) tasks had no items or dimensions specified and were therefore the broadest in scope. In this respect, we expected task version U to be the least *a priori* determinable (i.e., to have the most uncertainty about the task outcomes and processes involved). The IO and IS tasks were the most narrowly focused of the six task versions, and possibly the most *a priori* determinable. We also anticipated that specifying an objective dimension would involve more determinability than specifying a subjective dimension. Table 1 lists the different topics, items, objective dimensions, and subjective dimensions used for each of our 12 task topics.

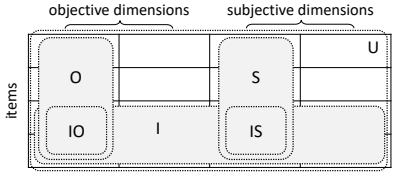



**Figure 1: Conceptual representation of comparative tasks**

## 4 METHOD

We conducted a within-subject user study ($N$=144) in which each participant completed six search tasks, one for each of our six task versions (U, I, O, S, IO, IS). Participants were asked to use a custom-built search system to find and bookmark pages that would be useful for addressing the task, and to provide a brief justification for why each bookmarked page was useful.

There were 12 task topics and 6 task versions (U, I, O, S, IO, IS) (see Section 3). These were fully crossed to create a total of 72 search tasks. Each individual participant was assigned 6 search tasks that included *all* 6 task versions and a subset of 6 of the 12 task topics. We used Latin squares to balance the order of presentation for both task versions and task topics.

The study was conducted using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. To accomplish the within-subjects

Table 1: Task topics, items, and dimensions used in our task descriptions.

| Topic | Items | Objective Dimension | Subjective Dimension |
|---|---|---|---|
| motor oil for cars | synthetic and organic oil | price range | cost-effectiveness |
| types of rice | white and brown rice | fiber content | noticeably affect insulin levels |
| types of ballet | classical and neo-classical | historical origin | difficulty of postures and movements |
| music speaker materials | polypropylene and paper | price | sound quality |
| garden fertilizers | organic and chemical | nutrient content | safety for growing vegetables |
| types of paint thinner | linseed and poppyseed oil | time for paint to dry | how well-suited for beginner |
| wifi routers | single band and dual band | signal interference | privacy and security issues |
| types of plastic | PET and PVC | how they can be recycled | risks involved in household use |
| indoor dog breeds | Pug and Bichon Frise | size at adulthood | ability to be left alone during the day |
| smoking cessation methods | nicotine gum and nicotine patch | average treatment length | difficulty to stop treatment |
| water purification methods | boiling and charcoal filter | micro-organisms eliminated | tradeoffs of safety and convenience |
| cooking skillet materials | aluminum and cast iron | how rapidly they heat up | what foods should/not be cooked |

design, each MTurk Human Intelligence Task (HIT) required participants to complete a full set of six search tasks on our server (external to MTurk). After completing all six search tasks, participants received a completion code that they used to receive payment within MTurk. Participants were paid $10 USD per HIT (six search tasks) and were only allowed to complete one HIT. To help ensure English language proficiency and quality control, we only recruited MTurk workers located in the U.S. with a $\geq$ 95% acceptance rate.

## 4.1 Study Protocol

In the preview description for our HITs, we informed participants about the study protocol and emphasized they would need to devote 70-90 minutes to complete all six search tasks. We also provided links to a video describing the HIT and to an informed consent form for the study. To complete the HIT, participants were required to add a set of "bookmarklet" buttons to their web browser toolbar. These buttons interfaced with our server and allowed participants to: (1) bookmark a page (including writing a justification of why the page was selected), (2) view (and possibly delete) the current set of bookmarks, (3) return to the most recent search engine results page (SERP), and (4) indicate when they were "done with the task". After accepting the HIT, participants' browsers were directed to a "main page" on our server that allowed them to start the first/next search task, track their progress on the six tasks, and resume a task if they encountered a problem (e.g., accidentally closing their browser).

For each search task, participants completed the same procedure. First, they were shown the task description and given instructions:

> "Your goal is to do a thorough search for information to address the task. You don't need to produce a written response to the task. Instead, bookmark pages that could be used to create an accurate and comprehensive response. For each bookmark, explain why the page is useful AND how it relates to the other information you have already found. Bookmark as many pages as you think are needed."

After reading the task description, participants completed a pre-task questionnaire (Section 4.2), and were then directed to a custom-built search engine that returned results using the Bing Web Search API. The task description remained visible at the top of the SERP. Participants were free to interact with the search system as they wished, and to use the toolbar buttons to bookmark pages (adding a justification) and to view or delete the current of set of bookmarks. After clicking the "done with the task" toolbar button, participants were directed to the User Engagement Scale (UES) questionnaire followed by a second post-task questionnaire (Section 4.2).

## 4.2 Questionnaires

Participants completed a pre-task questionnaire before starting each search task (denoted as *PreTask*), and two questionnaires after completing each task: a short form of O'Brien's User Engagement Scale [18] (denoted as *UES-SF*) and a post-task counterpart to the pre-task questionnaire (denoted as *PostTask*). All questions asked participants to report their level of agreement on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree).

**PreTask and PostTask Questionnaires:** The PreTask and PostTask questionnaires were designed to be counterparts of each other—both questionnaires consisted of the same or similar statements. The 14 questions in the PreTask and PostTask questionnaires (Table 2) may be categorized according to five main themes: (1) prior knowledge/knowledge increase, (2) interest/interest increase, (3) expected/experienced difficulty, (4) determinability, and (5) subjectivity. The first three aspects (knowledge, interest, difficulty) are commonly included in IIR studies and based on prior work [12], we included one question about each. Determinability and subjectivity are more novel measures and to investigate them, we designed and included 11 additional questions.

