Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies

Kevyn Collins-Thompson[1], Soo Young Rieh[1], Carl C. Haynes[2], Rohail Syed[1]

[1]School of Information, University of Michigan, Ann Arbor, MI U.S.A. 48109

[2]School of Information Studies, Syracuse University, Syracuse, NY U.S.A. 13244

{kevynct, rieh, rmsyed}@umich.edu, cchaynes@syr.edu

ABSTRACT

Users make frequent use of Web search for learning-related tasks, but little is known about how different Web search interaction strategies affect outcomes for learning-oriented tasks, or what implicit or explicit indicators could reliably be used to assess search-related learning on the Web. We describe a lab-based user study in which we investigated potential indicators of learning in web searching, effective query strategies for learning, and the relationship between search behavior and learning outcomes. Using questionnaires, analysis of written responses to knowledge prompts, and search log data, we found that searchers' perceived learning outcomes closely matched their actual learning outcomes; that the amount searchers wrote in post-search questionnaire responses was highly correlated with their cognitive learning scores; and that the time searchers spent per document while searching was also highly and consistently correlated with higher-level cognitive learning scores. We also found that of the three query interaction conditions we applied, an intrinsically diverse presentation of results was associated with the highest percentage of users achieving combined factual and conceptual knowledge gains. Our study provides deeper insight into which aspects of search interaction are most effective for supporting superior learning outcomes, and the difficult problem of how learning may be assessed effectively during Web search.

Keywords: Learning; search behavior; exploratory search; user study.

1. INTRODUCTION

Users often turn to Web search when their goal is to learn [5]. These learning-related search tasks range from basic factual knowledge questions, to more in-depth needs that seek information about 'how' or 'why' [8]. While researchers have recognized the importance of learning as a search outcome, [2][3] current Web search engines are optimized for generic relevance, not learning outcomes. To build search engines that provide better support for learning-related tasks requires progress in several areas. First, systems need more effective algorithms for retrieving documents that are optimal for a particular learning goal. Second, we need better understanding of how different query strategies affect different types of learning. Third, we need reliable assessment methods that can detect when and how different types of learning occur.

Toward these goals, we conduct a user study of an interactive search system with which searchers accomplish different learning tasks within one of three between-subjects querying frameworks. Using data from background surveys, pre- and post-search questionnaires, written responses to prompts, and search interaction logs, we explore the effect of different query and exploration strategies on learning, and characterize effective indicators of learning outcomes in Web search. Specifically, this study addresses the following research questions:

- **RQ1**: What kinds of measures and indicators can be developed to demonstrate learning experiences and outcomes in interactive search systems?
- **RQ2**: What query strategies – submitting a single query, multiple queries, or multiple queries with intrinsically diverse results – do best support human learning experiences and outcomes?
- **RQ3**: To what extent is searchers' search behavior correlated with learning experiences and outcomes?

The idea that search technology can and should play a more central role in supporting deeper learning experiences has been attracting renewed interest with researchers. At both the SWIRL 2012 workshop [3] and the 2013 Dagstuhl Seminar on Evaluation in IR [2] participants proposed ideas for moving "from searching to learning" that emphasized the importance of learning as a search outcome. While these venues discussed the possibilities of a new research agenda for searching as learning, they did not present specific research advances.

This study aims to conceptualize searching as learning by expanding the concept of learning not only in terms of search tasks but also as a part of cognitive activities occurring during the search process. Thus, this study explores new methods and measures for assessing learning across multiple stages of the search process, starting from query formulation, selection of documents, and saving documents to writing summaries at different learning levels, reflecting users' perceived learning and searching experiences and outcomes after searching.

2. RELATED WORK

Previous studies providing context for our work can be divided into four major themes: learning-related search tasks, learning-

CHIIR '16, March 13-17, 2016, Carrboro, NC, USA
 ISBN 978-1-4503-3751-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2854946.2854972

oriented exploratory search, expertise and learning, and assessment of learning in searching.

Learning-related search tasks. Learning-related search tasks can be complex, requiring multiple queries and significant time spent searching and browsing. A study by Bailey et al. [5] of how users engage in such tasks using commercial search engines described a task taxonomy of web search, based on 4 months of search log data that captured query events to Google, Yahoo, and Bing. This taxonomy included some learning-related tasks, spanning topic exploration, fact-finding, and procedural learning. In a later study, Eickhoff et al. [8] analyzed the fraction of sessions that involved a procedural or declarative knowledge intent. Both studies found that learning-related tasks accounted for a non-trivial proportion of all search sessions, and a disproportionally larger fraction of time spent searching: many learning-related tasks each accounted for 1-2% of all search sessions, but 4-5% of time spent searching. Raman et al. [18] had a similar finding for *intrinsically diverse* (ID) search tasks, which are exploratory Web searches intended to explore and learn about multiple aspects of a specific topic. Jansen et al. [12] applied revised Bloom's taxonomy of learning [4] to classify search tasks and described how searching needs could be classified into an appropriate learning model based on searching behavior. Other recent work has attempted to assess the motivation that users exhibit in completing information-seeking tasks, e.g. Kim et al. [14] characterized this motivation in terms of how willing a user was to search and browse documents that were far above their 'typical' reading level.

Learning-oriented exploratory search. Learning-based search activities often involve multiple interactions in the search process and processing of multiple sets of search results that need to be interpreted deeply by searchers. Marchionini [17] claims that search activities that support learning in particular focus on "knowledge acquisition, comprehension of concepts or skills, interpretation of ideas, and comparisons, or aggregations of data and concepts" (p. 43). Therefore, searching activities that support learning require human participation in more continuous and exploratory ways during the search process.

Exploratory search, which focuses on broader information-seeking strategies that emphasize deeper understanding over quick factual answers, has emerged as an alternative paradigm to foster learning and investigation in search [17][22]. Heinström [9] gathered empirical evidence comparing exploratory vs precise information-seeking patterns among students and their relationship to students' learning. She observed that students often undertake broad explorations when exploring new research topics or to get a wide overview of the topic, switching to more precise strategies to fill in specific facts once a topic has been initially explored. Learning outcomes have been proposed as an important future evaluation method for exploratory search [22].

Expertise and learning. Expertise is a dynamic characteristic of users that reflects learning over time. Wildemuth [24] examined how domain expertise was reflected in users' choice of search strategies, finding that domain novices tended to exhibit increasingly similar search strategies to those of more expert users as the novices learned more about the topic. Previous work [23][26] has characterized domain expertise and search behavior in terms of metrics that can be derived from search logs, typically focusing on longer-term behavior patterns across sessions. In one of the first large-scale search log-based studies to examine session-level features of tasks where people are explicitly searching for new knowledge, Eickhoff et al. [8] looked at within-session changes for these expertise metrics. They focused on two specific types of knowledge acquisition: procedural knowledge (how to do something) vs. declarative knowledge (knowing facts about something). The authors found evidence both for learning progress within single session, and for persistence of learning across sessions. Significant proportions of new query terms came from result page snippets and recently visited pages, showing that the search process itself contributed to augmenting the user's domain knowledge. Other recent studies, e.g. [26] have attempted to predict domain knowledge from user search behavior.

