Measuring Learning During Search: Differences in Interactions, Eye-Gaze, and Semantic Similarity to Expert Knowledge

Nilavra Bhattacharya
School of Information
The University of Texas at Austin
nilavra@ieee.org

Jacek Gwizdka
School of Information
The University of Texas at Austin
chiir2019@gwizdka.com

ABSTRACT

We investigate the relationship between search behavior, eye - tracking measures, and learning. We conducted a user study where 30 participants performed searches on the web. We measured their verbal knowledge before and after each task in a content-independent manner, by assessing the semantic similarity of their entries to expert vocabulary. We hypothesize that differences in verbal knowledge-change of participants are reflected in their search behaviors and eye-gaze measures related to acquiring information and reading. Our results show that participants with higher change in verbal knowledge differ by reading significantly less, and entering more sophisticated queries, compared to those with lower change in knowledge. However, we do not find significant differences in other search interactions like page visits, and number of queries.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; *Web-based interaction.*

KEYWORDS

search as learning; measuring knowledge change; human information behavior; eye-tracking

ACM Reference Format:
Nilavra Bhattacharya and Jacek Gwizdka. 2019. Measuring Learning During Search: Differences in Interactions, Eye-Gaze, and Semantic Similarity to Expert Knowledge. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3295750.3298926

1 INTRODUCTION

We investigate the relationship between search behavior, eye gaze, and learning. Marchionini described information seeking as "a process, in which humans purposefully engage in order to change their state of knowledge" [19]. Thus information search is driven by higher-level human needs. We can consider information seeking as a process that changes the state of a searcher's knowledge. This consideration points us to works that draw upon strong ties between information search and learning. For example, exploratory search has been described as search as learning [20]. More recently, Jansen et al. [16] noted that "a learning theory may better describe the information searching process than more commonly used paradigms of decision making or problem solving". I-LEARN model created by Neuman, from the learning theory perspective, attempts to connect learning theory with information science perspectives [22]. The model posits "the use of information as the fundamental building block for learning". If we consider learning as an integral part of information search process, the challenge then is how to measure learning. We can turn to the educational psychology literature, but we should first note that measuring a searcher's learning is actually not new in our field. For example, Pirolli et al. [26] measured user learning in their evaluation of Scatter/Gather exploratory search interface. Learning was assessed by examining gains in a user's understanding of the topic structure, and in a user's ability to formulate effective queries. Possible operationalizations of learning measurement include measures of a user's ability to formulate more effective queries, a user's familiarity with concepts, and relationships between concepts.

Our interest is in the learning that takes place at the bottom level of the modified Bloom's taxonomy – the remembering and factual knowledge level [1]. Therefore, we operationalize learning as changes in verbal knowledge [17] from before to after a search session. One goal of our work is to construct learning measures that require minimal input from users, and, for example, do not require users to answer topic-specific comprehension tests. We use two types of learning measures, a simple topic-independent measure, and a measure based on semantic similarity with expert vocabulary. Our expectation is that searchers who invest more search-effort, and who consume more result pages, learn more. In other words, their topical vocabulary improves and becomes more similar to expert vocabulary. Search-effort is conceptualized as a two part, multiple-component construct: (a) search interaction (i.e., visiting SERPs and result pages, entering queries); and (b) acquiring text from web pages by reading (i.e., number and duration of reading fixations, length of reading sequences, number and length of eye regressions).

2 BACKGROUND

Learning, in the context of interactive information retrieval, is interesting for a few reasons. Assessment of learning outcomes resulting from search, are a good candidate for more comprehensive, user-oriented evaluation measure of information retrieval systems. In particular, the need to perform evaluations that go beyond individual query interaction has been noted by many researchers, and some approaches have been proposed (e.g., most recently by Raman

CHIIR '19, March 10–14, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298926

et al. [27]). Measuring learning on whole, or multiple-session search offers one possible approach.

However, assessing learning typically requires collecting explicit (knowledge-based) responses from users. Collecting such responses may be disruptive, and while it works well in controlled research conditions, it is hard to transfer this approach to the field. Two questions arise: 1) which implicit measurement techniques of learning could be used; and 2) which techniques work outside the laboratory.