**User Engagement Scale (UES-SF):** Recently, O'Brien *et al.* [18] developed and tested a briefer version of the 31-item User Engagement Scale (UES). The UES-Short Form (UES-SF) (Table 3) contained 12 questions designed to capture four dimensions of engagement (3 questions per dimension): (1) focused attention (FA), (2) perceived usability (PU), (3) aesthetic appeal (AE), and (4) reward (RW).

# 5 DATA ANALYSIS AND MEASURES

Out of the 864 total task sessions (144 participants x 6 tasks), we omitted 25 sessions based on missing or careless responses (based on "attention check" questions that we included in the questionnaires). Shapiro-Wilk tests revealed that the PreTask, PostTask, UES-SF questionnaire data were not normally distributed. In subsequent statistical analyses, we used methods to account for this non-normality (e.g., log transformation for our ANOVAs; Principle Axis Factoring).

**Pre- and Post-Task Measures:** We measured knowledge, interest, and difficulty in the PreTask and PostTask questionnaires using a single question about each measure (Section 4.2). To measure determinability and subjectivity, we included 11 questions. To test the grouping of the questions into these two dimensions, we performed Principle Axis Factoring (PAF) with Direct Oblimin rotation. This analysis identified one factor in the PreTask and PostTask questionnaires that included seven items related to the *determinability* of the task (exactitems, exactdims, details, specificity, narrow, lookfor,

**Table 2: PreTask and PostTask Questionnaires.**

| Measure | Pre-task question | Post-task question |
|---|---|---|
| knowledge | I already know a lot about this topic. | My knowledge of this topic has increased. |
| interest | I am interested in the topic. | My interest in the topic has increased. |
| difficulty | I think the task will be difficult. | The task was difficult. |
| specificity | The task is specific. | The task was specific. |
| narrow | The information requested in narrowly focused. | The information requested was narrowly focused. |
| newinfo | The task description provides me with new information that I did not already know about this topic. | The task description provided me with new information that I did not already know about this topic. |
| figureout | I will need to figure out things that are not specified in the task description. | I needed to figure out things that were not specified in the task description. |
| details | The task description has details that will help me complete the task. | The task description had details that helped me complete the task. |
| lookfor | Right now, I know some specific things to look for to address the task. | I knew some specific things to look for to address the task. |
| exactitems | The task description tells me exact items that I need to compare. | The task description told me exact items I needed to compare. |
| exactdims | The task description tells me exact criteria that I need to consider in understanding the differences between items. | The task description told me exact criteria I needed to consider when comparing items. |
| openended | The task is open-ended. | The task was open-ended. |
| opinions | The task will require gathering information regarding people's feelings, tastes, and/or opinions. | The task required gathering information regarding people's feelings, tastes, and/or opinions. |
| factualinfo | The task will require gathering factual information. | The task required gathering factual information. |

**Table 3: User Engagement Scale (Short Form)**

| |
|---|
| FA1: I lost myself in this search experience. |
| FA2: The time I spent searching just slipped away. |
| FA3: I was absorbed in the search task. |
| PU1: I felt frustrated while doing the search task. |
| PU2: My search experience was taxing. |
| PU3: I found the search system confusing to use. |
| AE1: The search system was attractive. |
| AE2: The search system was aesthetically appealing. |
| AE3: The search system appealed to my senses. |
| RW1: My search experience was worthwhile. |
| RW2: My search experience was rewarding. |
| RW3: I felt interested in the search task. |

and newinfo). Furthermore, the PAFs identified a second factor related to the *subjectivity* of the task that had two items in common between the PreTask and PostTask questionnaires (opinions, openended).[1]

We examined the internal consistency of the identified PreTask and PostTask determinability and subjectivity factors using Cronbach's alpha. The determinability factor had acceptable pre- (0.77) and post-task (0.81) Cronbach's alpha values [9]. Cronbach's alpha values of the subjectivity factor were lower (0.48, 0.54), but the items loaded sufficiently ($> 0.45$) and fit conceptually with the construct of subjectivity. As a result, we averaged each participant's responses to the items for each of these factors to create pre- and post-task determinability and subjectivity scores for each task.

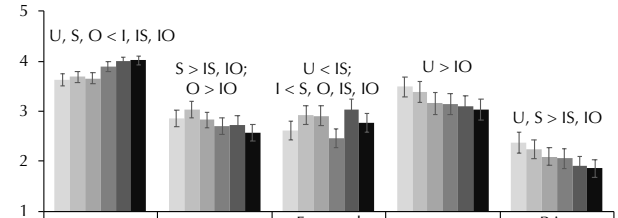
**User Engagement Measures:** We performed PAF with Direct Oblimin rotation and found the expected four-factor solution for the UES-SF (three items per factor) [18]. All item loadings were $\geq .51$. Thus, we averaged participants' responses for each three-item factor. The perceived usability (PU) items were reversed-coded.

# 6 RESULTS

We present results in terms of our five research questions (RQ1-RQ5). For RQ1-RQ4, we conducted repeated-measures ANOVAs to analyze the effect of task version on each measure related to participants' pre-task perceptions (RQ1), post-task perceptions (RQ2), level of engagement (RQ3), and search behaviors (RQ4).[2] We used Bonferroni-corrected post-hoc tests to compare all pairs of task versions for each measure. For RQ5, we present a *qualitative* analysis of participants' querying strategies across task version.