Assessment of learning in searching. A few studies have attempted to identify indicators of learning during the search process. Vakkari et al. [21] found that students' level of knowledge about their topic can predict patterns of search queries, in that students who know less about the topic are likely to use fewer, broader, more vague search terms as their queries. In a study with medical students, Vakkari and Huuskonen [20] found that effort put into the search process did not lead to better search outcomes (the products delivered by a system), but did improve task outcomes (the benefits the system produced). Several IR researchers designed research methods to investigate learning as a measure of search outcome. In one of the earlier studies measuring learning, Hersh et al. [10] showed how searching enabled students to answer more questions in a post-search quiz. Instead of a quiz, Kammerer et al. [15] asked their study participants to write a summary about the topic after using the exploratory search interface, MrTaggy, to assess learning. Summary quality was evaluated based on topic-specific criteria including the number of reasonable topics, overall quality of the topic description, and number of arguments. Wilson and Wilson [25] developed systematic techniques to measure the depth of learning at three levels: quality of facts, interpretation of data into statements, and use of critique. These previous studies indicate that traditional measures in information retrieval, such as recall and precision, can be effectively complemented with alternative measures that pay more attention to search behavior, process, and task outcomes beyond basic search results.

3. METHODS

We conducted a user study with an interactive Web search system in a laboratory setting, controlling two conditions: query strategy and search tasks. By controlling 'query strategy' we mean constraining or expanding the space of possible query interactions available to users in the search environment. As we were interested in finding out to what extent query strategies have an effect on users' learning outcomes and experiences from searching, we decided to compare search-related and learning-related measures by controlling users' query strategies as follows:

Single Query (SQ) condition: subjects are asked to select a single query from a given set and to use the results of this query for the remainder of the search session;

Multiple Query (MQ) condition: after selecting an initial query from a given set, subjects may run and use the results from multiple queries of their own design;

Intrinsic Diversity (ID) condition: subjects are allowed to run multiple queries, as in the MQ condition, but additionally the results are *intrinsically diverse*, covering a range of subtopics related to the query and providing query suggestions associated with subtopics that they may have not initially considered [18].

For all three conditions, subjects were instructed to select an initial query out of 10 queries that were pre-determined by the authors for each of two search tasks. This restriction was to minimize the variance of results that might occur due to differences in users' abilities to formulate a good initial query which in turn could influence the assessment of learning.

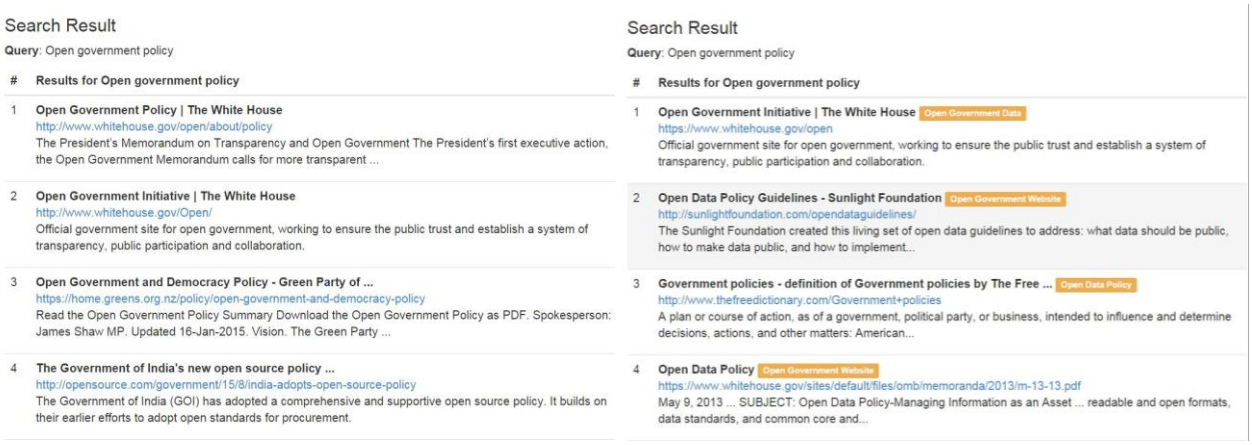

Figure 1. Search results interface for single & multiple query conditions (left) and intrinsically diverse condition (right).

After the initial query, subjects were allowed either to create their own queries or select one of the topics offered by the Intrinsic Diversity (ID) condition. The study used a mixed between- and within-subject design, the between-subject factor being query strategy and within-subject factor being search tasks. The order of the two search tasks was rotated to avoid any ordering bias, learning effect or potential fatigue issues. Subjects were randomly assigned to one of the three query formulation conditions and compensated US $20.00 for participation.

3.1 Study Participants

A recruiting email was sent to a University of Michigan School of Information mailing list. Undergraduate students, graduate students, and alumni can opt in to subscribe to the list. A total of 44 study subjects (30 female, 14 male) signed up for the study. We offered 10 different session timeslots in which we accepted up to 10 people for each session. Participants ranged in age from 19 to 38 years old. Participants also varied in their academic standing: there were 34 graduate students, 7 undergraduate students, 2 doctoral students, and 1 alumnus; 36 of the recruited subjects were affiliated with the University of Michigan.

Subjects were assigned to one of the three query conditions in a randomly-initialized, round-robin fashion to ensure balanced numbers for each condition. Two subjects were later removed due to technical issues with incomplete data gathering, leaving 42 subjects for analysis. The final counts of subjects in each query condition were: Single Query (SD): 12, Multiple Query (MQ): 15, Intrinsic Diversity (ID): 15.

3.2 Search System and Interface

The search system was hosted on Amazon EC2 and used an architecture derived from uFindIt [1] that logs user events such as queries and clicks to a MySQL database. The baseline ranked document lists for the single and multiple query conditions were provided by the Google Custom Search API[1]. The intrinsically diverse (ID) condition was implemented using the ranking algorithm by Raman et al. [18] that jointly finds a set of diverse subtopics, and a set of main results representing the best results that cover those subtopics[2]. The ID result candidates were obtained using the same Google Custom Search API to obtain ranked lists for the main query and subtopics. Since the ID algorithm requires a source of subtopic candidates, in our implementation we used section headings of Wikipedia articles if one existed for a given query, and otherwise used query suggestions provided by the Bing Related Query API.

In the user interface, each ID subtopic was displayed as a rectangular button to the right of the corresponding document title in the search engine results page (SERP). Users could click on the subtopic button to launch a new query using that subtopic as the query, whose results would be a simple baseline ranking (not another ID ranking). Figure 1 shows screenshots of the search results interface used for the study conditions.