With recent technological advances in (psycho-physiological and brain) sensing techniques, tools that enable implicit assessment of changes in a user's cognitive state are now more readily available. One such tool is eye-tracking. Prior work [4] demonstrates feasibility of using eye gaze patterns to assess differences in the levels of users' domain knowledge (at least for text search). This work indicates a possibility of using eye tracking to measure changes in a searcher's learning. One advantage of using measurement techniques based on eye tracking is the possibility of scaling up to larger numbers of users. If the current trend in dropping eye tracker prices continues, we can expect that in-not-so-distant-future, eye trackers will become a selectable option for many computers, just like getting a larger hard drive. (A reasonably good, but slow, eye tracker currently costs few hundred USD). This line of research is still in early development and we need more studies to confirm reliability and to examine validity of using eye gaze patterns in measuring learning. Our work aims to contribute to these developments.

Past efforts have tried to gauge the existing domain knowledge or expertise of a user from interaction features. Wildemuth [30] observed that with change in expertise, novices tend to replicate the search tactics employed by domain experts. White et al. [29] predicted domain expertise from interaction measures like website visits, dwell time, focusing on single vs. multiple topics on search engine result pages (SERPs), etc. Cole et al. [4] identified that behavioral features were topic-agnostic predictive cues of a user's domain knowledge. Further, Zhang et al. [33] identified a separate set of features, like the rank of search-result rank considered relevant, and average query length, as being predictive of a user's knowledge. Other works have tried to measure the change in knowledge over a task-span, and correlate it with interaction measures. Collins-Thompson et al. [6] observed that diversity in a user's queries is an indicator of increased knowledge gain. Eickhoff et al. [7] studied correlations between search interactions, visited SERP features, and learning needs related to declarative or procedural knowledge. Vakkari [28] presented a set of features that are predictive of knowledge change during searches.

With the advent of the 'Search-as-Learning' (SAL) sub-field, new research has also investigated the measurement of knowledge change during online information search tasks [11, 12, 31, 32]. For instance, Gadiraju et al. [11] and Yu et al. [32] measured pre- and post-task knowledge levels using online tests, where participants had to choose between 'True', 'False' and 'I don't know' options for a series of factual questions related to the topic of the search task. Knowledge change was quantified as the difference between the pre- and post-task scores. Yu et al. [32] also proposed an automated system to predict this differential knowledge-change measure using a variety of interaction features as predictors. However, the drawbacks of this approach are that (a) it requires creation of domain specific knowledge-tests, (b) users get exposed to the topics of the task before searching starts, which may interact with the pre-task knowledge levels, and (c) in multiple-choice questions, users may select the correct answers by guessing. Ghosh et al. [12] takes a different approach for pre- and post-task assessments, by asking the participants about their perceived levels of existing knowledge, task difficulty, newly gained knowledge, and interest in the topic, all measured using five-point Likert scales. Knowledge change was operationalized as a paired sample t-test to determine if self-perceived existing knowledge and self-perceived new knowledge differ significantly. Although this approach avoids the need of domain specific tests, and exposing users to task-topics, it relies heavily on the user's perception of his/her own knowledge change. Such perceived measures are subjective to users, and may not provide a truly quantifiable indication of knowledge change.

In light of the above, we see that past works have either studied a limited range of learning objectives, or considered limited predictive features. Many of them required domain-specific information as well. In this work, we aim to contribute by investigating knowledge-change assessment metrics that (a) do not require domain-specific comprehension tests, (b) do not expose the user to the topic of the search in the pre-task assessment, and (c) attempt to assess a user's true knowledge-level (with respect to expert knowledge) with minimal scope for guessing or subjective differences.

3 METHOD

3.1 Experimental Design

We conducted a controlled, within-subject user study with 30 participants (16 females; mean age 24.5 years), who were asked to search for health-related information on the internet. We pre-screened our participants for (a) native-level English familiarity (to minimize influence of varied levels of English familiarity on their eye movements and fixations), (b) non-expert topic familiarity (so that they did not have extensive prior knowledge of the topic for which we were trying to measure their learning on search), and (c) uncorrected 20/20 vision (to minimize problems with eye-tracking). In the pre-screening survey, all participants reported using the internet for longer than an hour everyday, and daily usage of Google search engine. The majority of the participants had been using Google for longer than seven years, and considered themselves proficient in searching for information online.