## 6.1 RQ1: Pre-task perceptions

Figure 2 shows the means and 95% confidence intervals of participants' responses across task versions for the pre-task measures.




| | Determin. | Subjectivity | Expected Difficulty | Interest | Prior Knowledge |
|---|---|---|---|---|---|
| U | 3.617 | 2.856 | 2.619 | 3.489 | 2.374 |
| S | 3.690 | 3.029 | 2.929 | 3.393 | 2.243 |
| O | 3.658 | 2.832 | 2.907 | 3.157 | 2.093 |
| I | 3.891 | 2.711 | 2.457 | 3.150 | 2.057 |
| IS | 3.998 | 2.732 | 3.043 | 3.107 | 1.907 |
| IO | 4.022 | 2.568 | 2.771 | 3.036 | 1.857 |

**Figure 2: Mean (95% conf. int.) of participants' pre-task perceptions across task versions. Significant (Bonferroni-corrected) pairwise differences are displayed above each factor ($p < .05$).**

**Pre-task determinability**: Task version had a significant effect on pre-task determinability ($F(4.103, 566.171) = 20.431$, $p = .000$). Participants reported greater levels of determinability for task versions that specified the items (I, IS, IO) as compared to task versions that did not specify the items (U, O, S). Post-hoc tests found significantly higher levels of determinability for task versions I, IS, and IO as compared to U, O, and S ($p < .05$). The observed trend is that specifying items in the task description made the task more determinable, and that specifying an objective or subjective dimension had no effect.

**Pre-task subjectivity**: Task version had a significant effect on pre-task subjectivity ($F(5, 690) = 6.518$, $p = .000$). Figure 2 shows two main trends. First, *excluding* items from the task description increased the level of pre-task subjectivity. Post-hoc tests found significantly higher levels of subjectivity for task version S as compared to IS and for task version O as compared to IO ($p < .05$).

[1] After four iterations of PAF on the PreTask questionnaire data, we arrived at a two-factor structure that explained 48.42% of the variance (KMO=0.83; $\chi^2 = 1948.875(45)$, $p = .000$). Based on this PreTask solution, we specified a two-factor solution for the PostTask questionnaire. This explained 52.26% of the variance (KMO=0.849; $\chi^2 = 2383.165(45)$, $p = .000$).

[2] In cases where the assumption of sphericity was violated, we applied the Greenhouse-Geisser correction to the degrees of freedom.

Second, specifying a subjective dimension in the task description had a *slight* increase in the level of pre-task subjectivity. While the differences did not reach statistical significance, participants reported higher levels of subjectivity for task version S as compared to O and for task version IS as compared to IO.

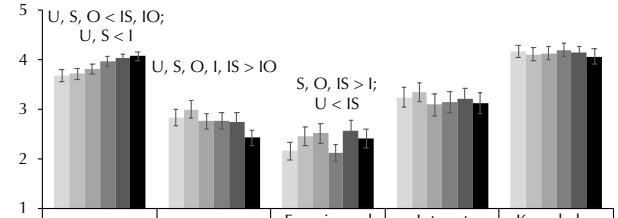
**Pre-task difficulty**: Task version had a significant effect on pre-task difficulty ($F(5, 690) = 7.071$, $p = .000$). The main trend observed in Figure 2 is that participants reported higher levels of expected difficulty when the task description specified an objective or subjective dimension. Post-hoc tests found significantly higher levels of expected difficulty for task version IS as compared to U, and for task versions S, O, IS, and IO as compared to I ($p < .05$).

**Pre-task interest**: Task version had a significant effect on pre-task interest ($F(5, 690) = 3.832$, $p = .002$). The main trend observed in Figure 2 is that participants reported *lower* levels of interest when the task description specified the items and/or a (subjective or objective) dimension. Post-hoc tests found significantly higher levels of interest for task version U as compared to IO ($p < .05$). Thus, participants reported higher levels of interest for the most open-ended task version (U) as compared to the most narrowly focused (IO). One possible explanation is that participants perceived task version U as allowing them to explore their own interests.

**Prior knowledge**: Task version had a significant effect for prior knowledge ($F(5, 690) = 6.441$, $p = .000$). As with pre-task interest, the main trend observed in Figure 2 is that participants reported *lower* levels of prior knowledge when the task description specified the items and/or a (subjective or objective) dimension. Post-hoc tests found significantly higher levels of prior knowledge for task versions U and S as compared to IS and IO ($p < .05$). One possible explanation is that the most narrowly-focused task versions (IS and IO) restricted the types of prior knowledge that might be useful for addressing the task.

## 6.2 RQ2: Post-task perceptions

Figure 3 shows the means and 95% confidence intervals of participants' responses across task versions for the post-task measures.




| | Determin. | Subjectivity | Experienced Difficulty | Interest Increase | Knowledge Increase |
|---|---|---|---|---|---|
| U | 3.692 | 2.838 | 2.173 | 3.252 | 4.180 |
| S | 3.727 | 3.004 | 2.464 | 3.357 | 4.114 |
| O | 3.823 | 2.768 | 2.521 | 3.100 | 4.136 |
| I | 3.981 | 2.771 | 2.136 | 3.150 | 4.207 |
| IS | 4.035 | 2.764 | 2.579 | 3.229 | 4.157 |
| IO | 4.083 | 2.436 | 2.421 | 3.136 | 4.071 |

**Figure 3: Mean (95% conf. int.) of participants' post-task perceptions across task versions. Significant (Bonferroni-corrected) pairwise differences are displayed above each factor ($p < .05$).**

**Post-task determinability:** Task version had a significant effect on post-task determinability ($F(4.161, 574.201) = 19.484$, $p = .000$). As with pre-task determinability, participants reported greater levels of determinability for task versions that specified the items (I, IS, IO) as compared to task versions that did not specify the items (U, O, S). Post-hoc tests found significantly higher levels of determinability for task versions IO and IS as compared to U, O, and S, and for task version I as compared to U and S ($p < .05$). The observed trend is that specifying items in the task description made the task more determinable, and that specifying an objective or subjective dimension had no effect.