3.3 Tasks and Procedures

We developed two tasks as "simulated work task situations" [7] in which we gave scenarios that simulate real life information needs. Subjects were asked to conduct searches on a topics for a course term paper. We developed two search tasks that could be characterized differently in terms of complexity and domain/non-domain knowledge. At the same time, we tried to create tasks that required students to explore multiple aspects of the topic. We made such expectations clear by saying "present your views on this topic" and "save all the webpages, publications, and other online sources that are helpful for you to write a paper." The task descriptions as shown to the participants were as follows.

> Task 1 description (Oil Spill): Suppose you are taking an introductory Environmental Science course this term. For your term paper, you have decided to write about what chemicals can be used to clean up oil spills. You also would like to learn what environmental effects oil spills have in the ocean and on shore.
>
> Task 2 description (Open Data): For a course you are taking this term, you have decided to write a term paper about government open data policy. You know that it is about how government agencies manage information as an asset throughout the life cycle to promote openness.
>
> General descriptions for both tasks: The professor requires all students to demonstrate what they learn about a particular topic by conducting searches online and presenting their views on this topic. To prepare your term paper, you need to collect and save all the webpages, publications, and other online sources that are helpful for you to write a paper. After your search is completed, you will be asked to answer six questions about this topic. Questions include answering

[1] With personalization and 'safe search' filtering not activated.

[2] Using default ID algorithm parameters of $\beta = 0.5$, $\lambda = 0.3$.

questions and writing an outline and completing a survey based on what you have learned from this search. To be able to answer these questions, you may want to take some notes during the searching.

The first "oil spill" task dealt with a scientific topic about which students were unlikely to have extensive domain knowledge, but for which the basic concept was not especially abstract or difficult to understand. The second "open data" task addressed a topic about which participants might have a certain level of prior exposure and some domain knowledge, but the basic concept itself was more abstract and could be rather complex.

The study took approximately 1 hour to complete. When subjects arrived at the laboratory, they received a one-page set of written instructions that outlined the 13 steps for subjects to take step-by-step. Subjects were presented with their first search task both on screen and in hard copy. Next, after they completed reading the task description, they were asked to fill out a Pre-Search Questionnaire which had three scale-based questions and one writing question about subjects' knowledge level in the topic. They were then guided to follow instructions for one of the three query conditions. Their search continued without time constraints. When subjects found a document they liked in the search results page (SERP), they would click a document URL to view the document. After they finished reading the document, they would come back to the SERP, at which point they were presented with a new button "Was this helpful?" that was associated with that particular document. They were given to a choice to save the document by clicking that button or to click on the X icon. They would then return to looking at the search results page. When they wanted to stop their searching, they would click "I completed my search for this topic" button, which would take them to the next screen displaying a Post-Search Questionnaire, which had 23 questions in total. Once they completed the Post-Search Questionnaire, they were allowed to start the same process for the second search task. When they completed searching and filling out questionnaires for both tasks, they were taken to the screen showing a Background Questionnaire, which was the last step in their participation.

3.3.1 Questionnaires

We now discuss the motivation and methods for our question design in each portion of the study. A complete inventory of the question set described here is given in Table 3. The content of the Pre-Search Questionnaire was identical for the two search tasks. The first three questions were closed questions designed to assess subjects' prior knowledge level (**P1**), interest in the topic (**P2**), and perceived difficulty of searching (**P3**). These three questions were assessed using a 5-point scale (1 = not at all and 5 = very likely). The fourth question (**P4**) asked participants to summarize their topic knowledge.

The Post-Search questionnaire was composed of two parts: (1) A set of 15 questions investigating learning and searching experiences on a 5-point scale (1= not at all and 5=very likely). Questions covered variables related to search experiences for the purpose of information exploration, user experiences with system usability, and learning attitudes focusing on interest, motivation, and willingness to learn more. The on-screen order of these 15 questions was rotated for each subject. (2) Two questions assessing subjects' perceived search and learning outcomes based on a self-reported learning score on a scale of 0 to 100. All responses were collected using Google Forms, and exported to Excel and Stata for analysis.

3.3.2 Post-Search Written Tests

In addition to the above 17 questions, the Post-Search Questionnaire also included a set of six learning assessment questions that were developed using Bloom's revised learning taxonomy [4]. Each question addressed one of Bloom's learning levels such as remembering (**Q1**), understanding (**Q2**), applying (**Q3**), analyzing (**Q4**), evaluating (**Q5**), and creating (**Q6**). Out of six questions, the first three questions focused on *lower-level cognitive learning* (Q1–Q3) while the other three questions focused on *higher-level cognitive learning* (Q4–Q6).

The lower-level learning questions Q1–Q3 were tailored for the nature of the specific search tasks. For the Oil Spill task, the first three questions designed to assess lower-level learning were:

- **Q1**: What are the kinds of materials that can be used as a sole cleanup method in small spills?
- **Q2**: When workers decide which methods are most effective to clean up oil spills, what are some factors that they should consider to make decisions for recovery methods?
- **Q3**: Why do you think that oil spills are important environment issues? Describe its effects and impacts on human and environment.

For the Open Data task, the corresponding questions were:

- **Q1**: Is copyright protection available for works of the United State Government?
- **Q2**: In 2007, a number of open government advocates got together and claimed that government data shall be considered open if it is made public in a way that complies with some fundamental principles. Others added more principles since then. What are some examples of principles of open government data?
- **Q3**: What kinds of individuals, communities, or organizations could be benefited as a result of accessing open data provided by government?

The higher-level learning questions Q4–Q6 were identical across tasks and were as follows:

- **Q4**: Based on what you have learned from your searching, please write an outline for your paper.
- **Q5**: Please write what you learned about this topic from your searching with 3-5 sentences.
- **Q6**: Based on your searching, what questions do you still have about this topic?

We used Google Forms to collect subjects' written responses to these six questions.

3.4 Coding of Written Summaries

To analyze the written responses to the one pre-search summary question (P4) and six post-search questions (Q1-Q6) described above, we derived the following three assessment variables, the first of which was based on a detailed coding scheme designed to be a sensitive measure of knowledge acquisition.

1. **Cognitive Learning Scores**. For each question Q1-Q6, we defined seven criteria for assessing different dimensions of knowledge that might be observed in the participant's response. These criteria ranged from assessing factual knowledge by checking whether subjects could recall factors, issues, and elements, to assessing conceptual knowledge by looking at subjects' written responses to see whether they could identify themes and integrate multiple concepts. The criteria were derived from cognitive processes identified by Anderson & Krathwahl [4] as being associated with each of the six main learning levels in Bloom's Revised Taxonomy.

Each written response was assigned a raw score in the range 0–7 by counting how many of these seven learning criteria were demonstrated. If a written response showed no evidence of knowledge about the topic, the lowest score (0) was given. If a response exhibited every knowledge dimension listed in all seven learning criteria, it was given the highest score (7).