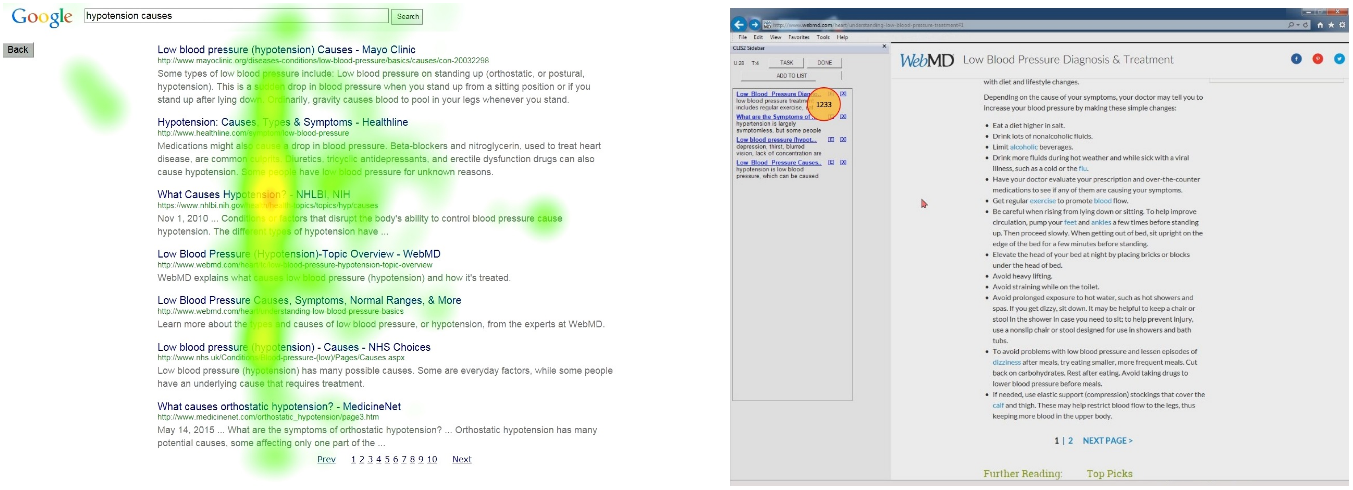
3.2 Task Description

To investigate changes in vocabulary knowledge, each participant performed two information search tasks on health-related topics in counterbalanced order. These two tasks simulated a work-task approach by triggering realistic information-need for participants [2], as they were asked to find useful information for helping a family member and a friend. The tasks were designed to be complex, and contained multiple facets. The prompts for each task are:

Task 1: *Vitamin A: Your teenage cousin has asked your advice in regard to taking vitamin A for health improvement purposes. You have heard conflicting reports about the effects of vitamin A, and you want to explore this topic to help your cousin. Specifically, you want to know:*

(i) What is the recommended dosage of vitamin A for underweight teenagers?

(ii) What are the health benefits of taking vitamin A? Please find at least 3 benefits and 3 disadvantages of vitamin A.



Low blood pressure (hypotension) Causes - Mayo Clinic
http://www.mayoclinic.org/diseases-conditions/low-blood-pressure/basics/causes/con-20032298
Some types of low blood pressure include: Low blood pressure on standing up (orthostatic, or postural, hypotension). This is a sudden drop in blood pressure when you stand up from a sitting position or if you stand up after lying down. Ordinarily, gravity causes blood to pool in your legs whenever you stand.

Hypotension: Causes, Types & Symptoms - Healthline
http://www.healthline.com/symptom/low-blood-pressure
Medications might also cause a drop in blood pressure. Beta-blockers and nitroglycerin, used to treat heart disease, are common culprits. Diuretics, tricyclic antidepressants, and erectile dysfunction drugs can also cause hypotension. Some people have low blood pressure for unknown reasons.