**Post-task subjectivity:** Task version had a significant effect on post-task subjectivity ($F(5, 690) = 8.877$, $p = .000$). There were significantly lower levels of subjectivity for the task version that included items and an objective dimension (IO) compared to *all* other task versions ($p < .05$).

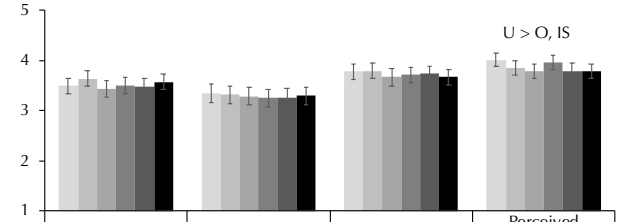
**Post-task difficulty:** Task version had a significant effect on post-task difficulty ($F(5, 690) = 4.239$, $p = .001$). As with (pre-task) expected difficulty, participants reported higher levels of (post-task) experienced difficulty when the task description specified an objective or subjective dimension. Post-hoc tests found higher levels of experienced difficulty for task version IS as compared to U, and for task versions S, O, and IS as compared to I ($p <= .05$).

**Post-task interest increase:** Task version did not have a significant effect on interest ($F(4.611, 636.343) = 1.590$, $p = .166$). Though not significant, participants reported slightly greater levels of increased interest when the task description specified a subjective versus objective dimension. This trend can be observed by comparing task version S vs. O and task version IS vs. IO.

**Post-task knowledge increase:** Task version did not have a significant effect on post-task knowledge ($F(4.478, 617.965) = .919$, $p = .460$). The trend observed in Figure 3 is that knowledge increase was fairly high (around 4.0 on a 5-point scale) and fairly consistent across task versions.

## 6.3 RQ3: Post-task Engagement

Figure 4 shows the means and 95% confidence intervals of participants' responses across task versions for each engagement factor.




| | Focused Attention | Aesthetic Appeal | Reward | Perceived Usability |
|---|---|---|---|---|
| U | 3.489 | 3.348 | 3.775 | 4.010 |
| S | 3.636 | 3.314 | 3.795 | 3.845 |
| O | 3.424 | 3.288 | 3.671 | 3.788 |
| I | 3.490 | 3.252 | 3.710 | 3.960 |
| IS | 3.479 | 3.260 | 3.736 | 3.786 |
| IO | 3.576 | 3.293 | 3.667 | 3.783 |

**Figure 4: Mean (95% conf. int.) of participants' post-task engagement factors across task versions. Significant (Bonferroni-corrected) pairwise differences are displayed above each factor ($p < .05$).**

Task version did not have a significant effect on focused attention ($F(4.495, 620.311) = 2.141$, $p = .066$), aesthetic appeal ($F(3.889, 536.707) = .924$, $p = .447$), and reward ($F(4.359, 601.524) = 1.017$, $p = .401$), but did have a significant effect on perceived usability ($F(4.601, 634.946) = 3.024$, $p = .013$). Post-hoc tests found that participants reported significantly higher levels of perceived usability for task version U as compared to O and IS ($p < .05$). The difference between task version U and IO was marginally significant ($p = .070$). The main trend observed in Figure 4 is that specifying an objective or subjective dimension in the task description (S, O, IS, OS) resulted in lower levels of perceived usability.

## 6.4 RQ4: Search Behaviors

To analyze the effects of task version on participants' search behaviors, we computed nine measures associated with search effort: (1) number of queries, (2) average query length (in words), (3) number of clicks, (4) number of clicks per query, (5) number of abandoned queries, (6) number of bookmarks, (7) number of queries without a bookmark, (8) number of clicks without a bookmark, and (9) task completion time. Additionally, we computed three measures associated with the extent to which participants' search strategies differed from all the other participants who completed the same combination of task topic and task version. Our three divergent search strategy measures included: (1) number of unique queries (not issued by any other participant), (2) number of unique query terms (not used by any other participant), and (3) number of unique URLs clicked on a SERP (not clicked by any other participant). Figure 5 shows the means and 95% confidence intervals for each measure across task version.

Task version had a significant effect on 11 of the 12 measures (Figure 5):

- queries ($F(4.254, 587.120) = 6.519$, $p = .000$)
- query length ($F(4.412, 608.792) = 6.608$, $p = .000$)
- clicks ($F(3.976, 548.649) = 2.824$, $p = .025$)
- clicks per query ($F(5, 690) = 3.755$, $p = .002$)
- abandoned queries ($F(4.002, 552.276) = 3.162$, $p = .014$)
- bookmarks ($F(4.658, 642.756) = 3.527$, $p = .005$)
- queries w/o a bookmark ($F(4.093, 564.883) = 4.858$, $p = .001$)
- clicks w/o a bookmark ($F(3.756, 518.385) = 2.836$, $p = .027$)
- unique queries ($F(4.273, 589.728) = 7.856$, $p = .000$)
- unique query terms ($F(3.879, 535.261) = 2.985$, $p = .020$)
- unique SERP clicks ($F(4.154, 573.279) = 4.515$, $p = .001$)

Task version did not have a significant effect on the task completion time ($F(3.757, 518.404) = 1.102$, $p = .354$).