Two coders independently applied this coding scheme comprising 84 different learning criteria (7 criteria for each of the 6 written tests, Q1–Q6, for both tasks) to analyze the written responses of all 42 valid subjects. There were 87 written summary responses in total to Q1–Q6. Inter-coder agreement was computed between the two coders. We used Holsti's coefficients [11] to measure the consistency of coder judgments by calculating the ratio of coding agreements to the total number of coding decisions for each of the six questions. For both search tasks, we reached a high level of agreement: the mean inter-coder reliability across Q1–Q6 for the Oil Spill Task was 0.914 and for the Open Data task was 0.797.

From the raw scores we derived 3 cognitive learning scores:

- *Lower-level cognitive learning score*: The sum of raw scores coded from Q1–Q3 responses.
- *Higher-level cognitive learning score*: The sum of raw scores coded from Q4–Q6 responses.
- *Overall cognitive learning score*: The sum of raw scores coded from Q1–Q6 responses.

2. **Knowledge Level Gain**. To capture the nature of a subject's gain in knowledge during a task, we also coded written responses for P4 (prior knowledge) and Q5 (current knowledge) with a level score based on the highest of three levels of knowledge judged to have been exhibited in the writing (0=no knowledge, 1=factual knowledge, and 2=conceptual knowledge). If we observed a gain in this score from P4 to Q5, we considered that the subject had increased their level of knowledge for that task.

3. **Written Response Length**. This was simply the number of characters submitted by a subject for each of the written tests.

In Section 4, we examine these observed variables and their relationships to those derived from other sources in the study.

3.5 Logging of Search Interaction

To identify implicit indicators of learning from search interaction, we collected and analyzed the following variables that captured important aspects of time-related interaction, clicking behavior, and judgments of "usefulness" for each document viewed:

Unique docs clicked in search results: Total unique document clicks from the SERP (Search Engine Results Page) for each query entered.
Total time spent on assessing SERP: The time difference between when subjects returned to a SERP and when they switched to another page.
Average time spent on assessing SERP: Total time spent on assessing SERP divided by total queries entered.
Total time spent viewing documents: The difference from the time that subjects left a SERP to the time that they returned to the SERP.
Average time spent viewing documents: Total time spent viewing documents divided by count of unique clicks.
Average time spent viewing documents per query: Total time spent viewing documents divided by total queries.
Number of useful documents saved: Total documents users marked as useful.

Table 1. Log-based variables

Variable	Task	SQ	MQ	ID
Unique documents clicked in search results	Open	5.07	6.40	7.00
	Oil	4.91	6.66	5.73
Total time assessing SERP (sec)	Open	146.85	98.13	276.77
	Oil	83.42	137.07	176.07
Average time assessing SERP per query (sec)	Open	146.85	45.38	133.93
	Oil	83.42	85.42	81.95
Total time viewing documents (sec)	Open	973.64	964.4	1810.38
	Oil	787.33	1410.4	1399.06
Average time spent per document (sec)	Open	196.66	228.38	270.23
	Oil	146.52	272.29	246.87
Average time viewing documents/query (sec)	Open	973.64	505.18	865.72
	Oil	787.33	652.69	723.52
Number of useful documents saved	Open	3.00	4.00	4.08
	Oil	2.33	4.00	4.00

In analyzing document selection behavior, we investigated how many links on the SERP page the user clicked, what position each link was ranked at and how a binary judgment (a document was useful or it was not useful) the user gave to the corresponding document after viewing it. Table 1 summarizes observed values of these variables per task and condition, averaged across users.

4. DATA ANALYSIS

We now summarize our main results regarding self-reported outcomes, learning measures and indicators, the relationship of query strategy conditions to learning outcomes, and the relationship between learning outcomes and search behavior.

4.1 Self-Reported Searching and Learning Experiences

Before searching, subjects were asked about their topic knowledge and perception for each task (P1-P3). There was a significant difference in their perceived difficulty (P3) between the two tasks. Subjects perceived that the information searching required for the Open Data Task (M=3.10, SD=.81) would be more difficult than that required for the Oil Spill Task (M=2.53, SD=.93), $p<0.01$. However, their perceived interest and prior knowledge did not show any difference between the two tasks.

After subjects completed their searches, they were asked to respond to 15 items in the Post-Search Questionnaire. Overall, subjects seemed have more positive search and learning experiences with the Oil Spill Task than the Open Data Task across the three query conditions: this is not surprising, as subjects perceived the Open Data Task would be more difficult even before they began searching. Subjects did not report that their search experiences differed significantly depending on the perceived difficulty of tasks. For instance, subjects responded that they felt that search time was spent productively (Oil Spill M=3.87, Open Data M=3.62) and they were cognitively engaged in the search task (Oil Spill M=3.85, Open Data M=3.75), showing no difference between the tasks. However, they responded that they had more positive learning experiences after searching when they perceived the task as easier. For instance, they rated that they were able to develop new ideas and perspectives more highly in the Oil Spill Task (M=3.65) than in the Open Data Task (M=3.17), $p<.05$. They also reported understanding the topic at a higher level in the Oil Spill Task (M=2.82) than in the Open Data Task (M=2.37), $p<.05$, although

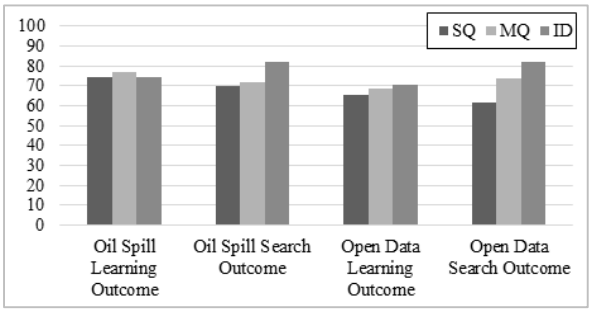

Figure 2. Users' self-reported perceived search and learning success by query condition and task.

their overall rating for this question was lowest out of the 15 items in the Post-Search Questionnaire.

We next examined how subjects' responses to searching and learning experiences might differ depending on the query condition to which they were exposed. An analysis of individual items revealed that subjects' learning and search experiences were not significantly different across the three query conditions (SQ, MQ, and ID) except in relation to one question. After subjects completed searching about Open Data, those subjects who were assigned to the ID condition (M=3.85) reported feeling able to synthesize the various pieces of information together at significantly higher levels than those subjects who were in the other two conditions (SQ=3.14, MQ =2.85), $p<.05$.

Cross-question factor analysis. To investigate possible aspects that shaped subject's responses to questions about their searching and learning experiences, we conducted a factor analysis of responses across questions: specifically, principal component factor analysis using Varimax with Kaiser Normalization, based on the responses to the 15 items with 5-scale rating (Q7-Q21) from the Post-Search Questionnaire. This analysis revealed three distinct factors, which we term 'Experience Factors' that characterized aspects of subjects' responses:

Experience Factor 1 (Search for Information Exploration) focuses on users' experiences with searching itself, examining their effort, engagement, feeling of time well-spent, and perception of knowledge expansion as a result of searching.