What Causes Hypotension? - NHLBI, NIH
https://www.nhlbi.nih.gov/health/health-topics/topics/hyp/causes
Nov 1, 2010 ... Conditions or factors that disrupt the body's ability to control blood pressure cause hypotension. The different types of hypotension have ...

Low Blood Pressure (Hypotension)-Topic Overview - WebMD
http://www.webmd.com/heart/tc/low-blood-pressure-hypotension-topic-overview
WebMD explains what causes low blood pressure (hypotension) and how it's treated.

Low Blood Pressure Causes, Symptoms, Normal Ranges, & More
http://www.webmd.com/heart/understanding-low-blood-pressure-basics
Learn more about the types and causes of low blood pressure, or hypotension, from the experts at WebMD.

Low blood pressure (hypotension) - Causes - NHS Choices
http://www.nhs.uk/Conditions/Blood-pressure-(low)/Pages/Causes.aspx
Low blood pressure (hypotension) has many possible causes. Some are everyday factors, while some people have an underlying cause that requires treatment.

What causes orthostatic hypotension? - MedicineNet
http://www.medicinenet.com/orthostatic_hypotension/page3.htm
May 14, 2015 ... What are the symptoms of orthostatic hypotension? ... Orthostatic hypotension has many potential causes, some affecting only one part of the ...

(a)

http://www.webmd.com/heart/understanding-low-blood-pressure-treatment#1

Low Blood Pressure Diagno..
low blood pressure treatment includes regular exercise, eat
What are the Symptoms of ..
hypertension is largely symptomless, but some people
Low blood pressure (hypot...
depression, thirst, blurred vision, lack of concentration are
Low Blood Pressure Causes..
hypotension is low blood pressure, which can be caused

Low Blood Pressure Diagnosis & Treatment

with diet and lifestyle changes.

Depending on the cause of your symptoms, your doctor may tell you to increase your blood pressure by making these simple changes:

- Eat a diet higher in salt.
- Drink lots of nonalcoholic fluids.
- Limit alcoholic beverages.
- Drink more fluids during hot weather and while sick with a viral illness, such as a cold or the flu.
- Have your doctor evaluate your prescription and over-the-counter medications to see if any of them are causing your symptoms.
- Get regular exercise to promote blood flow.
- Be careful when rising from lying down or sitting. To help improve circulation, pump your feet and ankles a few times before standing up. Then proceed slowly. When getting out of bed, sit upright on the edge of the bed for a few minutes before standing.
- Elevate the head of your bed at night by placing bricks or blocks under the head of bed.
- Avoid heavy lifting.
- Avoid straining while on the toilet.
- Avoid prolonged exposure to hot water, such as hot showers and spas. If you get dizzy, sit down. It may be helpful to keep a chair or stool in the shower in case you need to sit; to help prevent injury, use a nonslip chair or stool designed for use in showers and bath tubs.
- To avoid problems with low blood pressure and lessen episodes of dizziness after meals, try eating smaller, more frequent meals. Cut back on carbohydrates. Rest after eating. Avoid taking drugs to lower blood pressure before meals.
- If needed, use elastic support (compression) stockings that cover the calf and thigh. These may help restrict blood flow to the legs, thus keeping more blood in the upper body.

(b)

Figure 1: Screenshots of our system: (a) Customized Google SERP, with 7 results per page, and no advertisements. (b) A 'CONTENT' page showing our customized left-sidebar, for viewing the task-prompt, creating bookmarks and taking notes. Green/yellow patches in (a) are eye-tracking fixation heatmaps. The circle with number in (b) is an eye-fixation with duration.

(iii) What are the consequences of vitamin A deficiency or excess? Please find 3 consequences of vitamin A deficiency and 3 consequences of its excess.

(iv) Please find at least 3 food items that are considered as good sources of vitamin A.

Task 2: *Hypotension: Your friend has hypotension. You are curious about this issue and want to investigate more. Specifically, you want to know:*

(i) What are the causes of hypotension?

(ii) What are the consequences of hypotension?

(iii) What are the differences between hypotension and hypertension in terms of symptoms? Please find at least 3 differences in symptoms between them.

(iv) What are some medical treatments for hypotension? Which solution would you recommend to your friend if he/she also has a heart condition? Why?