The results in Figure 5 show four important trends. First, the unspecified task version (U) required the least amount of search effort. While completing task version U, participants issued fewer and shorter queries; had fewer clicks and more clicks per query; had fewer abandoned queries; and had fewer queries and clicks without a bookmark. Moreover, while completing task version U, participants adopted the most similar strategies to each other. Specifically, task version U was associated with the least number of unique queries, unique query terms, and unique URLs clicked on a SERP. One possible explanation is that participants satisficed when completing task version U and did not explicitly seek information about specific items and/or dimensions.

The second important trend is that *only* specifying the items in the task description (task version I) resulted in low levels of effort and few unique behaviors. As shown in Figure 5, there were no significant differences between task versions I and U across any of the 12 measures. One possible explanation is that items tend to be concrete (rather than abstract) concepts (e.g., synthetic oil, white rice, charcoal filter), which may be referred to using consistent terminology and are therefore easy to include in queries and identify in relevant documents.

Third, specifying an objective or subjective dimension in the task description (task versions S, O, IS, and IO) resulted in greater levels of search effort and divergent search strategies. This trend can be observed in Figure 5 by comparing task versions S, O, IS, and IO with task versions U and I. We speculate that both objective and subjective dimensions tended to be abstract (rather than concrete) concepts (e.g., cost effectiveness, suitability for a beginner, difficulty to stop treatment), which can be referred to in many different ways. Thus, compared to items, dimensions may have been more difficult to express in queries and identify in relevant documents.

Finally, participants expended slightly more effort and had more divergent behaviors when the task description included a subjective versus objective dimension (Figure 5). This trend can be observed by comparing task version S versus O, and IS versus IO. While the differences were not significant, task version S had longer queries than task version O. Similarly, task version IS had more clicks, abandoned queries, queries/clicks without a bookmark, and longer completion times than task version IO. Lastly, task version IS had more unique queries, query terms, and SERP clicks than task version IO. A possible explanation of this trend is that subjective dimensions required gathering and synthesizing information from different sources, as well as judging credibility.

## 6.5 RQ5: Search Strategies

As part of RQ5, we consider whether participants employed different search strategies across task versions. To address this question, we performed qualitative coding of all queries issued by participants. Specifically, two of the authors coded the queries based on the presence of items and dimensions (i.e., a two-dimensional coding scheme). With respect to items, each query was assigned a code of 'I' if it contained at least one item and a code of '*' otherwise. Similarly, with respect to dimensions, each query was assigned a code of 'D' if it contained at least one dimension and a code of '*' otherwise. Table 4 illustrates a few example queries and their assigned codes. We refer to (*,*) queries as having the broadest intent, (I,*) and (*,D) queries as having a narrower intent, and (I,D) queries has having the narrowest intent. Initially, both authors independently coded a common set of approximately 10% of all search sessions. Cohen's Kappa was $\kappa = .966$ for items (i.e., codes 'I' vs. '*'), and $\kappa = .983$ for dimensions, (i.e. codes 'D' vs. '*'). Given this high level of agreement, the remaining 90% of all search sessions were coded independently (45% per coder) to complete the dataset.

We investigated the effect of task version on participants' search strategies from two perspectives. First, we considered whether task version influenced the types of queries issued by participants. For example, were participants more likely to issue (*,*) queries for task version U than for other task versions? Second, we examined whether task version influenced participants to issue *multiple* types