Experience Factor 2 (User Experience with Search Systems) deals with users' experience with respect to learning by investigating whether searching helps people to increase their interest in the topic or develop new ideas and a willingness to find and share more information.

Experience Factor 3 (Learner Interest and Motivation) is related to the use and usability of search systems.

Internal consistency for each of the scales was examined using Cronbach's alpha. The alphas were 0.855 for Search for Information Exploration (6 items), 0.846 for User Experience with Search Systems (4 items), and 0.814 for Learner Interest and Motivation (5 items). Factor 1 explains 25.76% of the total variance, and Factor 2 explains 20.52% of the total variance. Factor 3 explains 17.60% of the total variance. Table 3 (the question inventory), shows questions (Q7-Q21) grouped by these three Experience Factors.

Subjects rated their learning experience with the search systems (Experience Factor 2) lower than their search experience for information exploration (Experience Factor 1) and their perceived learner interest and motivation (Experience Factor 3) across the two tasks. The results of ANOVA (df=39) comparing the three factors across the three query conditions (SQ, MQ, and ID) showed that there was no statistical difference in subjects' experiences of search for information exploration, user experience with search systems, and learner interest and motivation across query strategies.

Self-reported outcomes. Two questions in the Post-Search Questionnaire (Q22, Q23) asked subjects to grade their own learning and search outcomes on a scale of 0-100. Overall, subjects gave lower scores to both learning and searching outcomes related to the more difficult task – Open Data. The results of ANOVA showed that subjects self-reported learning outcomes did not differ significantly in both tasks depending on search strategies. However, while the mean of self-reported searching outcomes across the three query formulation conditions (SQ, MQ, and ID) were not significantly different in the case of the Oil Spill task, they were different for the Open Data task ($F(2,37) = 4.68$, $p<.02$): subjects who used the intrinsically diverse (ID) search system reported the highest search outcomes (M=81.92), compared with those subjects from the SQ (M=73.69) and MQ conditions (M=61.50). Also, those subjects who were allowed to reformulate their queries (MQ) reported higher search outcome scores than those who had to use a limited query formulation (SQ). Figure 2 summarizes the analysis of perceived learning and search outcomes across tasks and query conditions.

4.2 Learning measures and indicators

The analysis of data collected from the logs, questionnaires, and written tests described in Section 3 revealed a number of explicit and implicit indicators potentially useful for measuring learning in web searching (RQ1).

4.2.1. Explicit Measures

(1) Assessment of High- and Low-Level Cognitive Learning. When subjects engaged in searching for an easier task (Oil Spill), they demonstrated more evidence of lower-level cognitive learning (M=7.21) than higher cognitive learning (M=5.88) in their written summaries. When searching for a more difficult task (Open Data), they provided summaries with slightly more evidence of higher cognitive learning (M=5.31) than that of lower-level cognitive learning (M=4.55). Overall, we did not find that there was a significant difference in overall cognitive learning scores across the three query conditions.

(2) Perceived Learning and Searching Outcomes. When we examined the correlation between perceived learning outcome scores (Q22) and actual cognitive learning scores, we found that for both tasks, perceived learning outcome positively correlated with both lower-level cognitive learning scores on Q1–Q3 (r=.33, r=.38) and higher-level cognitive learning scores on Q4–Q6, (r=.32, r=.37). Aggregating both lower and higher cognitive learning scores, the overall correlation was even stronger (r=.40, r=.45). This result implies that subjects were able to assess their own learning outcomes reasonably well, and thus perceived learning outcome scores could be used as a measure for learning in searching. While the perceived searching outcome variable (Q23) was useful in comparing subjects' perceived outcomes across two tasks and three query conditions, we found that perceived searching outcomes were not correlated with learning scores from written responses.

4.2.2. Implicit Indicators

(1) **Knowledge Level Gain.** Of the 42 subjects who wrote valid answers to questions P4 and Q5, 16 showed no knowledge level

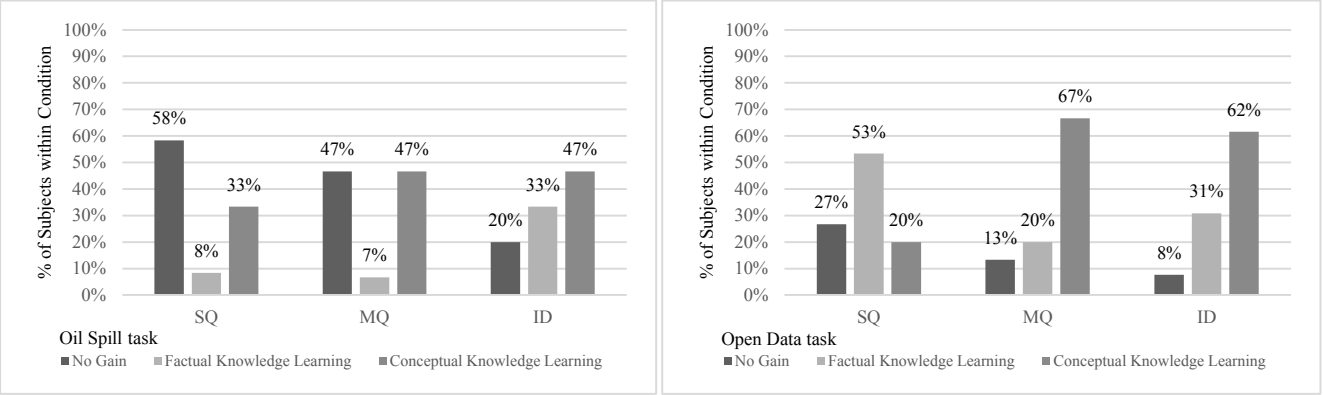

Figure 3. Percentage of Subjects Gaining in Knowledge Level in each Query Condition, for Oil Spill and Open Data tasks.

gain, 15 had a +1 gain, and 10 had a +2 gain. One participant had a negative gain (exhibiting conceptual knowledge in the pre-test but only describing factual knowledge in the post-test). We found a strong positive correlation between perceived learning outcomes and actual knowledge level gain in the ID condition, for both the Open Data (r=0.69) and Oil Spill (r=0.64) tasks.

(2) **Length of Written Responses.** We found that the length of written responses was another potential indicator of learning: the total combined length of all six post-written responses (Q1-Q6) had a strong positive correlation with the overall cognitive learning score for each task (r=.75, r=.90). The total combined response length for the lower-level cognitive learning assessment questions (Q1-Q3) was strongly correlated with the corresponding lower-level cognitive learning score for both the Open Data and Oil Spill tasks (r=.75, r=.86). Similarly, the total combined length of the written responses to the higher cognitive learning questions (Q4-Q6) was correlated with the higher-level cognitive learning score for both tasks (r=.66, r=.83). In sum, longer responses were more likely to exhibit more evidence of cognitive learning as measured by the coding scheme.