3.3 Apparatus

The main search tasks were performed using Internet Explorer. The tasks started from a customized version of the Google search engine interface (Fig. 1a), and the browser had an additional sidebar on the left (Fig. 1b). In our custom-written search engine interface, search results were retrieved from Google in real-time in the background by a proxy server. A search engine result page (SERP) controlled the display of the search results, by showing only seven results per page. This ensured that eye fixations were accurately tracked on each individual result in the SERP, and no advertisements could distract our participants. We chose seven results per page as this allowed the search results to have an optimum increased font size (and thereby increased visual angle), so that we could track eye-movements at the level of individual elements of search results surrogates. All other webpages (which we call 'CONTENT' pages) were displayed in their true form.

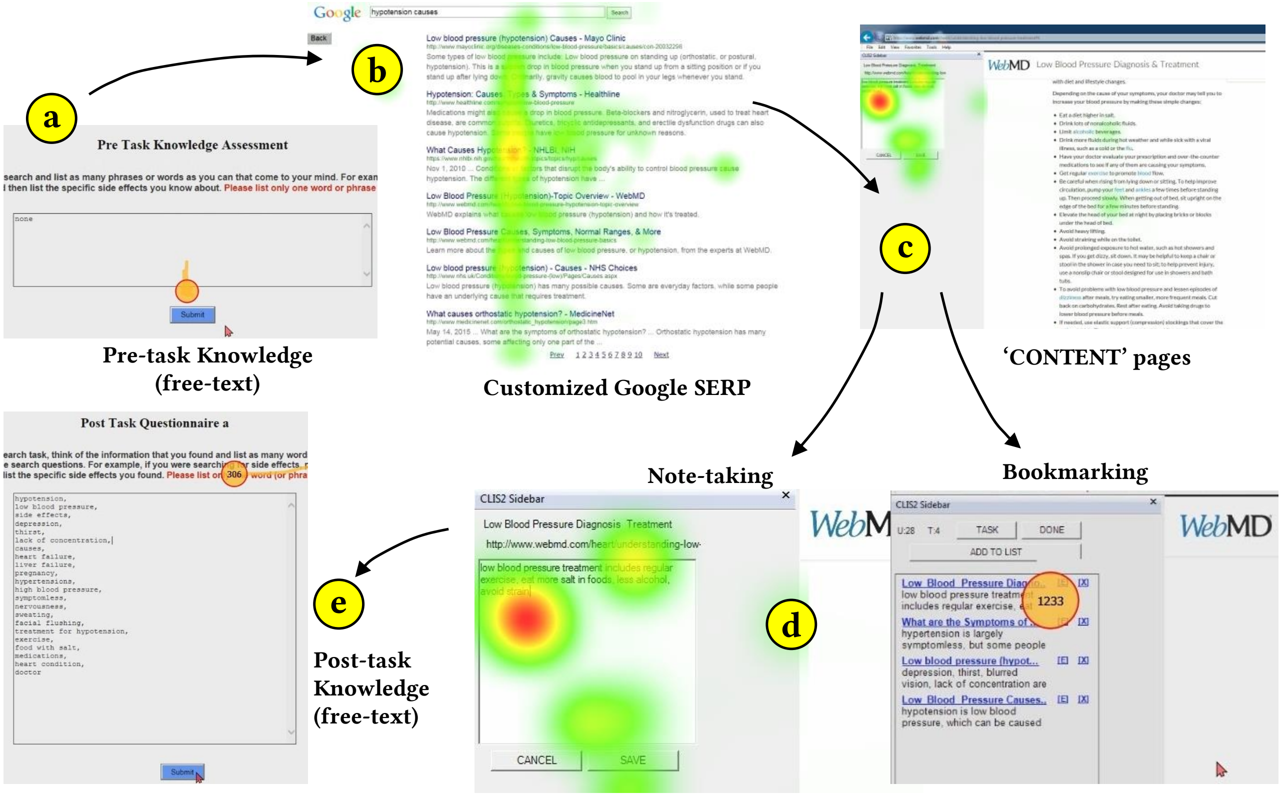
The sidebar on the left showed (on demand) the current search task prompt on the top, and had bookmarking and note-taking sections below. Bookmarking allowed the participants to save the URLs of webpages they opened and considered relevant. The list of bookmarked pages were available on the sidebar all throughout a search-task session, and a bookmarked page could be re-opened instantly by clicking the bookmark. The note-taking feature allowed participants to write and / or copy-paste relevant text from the webpages they visited. When the participants were performing the search tasks, we recorded their interactions with the computer system. This included eye gaze, keystrokes, mouse clicks, and other activities like bookmarks, notes, search queries, URLs of pages visited, etc., as well as the timing and duration of these activities. Eye gaze was captured using a Tobii TX-300 eye-tracker (Tobii Technology AB, Sweden) controlled by iMotions software (iMotions A/S) that also captured all user interactions.