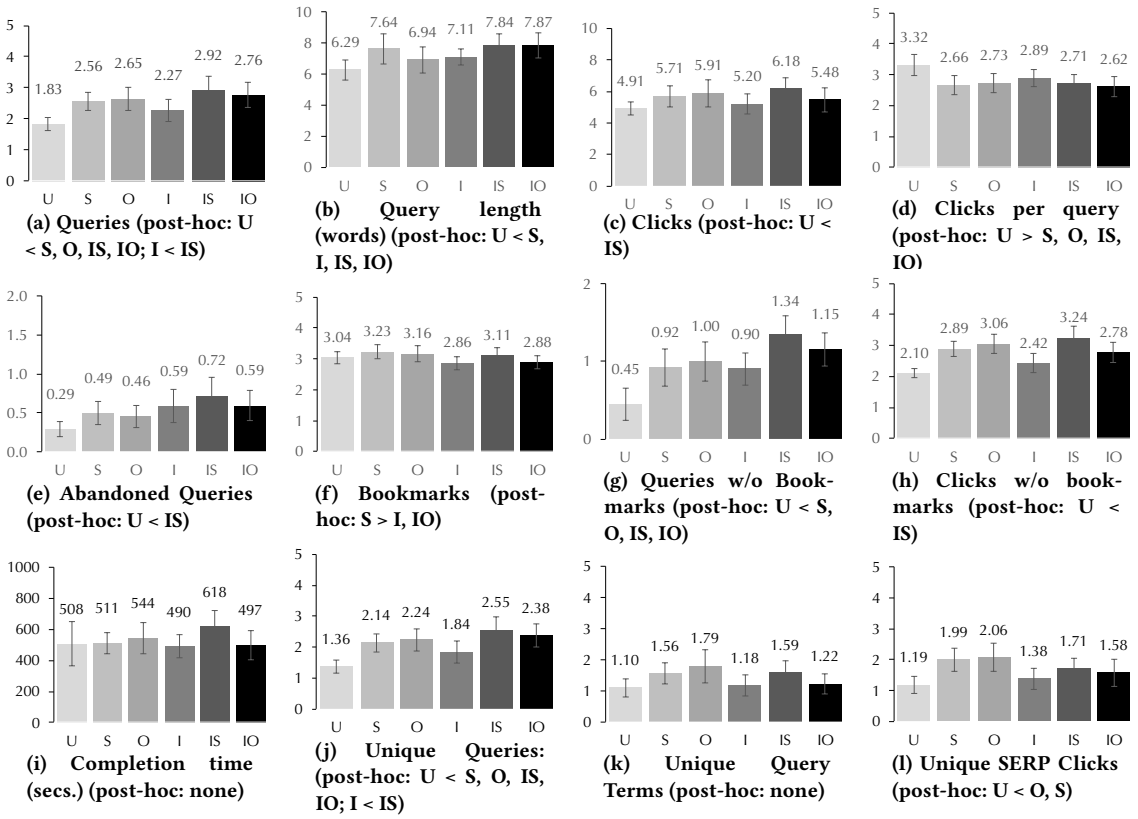



**Figure 5: The effects of task version on search behaviors. Bonferroni-corrected pairwise differences are displayed below each measure ($p < .05$).**

**Table 4: Example codes based on items and dims.**

| Query | Item(s) | Dim(s) |
|---|---|---|
| methods to quit smoking | * | * |
| how do smoking cessation methods differ | * | * |
| **nicotine gum** vs **patches** | I | * |
| *average treatment length* in quitting smoking | * | D |
| *average* **nicotine patch** *treatment length* | I | D |

**Table 5: Query-type distribution per task version. Values are macro-averaged across search sessions (Mean, SD). Highlighted cells show most frequent query-types per task version.**

| | (*,*) | (I,*) | (*,D) | (I,D) |
|---|---|---|---|---|
| U | 0.937 (0.192) | 0.020 (0.092) | 0.044 (0.171) | 0.000 (0.000) |
| S | 0.413 (0.368) | 0.004 (0.026) | 0.558 (0.371) | 0.025 (0.111) |
| O | 0.407 (0.365) | 0.012 (0.061) | 0.567 (0.368) | 0.013 (0.059) |
| I | 0.037 (0.114) | 0.915 (0.211) | 0.002 (0.023) | 0.046 (0.179) |
| IS | 0.018 (0.081) | 0.417 (0.406) | 0.051 (0.176) | 0.513 (0.420) |
| IO | 0.016 (0.101) | 0.318 (0.396) | 0.035 (0.132) | 0.631 (0.403) |

of queries during the same search session. For example, were participants more likely to switch strategies (i.e., by broadening or narrowing queries) for certain task versions than others?

**Query Type Distribution:** Table 5 shows the query-type distribution per task version. The values indicate the fraction of queries of each type issued by participants during a specific task version (rows sum to one).

Table 5 shows the following trends. First, for task version U, participants issued an overwhelming proportion of (*,*) queries; items and/or dimensions were very rarely included in queries for task version U. Second, for task versions S and O, participants frequently issued two types of queries (i.e., (*,*) and (*,D)). As might be expected, (*,D) queries were slightly more frequent. Third, for task version I, participants issued an overwhelming proportion of (I,*) queries. Participants very rarely included dimensions in queries for task version I. Finally, for task versions IS and IO, participants frequently issued two types of queries (i.e., (I,*) and (I,D)). Again, as might be expected, (I,D) queries were slightly more frequent.

To summarize, for task versions U and I, participants mostly adopted one strategy—issuing (*,*) and (I,*) queries, respectively. For task versions S, O, IS, and IO, participants adopted two strategies—issuing (*,*) and (*,D) queries for task versions S and O, and issuing (I,*) and (I,D) queries for task versions IS and IO. These results suggest that querying for dimensions was not trivial. For task versions S, O, IS, and IO, participants often issued queries that were broader than the scope of the task and ignored the dimension (i.e., (*,*) queries for S and O, and (I,*) queries for IS and IO).

**Multiple Strategies:** The previous results showed that participants frequently issued two different types of queries for task versions that included a dimension (O, S, IO, and IS). We investigated whether this trend was caused by participants adopting one strategy or the other, or by *switching* between strategies within the same session. Table 6 shows the percentage of search sessions with $n$ type(s) of queries (i.e., (*,*), (I,*), (*,D), (I,D)) , where $n = 1, 2, 3, 4$.

**Table 6: Percentage of participants who issued *n* types of queries during the same session.**

| | *n*=1 | *n*=2 | *n*=3 | *n*=4 |
|---|---|---|---|---|
| U | 89.21% | 10.07% | 0.72% | 0.00% |
| S | 48.57% | 47.14% | 4.29% | 0.00% |
| O | 50.36% | 41.73% | 7.19% | 0.72% |
| I | 84.29% | 14.29% | 0.71% | 0.71% |
| IS | 59.29% | 37.14% | 2.86% | 0.71% |
| IO | 64.29% | 30.71% | 5.00% | 0.00% |

As expected, for task versions U and I (no dimension), participants typically did not switch strategies. Conversely, for tasks S, O, IS, and IO, strategy switching was more common—between 36-50% of participants switched querying strategies during the same session.

## 7 DISCUSSION

Our results showed complex relationships between task scope, determinability, and search behaviors and experiences based on our task manipulation.