(3) Interaction Speed. Although time has been used in IR evaluation as an indicator of efficiency [13][19], we took a different approach to time-related measures, focusing on the relationship between interaction speed and learning outcomes. We hypothesized that time spent searching might be positively correlated with increased learning. Indeed, we found that the average viewing time spent per document had positive correlations with overall cognitive learning scores in both Open Data (r=.44) and Oil Spill (r=.39) conditions. Thus, regardless of the nature of tasks and query strategies, subjects who spent more time reading documents were more likely to receive higher scores on their writing summaries.

(4) Interaction with Documents. We also hypothesized that the more subjects saved documents they judged to be useful, the more they were learning through searching, which would lead to higher quality written responses. We observed a correlation between interaction with documents and learning only for the Open Data task: there was a positive correlation (r=0.385) between the number of unique documents clicked from the SERP and lower-level cognitive learning score. The number of useful documents saved was also strongly correlated with the lower-level cognitive learning score (r=.45) and with perceived search outcome (Q23), (r=0.34). This provides some evidence that document interaction variables could be useful as implicit indicators of learning.

4.3 Relationships of query strategies to learning outcomes

To help answer RQ2, we compared knowledge level gain across query conditions. In the Open Data task, subjects' gain in knowledge level was relatively consistent across interaction conditions, with ANOVA results showing no significant differences (SQ=0.928, MQ=1.33, ID=1.153). However, in the Oil Spill task, we found a statistically significant difference between conditions (SQ=0.41, MQ=0.73, ID=1.20: p=.031): the ID condition offered the highest learning gains compared with both the SQ and MQ conditions. We also observed that prior knowledge scores in this task also showed significant differences (SQ=1.08, MQ=0.8, ID=0.2, p=.004). For all three conditions, the post-search knowledge level (Q5) is almost the same (SQ=1.5, MQ=1.53, ID=1.4).

We also found a general trend that the average time users spent per query in terms of reading documents and in terms of assessing the SERP page was much higher in the SQ and ID conditions compared to the MQ condition (Table 1). Users in the MQ and ID conditions also tended to select more unique documents in the SERP than users in the SQ condition.

An analysis of knowledge level gain across the three query conditions, shown in Figure 3, shows that subjects exhibited different patterns in gaining knowledge depending on the query condition. For the Oil Spill task, 58% of the subjects who were assigned to the SQ condition did not gain any new knowledge as a result of searching, while 33% gained knowledge at the conceptual level of learning. In contrast, only 20% subjects in the ID condition showed no gain in knowledge after searching, and 47% achieved conceptual-knowledge-based learning for the same task. For the Open Data task, most subjects showed conceptual knowledge gains in both the MQ condition (67%) and ID condition (62%), compared to the SQ condition (20%). The ID condition gave the best combined factual + cognitive knowledge gain score on both tasks in terms of the percentage of users achieving a gain (Oil Spill: 80%; Open Data: 93%). The SQ and MQ conditions achieved gains of (41%, 73%) and (54%, 80%) of users respectively. Thus, of the three query conditions, there was some evidence that the intrinsically diverse (ID) query condition gave the best support to learning for these tasks.

One limitation of our study was that, while subjects might exhibit similar levels of knowledge about a topic, they might also have

different learning abilities, which could interact with variables such as their assigned query strategy condition. A future study could add specific assessments to track and account for individual differences in learning ability. Also, since the intrinsically diverse (ID) condition both modified the results ranking and added query suggestions, additional experiments would help understand how each of these modifications contributed to our observed results.

4.4 Relationships between key variables and learning outcomes: user factor analysis

Finally, to understand the relationships between key learning-related variables in this study as they relate to users' search behavior (RQ3), we conducted a second factor analysis across users based on their search behavior and learning outcomes (in contrast to the first factor analysis in Sec. 4.1 based on experience-oriented questions Q7-Q21). Specifically, each participant was represented by their responses to learning-oriented questions P1-P4, Q1-Q6 and the log-based search behavior variables TimePerDoc (Table 1, average time spent viewing documents per query) and TotalClicks (Table 1, unique documents clicked in search results). For space reasons we omit analysis of the remaining 15 variables, which also showed less consistent contrasts between factors than the ones included here. We used k=2 factors in order to examine whether at least two main groups of potentially different types of users were evident from the data. We did one factor analysis for each query formulation condition.

Figure 4 shows the resulting factor biplots. A biplot shows users and variables in the same factor space: users are shown as numeric points plotted by their factor scores, and each variable is shown as a vector whose coordinates are the factor coefficients of the variable. Thus, points that are close together represent users with similar factor scores, and vectors of similar length with small angle between them represent highly correlated variables.

Across all conditions, the lower-level cognitive learning variables (Q1-Q3) generally clustered together, as did higher-level cognitive learning variables (Q4-Q6). Of these two groups, the lower-level cognitive learning scores were most consistently correlated across conditions (Q1,Q2: r= 0.42; Q2,Q3: r=0.33; Q1,Q3: r=0.46; all p< 0.001). For higher-level cognitive learning scores, there was consistent but weaker correlation (Q4,Q5: r=0.41, p<0.001; Q4, Q6: r=0.24, p<0.04). For pre-search scores, none of P1-P4 were correlated with either TimePerDocs or TotalClicks. However, P2 scores (interest in learning more about the topic) were correlated with Q2-Q5 scores. For the log-based variables, we can refine our initial finding in Sec. 4.2.2 that time

Table 2. User factor analysis of key variables, for each of the three query conditions.

Condition	SQ		MQ		ID	
User Factor	F1	F2	F1	F2	F1	F2
TotalClicks	**0.71**		**-0.47**	-0.10	-0.31	**0.70**
TimePerDoc		**-0.19**	**0.76**	0.31	**0.36**	
P1		**0.44**	0.14	0.20	-0.10	**0.36**
P2	0.33	0.65	-0.22	-0.29	**0.39**	
P3	-0.20	**-0.74**		**-0.39**	-0.20	
Q1	0.53	0.78		**0.37**	0.11	**0.61**
Q2	**0.62**		0.35	0.36	0.19	**0.85**
Q3	**0.57**		0.17	**0.42**	0.32	0.38
Q4	**0.56**	0.22	0.15	**0.73**	**0.70**	
Q5	**0.58**		0.34	**0.78**	**0.65**	
Q6	**0.16**	-0.32	**0.98**	0.18	**0.10**	
%Variance	*21%*	*18%*	*19%*	*18%*	*17%*	*13%*