3.4 Procedure

Each experimental session started with the assessment of participant's working memory capacity (WMC) using memory span task [8] and health literacy using the eHealth Literacy Scale (eHEALS) [23]. Next, the participants performed a training task to familiarize themselves with the custom user interfaces (bookmarking and note-taking), and the study procedure. After the training, participants started the main search tasks. Task steps are illustrated in Fig. 2.

Each task started with a Pre-task knowledge assessment, to gauge the existing or initial knowledge of the participant for the task (Fig. 2a). The prompt for the pre-task assessment was as follows:

> *Think of what you already know on the topic of this search and list as many phrases or words as you can that come to your mind. For example, if you know about side effects, please do not just type the phrase "side effects", but rather type "side effects" and then list the specific side effects you know about. Please list only one word or phrase per line and end each line with a comma.*


a
b
c
d
e
Pre Task Knowledge Assessment
Submit
Pre-task Knowledge (free-text)
hypotension causes
Search
Back
Customized Google SERP
WebMD Low Blood Pressure Diagnosis & Treatment
'CONTENT' pages
Note-taking
Bookmarking
CLIS2 Sidebar
CANCEL
SAVE
U:28 T:4
TASK
DONE
ADD TO LIST
1233
306
Post Task Questionnaire a
Post-task Knowledge (free-text)

earch and list as many phrases or words as you can that come to your mind. For exan
l then list the specific side effects you know about. Please list only one word or phrase

none

Low Blood Pressure Diagnosis Treatment
http://www.webmd.com/heart/understanding-low-

low blood pressure treatment includes regular exercise, eat more salt in foods, less alcohol, avoid strain

Low Blood Pressure Diagno... [E] [X]
low blood pressure treatment includes regular exercise, eat
What are the Symptoms of ... [E] [X]
hypertension is largely symptomless, but some people
Low blood pressure (hypot... [E] [X]
depression, thirst, blurred vision, lack of concentration are
Low Blood Pressure Causes.. [E] [X]
hypotension is low blood pressure, which can be caused

earch task, think of the information that you found and list as many word
e search questions. For example, if you were searching for side effects, r
list the specific side effects you found. Please list on word (or phra

```
hypotension,
low blood pressure,
side effects,
depression,
thirst,
lack of concentration,
causes,
heart failure,
liver failure,
pregnancy,
hypertensions,
high blood pressure,
symptomless,
nervousness,
sweating,
facial flushing,
treatment for hypotension,
exercise,
food with salt,
medications,
heart condition,
doctor
```

Figure 2: Procedure for each of the search tasks: (a) pre-task knowledge assessment, (b) searching for information using Google, (c) visiting search result webpages, (d) bookmarking webpages that contain relevant information, and also taking notes, and (e) assessing the knowledge change through a post-task questionnaire.

After the pre-task assessment, the online searching started. Participants searched for, and visited publicly available webpages using Google. We asked them to bookmark a webpage if they considered it relevant to their task, and (optionally) to take notes of the information found on the webpage with respect to information need prompted by a task scenario. The bookmarking and note-taking features were mechanisms to help participants engage more in the simulated tasks in the lab (motivating task scenarios, as discussed in [2]). Since the tasks motivated the participants to find information for helping their friends and family, the notes were meant to help them remember and share the information they found.