**Unspecified Tasks:** Our unspecified (U) tasks did not specify items or dimensions and were the broadest in scope. We expected these tasks to have the lowest determinability (i.e., greatest uncertainty about the outcomes and processes) and to be the most challenging. Our results found a different outcome. While task version U was rated as having the *lowest* determinability before and after the task, it was also the easiest. Task version U had the least amount of search activity (e.g., queries, clicks), trial-and-error (e.g., abandoned queries, clicks without a bookmark), lowest expected and experienced difficulty, and highest ratings for the perceived usability factor of engagement. In addition, task version U also had the least amount of diversity in the search strategies adopted by participants. For example, task version U had the lowest number of unique queries, query terms, and SERP clicks. Furthermore, 94% of all the task U queries were of type (*,*) (no items or dimensions), and 89% of all task U search sessions *only* had (*,*) queries.

These results suggest that participants completed task version U through satisficing or self-defining a task scope that would reduce the uncertainty of the task. Participants did not attempt to cover the entire space of items and dimensions and did not deeply explore specific items and/or dimensions.

Our results for task version U indicate that an open-ended or broad task can be perceived to have low determinability, but may turn out to be a low-complexity task that requires little effort and creativity. The goals and motivations of our MTurk participants' may have played a role in this result. Unspecified tasks might result in very different trends if they were self-generated tasks, or if they were situated in scenarios that might influence participants to explore the unspecified space more comprehensively.

**Effects of specifying items:** Specifying items in the task description (e.g., U→I, O→IO, S→IS) made the task narrower in scope and significantly more determinable (pre- and post-task). In terms of search effort, task version I (only items) was not significantly different from task version U (the easiest) across almost every measure. The same was true for pre- and post-task difficulty.

These results suggest that querying for items was easy for participants. Similar to task version U, task version I had a low number of unique queries and query terms. Additionally, 91% of queries were of type (I,*) (items, no dimensions), and 84% of all search sessions *only* had (I,*) queries. One possible explanation is that items tend to be concrete concepts, which have specific names that are easy to include in queries and to recognize in relevant documents.

This suggests that specifying items in the task description increased the task determinability by reducing the number of outcomes (narrowing the scope) as well as reducing the uncertainty associated with some of the search processes. That is, the items provided noun-phrases that could be used to generate effective queries and identify relevant documents. These effects of specifying items are consistent with prior work [8].

**Effects of specifying dimensions:** Specifying a dimension (objective or subjective) in the task description (e.g., U→O, U→S, I→IO, I→IS) had very different effects than specifying items. Similar to the items, specifying a dimension also made the task narrower in scope. However, specifying a dimension did not influence participants to perceive the task as more determinable and it made the task *more* challenging. Tasks versions that included dimensions (O, S, IO, IS) were perceived to be more difficult (pre- and post-task), required more search effort (e.g., queries, clicks), more trial-and-error (e.g., abandoned queries, clicks without a bookmark), and had more diverse search behaviors (e.g., more unique queries, query terms, and SERP clicks). Furthermore, while the results were not significant, participants' ratings for the perceived usability factor of engagement were lower for tasks that specified a dimension.

Our results suggest that querying for dimensions was challenging for participants. For example, for task versions O and S, participants issued a large proportion of (*,*) queries, which were broader in scope than the task version and ignored the dimension. Similarly, for task versions IO and IS, participants issued a large proportion of (I,*) queries, which also ignored the dimension. These trends suggest that participants either had difficulty expressing the dimension in their queries or that they found that leaving out the dimension was a better strategy for finding relevant documents. One possible explanation is that dimensions tend to be abstract concepts that can be expressed in a variety of ways, making it difficult to include them in queries and identify them in relevant documents. Moreover, the varied language surrounding dimensions may also widen the vocabulary gap between queries that mention dimensions and relevant documents. The observed effects of specifying dimensions are also consistent with prior work [8]. The query analysis presented as part of RQ5 allowed us to gain insights about how and why the dimensions made the task more challenging.

**Effects on Engagement:** Our task manipulation did not have strong effects on user engagement. Task version only had a significant effect for perceived usability, with participants reporting greater levels of perceived usability for task version U (the most open-ended). As previously mentioned, challenge is one important factor that may influence user engagement [20]. One possible explanation is that our task versions were neither too easy to cause boredom nor too difficult to cause frustration and disengagement.

Our results do suggest a potentially interesting relationship between task scope and engagement. Participants reported significantly greater levels of pre-task interest and post-task perceived usability for task version U than some of the other more narrowly-focused task versions. A possible explanation is that open-ended tasks allowed participants to explore their own interests.

**Objective vs. Subjective Dimensions:** We did not observe strong differences between specifying an objective vs. subjective dimension in the task description. Across all measures, this trend can be observed by comparing task version O versus S and IO versus IS. While we did not find statistically significant differences between O and S (nor between IO and IS) for our measures, a few trends are worth noting. First, participants reported greater levels of pre- and post-task subjectivity for S vs. O and IS vs. IO, suggesting that they recognized that the subjective dimension added a degree of subjectivity to the task. Second, comparing IS vs. IO, the subjective dimension increased the level of search effort (e.g., queries, clicks, abandoned queries, bookmarks, query/clicks without a bookmark, and task completion time). Finally, our RQ5 results show that there were fewer dimensions included in queries for S vs. O and IS vs. IO.

In prior work [8], we used dimensions that tended to be subjective. Thus, we speculated that dimensions might have made the task more challenging due to a need to consult different perspectives. In this study, we directly compared subjective and objective dimensions. Both dimension types made the task more challenging and subjective dimensions required slightly more effort.