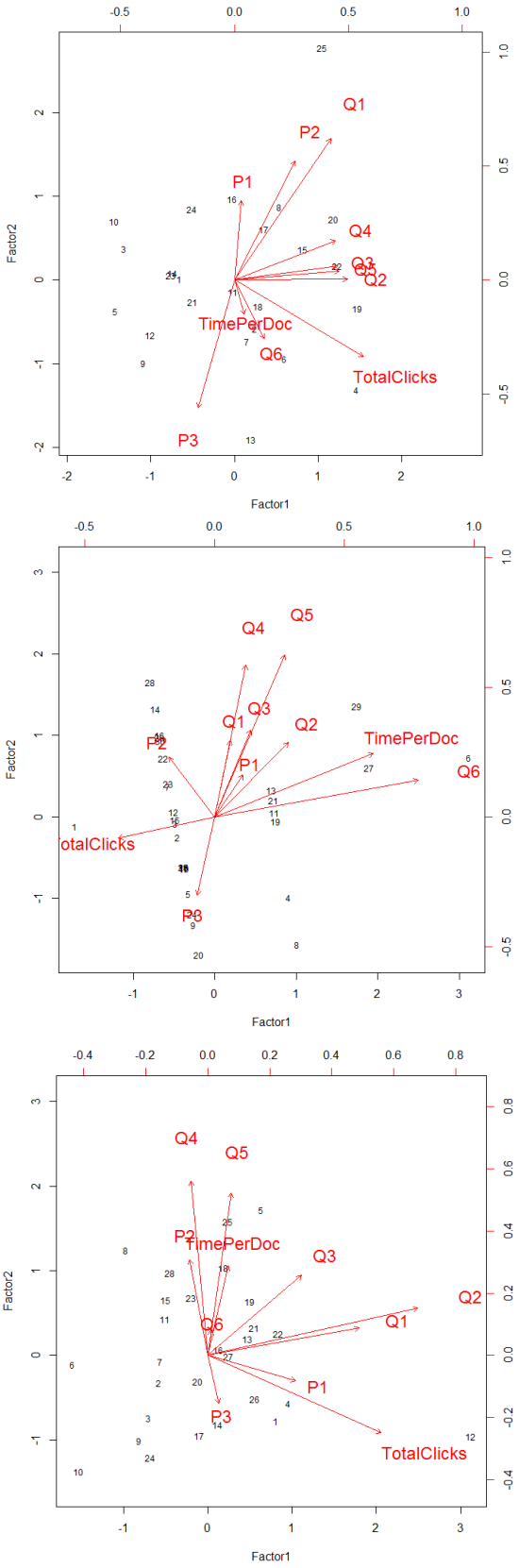

Figure 4. User factor biplots showing correlation of key learning and search behavior variables (red arrows) and clustering of subjects in factor space (numeric points) for Single Query (top), Multiple Query (center), and ID conditions (bottom). Vectors of similar length with small angle between them represent highly correlated variables.

spent per document is correlated with overall cognitive learning scores. Figure 4 makes clear that TimePerDoc is consistently correlated with Q6 scores (creative question-asking) across all query conditions ($r=0.56$, $p<0.001$), and to a lesser extent, the other high-level cognitive learning scores (Q4: $r=0.26$; Q5: $r=0.29$, $p < 0.01$) but not with lower-level cognitive learning scores (Q1-Q3). TotalClicks was most strongly correlated with Q2 ($r=0.22$; $p<0.05$) that assessed users' understanding of the topic. To examine the nature of the user groups found by the user factor analysis, we inspected the factor loadings (shown as User Factors F1, F2 in Table 2), with the following interpretations.

Single Query Condition. The group of users associated with User Factor 1 had greater positive loading on TotalClicks and cognitive learning scores (Q2-Q6). User Factor 2 subjects were characterized by high loadings on background knowledge (P1), interest level (P2), factual recall (Q1) and lower negative loading on perceived difficulty (P3).

Multiple Query Condition. User Factor 1 users had higher positive loading for time spent per document, and writing creative questions (Q6). User Factor 2 users had high positive loading on existing knowledge (Q1) and negative loading on perceived difficulty (P3), and strong positive loading on both lower and higher-level cognitive learning scores (Q3-Q5). These two clusters of users are evident on Fig. 4 (center).

Intrinsically Diverse Condition. Users associated with User Factor 1 were characterized by higher positive loading on time spent per document, level of interest (P2), and the three higher-level cognitive response scores (Q4-Q6). In contrast, User Factor 2 users had higher positive loading on clicks (TotalClicks), existing topic knowledge (P1), and response scores for the lower-level cognitive, factual questions (Q1, Q2).

In sum, our user factor analysis showed the existence of distinct groups of users exhibiting complementary search strategies: one group chose a broader strategy, tending to click and explore more results while obtaining higher lower-level cognitive learning scores, while the other group tended to read fewer results more deeply while obtaining higher scores on the higher-level cognitive learning assessment questions. These differences were evident in the richer query environments of both the MQ and ID conditions.

5. CONCLUSION

In this paper we described a lab-based user study in which we investigated potential indicators of learning in web searching, effective query strategies for learning, and the relationship between search behavior and learning outcomes. We developed and analyzed a rich set of implicit and explicit learning measures based on behavioral data from search logs, questionnaires, and written responses to knowledge questions. The written responses were coded using a new, carefully developed scheme based on Bloom's revised learning taxonomy. Our examination of potential learning indicators, and how search behavior correlated with learning outcomes, found that searchers' perceived learning outcomes closely matched their actual learning outcomes; that the amount searchers wrote in the post-search survey was highly correlated with their cognitive learning question scores; and that the time searchers spent per document while searching was also highly and consistently correlated with higher-level cognitive learning question scores. To investigate which search paradigms best support human learning experiences and outcomes our study incorporated three distinct between-subjects query strategies – submitting a single query, using multiple queries, and using multiple queries with intrinsically diverse (ID) subtopics. We found that the ID condition gave a large advantage over the SQ and MQ conditions, for both search tasks, in terms of the percentage of users able to achieve combined factual and conceptual knowledge gains. Our study provides deeper insight into the problem of how learning may be assessed effectively during web search, and which aspects of search interaction are most effective for supporting superior learning outcomes.

Acknowledgements. We thank Yi-Yin Alison Wang, Kuan-Chun David Cheng, and Gracie Mu-Yun Chien for their study contributions; the University of Michigan Office of Research and School of Information for support of this work; and the anonymous reviewers for their comments.

6. REFERENCES

[1] Ageev, M., Guo, Q., Lagun, D., and Agichtein, E. 2011. Find it if you can: a game for modeling different types of web search success using interaction data. In *Proc. of SIGIR '11*. ACM, 345-354.

[2] Agosti, M., Fuhr, N., Toms, E., and Vakkari, P. 2014. Evaluation methodologies in Information Retrieval Dagstuhl seminar 13441. *ACM SIGIR Forum.* 48, 1 (June 2014), 36-41.

[3] Allan, J., Croft, B., Moffat, A., & Sanderson, M. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012: the second strategic workshop on information retrieval in Lorne. *ACM SIGIR Forum*. 46, 1 (May 2012), 2-32.

[4] Anderson, L. W. and Krathwohl, D. R. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Longman, New York.

[5] Bailey, P., Chen, L., Grosenick, S., Jiang, L., Li, Y., Reinholdtsen, P., Salada, C., Wang, H., and Wong, S. 2012. User task understanding: a web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Kanagawa, Japan*. (Oct. 2012).

[6] Bloom B. 1956. *Taxonomy of Educational Objectives*. David McKay Company, New York.

[7] Borlund, P. 2000. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*. 56, 1, 71-90.

[8] Eickhoff, C., Teevan, J., White, R., and Dumais, S. 2014. Lessons from the journey: a query log analysis of within-session learning. In *Proc. of WSDM 2014*. ACM, New York, NY, 223-232.

[9] Heinström, J. 2006. Broad exploration or precise specificity: Two basic information seeking patterns among students. *JASIST*. 57, 11 (Sept. 2006), 1440-1450.

[10] Hersh, W. R., Elliot, D. L., Hickam, D. H., Wolf, S. L. and Molnar, A. 1995. Towards new measures of information retrieval evaluation. In *Proc. of SIGIR 1995*. ACM, 164-170.

[11] Holsti, O. R. 1969. *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley, Reading, MA.

[12] Jansen, B. J., Booth, D., and Smith, B. 2009. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*. 45, 6 (Nov. 2009), 643-663.

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

[14] Kim, J. Y., Collins-Thompson, K., Bennett, P. N., and Dumais, S. T. 2012. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proc. of WSDM 2012*. ACM, New York, NY, 213-222.

[15] Kammerer, Y., Nairn, R., Pirolli, P., and Chi, E. H. 2009. Signpost from the masses: Learning effects in an exploratory social tag search browser. In *Proceedings of SIGCHI 2009*. ACM, 625–634.

[16] Krathwohl, D. R. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice.* 41, 4, 212-218.

[17] Marchionini, G. 2006. Exploratory search: from finding to understanding. *Comm. of the ACM*. 49, 4 (April 2006), 41-46.

[18] Raman, K., Bennett, P. N., and Collins-Thompson, K. 2014. Understanding intrinsic diversity in Web search: Improving whole-session relevance. *ACM Trans. Info. Systems*. 32, 4 (Oct. 2014).
[19] Smucker, M., and Clarke, C. 2012. Time-based calibration of effectiveness measures. In *Proc.of SIGIR 2012,* ACM, 95-104.
[20] Vakkari, P., and Huuskonen, S. 2012. Search effort degrades search output but improves task outcome. *J. American Society for Information Science and Technology*. 63, 4 (April 2012), 657-670.
[21] Vakkari, P., Pennanen, M. and Serola, S. 2003. Changes of search terms and tactics while writing a research proposal. *Information Processing & Management*. 39, 3 (May 2003), 445-463.
[22] White, R. W., and Roth, R. A. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*. 1, 1 (2009), 1-98.
[23] White, R. W., Dumais, S. T., and Teevan, J. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of WSDM 2009.* (Barcelona, Spain, February 2009). WSDM '09. ACM, New York, NY, 132-141.
[24] Wildemuth, B. M. 2004. The effects of domain knowledge on search tactic formulation. *J. of the American Society for Information Science and Technology*. 55, 3 (Feb. 2004), 246-258.
[25] Wilson, M. J., and Wilson, M. L. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *J. of the American Society for Information Science and Technology*. 64, 2 (Feb. 2013), 291-306.
[26] Zhang, X., Cole, M., and Belkin, N. 2011. Predicting users' domain knowledge from search behaviors. In *Proceedings of SIGIR '11.* ACM, New York, NY, 1225-1226.

Table 3. Inventory of all questions used in this study, along with their response types and source

Category	Variable	ID	Question text	Scale or unit	Source
Topic knowledge	Perceived knowledge	**P1**	How much do you know about this topic?	1= nothing; … 5=I know a lot	Pre-search
	Interest in topic	**P2**	How interested are you to learn more about this topic?	1= not at all; … 5=very much	
	Perceived difficulty	**P3**	How difficult do you think it will be to search for information about this topic?	1= very easy; … 5=very difficult	
	Prior knowledge	**P4**	Please write what you know about this topic with 3-5 sentences.	Knowledge level coded as 0, 1, 2.	
Search for information exploration (Experience Factor 1)	Engagement in search	**Q8**	I was cognitively engaged in search task.	Rating on 5-point scale: 1= not at all; 2=unlikely; 3=somewhat; 4=likely; 5=very likely	Post-search
	Search effort	**Q9**	I made an effort at performing the search task.		
	Time well spent	**Q7**	The time for search was spent productively on meaningful tasks.		
	Concept relations	**Q11**	I was able to explore relationships among multiple concepts.		
	Topic scope expanded	**Q12**	I was able to expand the scope of my knowledge about the topic.		
	Synthesis	**Q20**	I feel that I was able to put together pieces of information into one big concept.		
User experience with search system (Experience Factor 2)	Like using system	**Q16**	I liked using this system to find information I needed.	Rating on 5-point scale: 1= not at all; 2=unlikely; 3=somewhat; 4=likely; 5=very likely	Post-search
	Needs well expressed in system	**Q17**	I feel that my needs were fully expressed using this system.		
	Easy to use the system	**Q18**	It was easy to use the system to express what I was looking for.		
	Topic understanding	**Q21**	I feel that I have full understanding of the topic of this task		
Learner interest and motivation (Experience Factor 3)	Increased interest	**Q13**	I became more interested in this topic.	Rating on 5-point scale: 1= not at all; 2=unlikely; 3=somewhat; 4=likely; 5=very likely	Post-search
	Willingness to find more information	**Q14**	I would like to find more information about this topic.		
	Willingness to share	**Q15**	I would like to share what I learned with my friends.		
	Learning useful information	**Q19**	I feel that I learned useful information as a result of this search.		
	Developing new ideas	**Q10**	I was able to develop new ideas or perspectives.		
Perceived learning	Self-reported learning score	**Q22**	How would you grade your learning outcome?	Score on 0-100 scale	Post-search
Perceived search success	Self-reported searching score	**Q23**	How would you grade your search outcome?	Score on 0-100 scale	Post-search
Lower-level cognitive learning assessment	Remember	**Q1**	Assessing to what extent a subject can remember specific elements about the topic.	Written response was analyzed by checking off 7 criteria	Post-search written test
	Understand	**Q2**	Assessing to what extent subjects could construct meaning about the topic.		
	Apply	**Q3**	Assessing to what extent subjects could carry out the concept in a given situation		
Higher-level cognitive learning assessment	Analyze	**Q4**	Assessing to what extent subjects could break content into an outline of a paper	Written response was analyzed by checking off 7 criteria	Post-search written test
	Evaluate	**Q5**	Assessing to what extent subjects could write what they learned from searching		
	Create	**Q6**	Assessing to what extent subjects could write creative questions		