## 8 CONCLUSIONS

We conducted a large-scale user study in which we manipulated the scope of comparative search tasks by varying the inclusion of specific items and (objective/subjective) dimensions. Our results found several trends. First, our unspecified tasks (most open-ended) were rated as the least determinable, but were the easiest in terms of search effort due to participants satisficing or self-defining a narrower task scope. Second, adding items to the task description increased participants' perceptions of determinability and made the task easier. Third, adding dimensions did not increase perceptions of determinability and made the task more difficult. A qualitative analysis of participants queries suggests that querying for dimensions is challenging for users. Lastly, we observed that subjective dimensions required slightly more effort than objective dimensions.

Our findings have implications for the design of search tasks in IIR studies and search tools to support users. With respect to task design, our results illustrated the complex relationship between task scope, determinability, and difficulty. A task may have a broad scope and low determinability based on the lack of details specified in the task description, but lend itself to satisficing and may ultimately become an easy task. Conversely, a task may have a narrow scope and high determinability based on the details specified in the description, but actually be a difficult task.

Task complexity has been viewed in terms of the number of task concepts (complex tasks have more concepts or concept-types) [24]. From this perspective, our results found that different concept-types can impact tasks in very different ways. In our case, specifying items made tasks *less* complex, while specifying a dimension made tasks more complex by increasing the level of uncertainty regarding the paths toward the solution. IIR experimenters are advised to consider the determinability of tasks in terms of both the outcomes *and* the processes involved, and to keep satisficing behaviors in mind.

Our results also suggest opportunities to develop tools to support users. One of the clearest findings from our study is that participants had difficulty querying for dimensions. This presents an opportunity to develop tools to support searchers. One possible approach is to incorporate dimensions into query suggestions. A second approach is to include dimensions as facets in a faceted search environment. Existing algorithms for dynamic facet prediction [14] could be extended to infer dimensions associated with the current search task. In the area of faceted search, prior work has shown that facet-values can help users advance a search even if the facets are not explicitly used [15]. Finally, one could imagine an application where a user adds information into a grid-like interface of items and dimensions, and the system attempts to complete the grid based on partial information.

**Acknowledgements:** This work was supported in part by SSHRC grant F15-04687 and NSF grants IIS-1552587 and IIS-1451668. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

## REFERENCES

[1] Lorin W. Anderson and David R. Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.*
[2] Jaime Arguello. 2014. Predicting Search Task Difficulty. In *ECIR*. Springer, 88–99.
[3] David J. Bell and Ian Ruthven. 2004. Searchers' Assessments of Task Complexity for Web Searching. In *ECIR*. Springer, 57–71.
[4] Katriina Byström and Preben Hansen. 2005. Conceptual framework for tasks in information studies. *JASIST* 56, 10 (2005), 1050–1061.
[5] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information Processing & Management* 31, 2 (1995), 191–213.
[6] Donald J. Campbell. 1988. Task Complexity: A Review and Analysis. *The Academy of Management Review* 13, 1 (1988), 40–52.
[7] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. 2015. Differences in the Use of Search Assistance for Tasks of Varying Complexity. In *SIGIR*. ACM, 23–32.
[8] Robert Capra, Jaime Arguello, and Yinglong Zhang. 2017. The Effects of Search Task Determinability on Search Behavior. In *ECIR*. Springer, 108–121.
[9] Robert F DeVellis. 2016. *Scale development: Theory and applications*. Vol. 26. Sage.
[10] Xiao Hu and Noriko Kando. [n. d.]. Task complexity and difficulty in music information retrieval. *JASIST* 68, 7 ([n. d.]), 1711–1723.
[11] Bernard J. Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *IPM* 45, 6 (2009), 643–663.
[12] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.
[13] 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 *ICTIR*. ACM, 101–110.
[14] Weize Kong and James Allan. 2013. Extracting Query Facets from Search Results. In *SIGIR*. ACM, 93–102.
[15] Bill Kules and Robert Capra. 2012. Influence of Training and Stage of Search on Gaze Behavior in a Library Catalog Faceted Search Interface. *JASIST* 63, 1 (2012), 114–138.
[16] Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *IP&M* 44, 6 (2008), 1822 – 1837.
[17] Heather L O'Brien. 2016. Theoretical perspectives on user engagement. In *Why Engagement Matters*. Springer, 1–26.
[18] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies* (2018).
[19] Heather L O'Brien and Jocelyn McKay. 2018. Modeling antecedents of engagement. In *The Handbook of Communication Engagement*. Wiley.
[20] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the Association for Information Science and Technology* 59, 6 (2008), 938–955.
[21] Eliane G. Toms. 2011. Task-based information searching and retrieval. In *Interactive information seeking, behaviour and retrieval*, Ian Ruthven and Diane Kelly (Eds.). Chapter 3, 43–59.
[22] Elaine G. Toms. 2015. Complex Tools for Complex Tasks. In *ECIR CEUR Workshop*.
[23] Barbara Wildemuth, Diane Kelly, Emma Boettcher, Erin Moore, and Gergana Dimitrova. 2018. Examining the impact of domain and cognitive complexity on query formulation and reformulation. *IP&M* 54, 3 (2018), 433 – 450.
[24] Barbara M. Wildemuth, Luanne Freund, and Eliane G. Toms. 2014. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation* 70, 6 (2014), 1118–1140.
[25] 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 *IIIX*. ACM, 254–257.