We classified each visited webpage as a 'SERP' (Search Engine Result Page) or a 'CONTENT' page, based on whether it was a Google SERP, or any other page, respectively. A CONTENT page was further marked as 'RELEVANT', if the participant bookmarked the page.

At the end of each task, the participants completed a Post-task knowledge assessment. This served as the final check of the participant's knowledge for the topic of the search session. The prompt for the post-task assessment was as follows:

> *Now that you have completed this search task, think of the information that you found and list as many words or phrases as you can on the topic of the search task. This will be short ANSWERS to the search questions. For example, if you were searching for side effects, please do not just type the phrase "side effects", but rather type "side effects" and then list the specific side effects you found. Please list only one word (or phrase) per line and end each line with a comma.*

After the Post-task assessment for each task, participants were presented with the NASA-TLX (Task Load Index) [15], which measured their perceived workload invested in the search task.

A key difference of our study from recent works like [11, 32] is that we aimed to capture knowledge change in the form of free-recall from memory, so that it is independent of the task topic. (Although participants took notes while searching, they were not allowed to consult these notes during the post-task assessment, as we were trying to capture how much of the information they learnt, and remembered). There was no time-limit on any portion of the task, including the Pre- and Post-task knowledge assessments. A session typically lasted for 1.5 to 2 hours. On completion of the entire study session, each participant received $25.

4 MEASURES

We have calculated our variables for each search task separately. We argue that for a user having minimal topical expertise, knowledge of disparate topics (e.g. Vitamin A vs. hypotension) are independent of each other. Therefore, a *user-task pair* is our unit of analysis, and all the measures described below are calculated for each of such user-task pairs.

4.1 Knowledge Change Measures (KC)

Our aim was to measure the difference in a user's knowledge of vocabulary on a searched topic, before and after a search task. So we needed two measurement points, and deliberated on a variety of possibilities for assessing participants' knowledge levels. Fact-checking questions before a task were considered inappropriate, as we wanted to avoid exposing the participant to the topic's content before they start the search. Since the tasks were conducted on the open web, we could not use methods like Sentence Verification Technique (SVT) [10], which requires creation of questions for each document. Our participants were not experts on the topics, hence concept maps and mind-mapping were deemed inappropriate, as they are particularly difficult to score for non-experts. Thus we settled for the Pre- and Post-Task knowledge assessments, where participants were asked to free-recall as many words or phrases on the topic of the task as they could, without time limit.

4.1.1 Simple Knowledge Change Measure:

Our first knowledge change measure, *KC_Simple*, is a simple, topic-independent measure. It was calculated as the relative difference in the number of items (words or phrases) entered by users before and after each task.

4.1.2 Expert Knowledge Creation:

In order to support more sophisticated measure of knowledge change, we created a vocabulary of expert words and phrases on the search task topics. The vocabulary was crowd-sourced on the Amazon Mechanical Turk (AMT) platform, and then verified in consultation with a medical doctor. A separate question for each facet of our search tasks was administered on AMT. The questions asked crowd-workers to enter as many words or phrases as they could on the topic of a task facet, along with providing links to web-sources for this information. We received responses from 156 and 91 crowd-workers on Task 1 and 2, respectively. The responses yielded 474 and 171 words / phrases on the topics of Task 1 and 2. After removing duplicates, the lists were reviewed and cleaned in consultation with a medical doctor. The final result was two lists of 115 and 105 words or phrases, for Task 1 and 2, respectively.

4.1.3 Semantic Similarity with Expert Knowledge:

Semantic similarity between two pieces of text (words / phrases / sentences / documents) measures how similar the two texts are, in terms of their meaning, rather than their syntactical representation (e.g. their string format) [14]. To measure the knowledge-change of our participants, we calculated how much their Pre- and Post-task free-recall entries were semantically similar to our curated expert vocabulary. For this purpose we used the state-of-the-art Universal Sentence Encoder [3] from Google's TensorFlow Hub, which encodes natural texts into a 512-dimensional embedding vector. We did not use popular word-embedding models like word2vec [21] and GloVe [25] because they are more suited for single word comparisons, while we were measuring similarity between phrases and/or sentences. The universal sentence encoder is pre-trained and optimized for greater-than-word length text, like sentences, phrases or short paragraphs. Before encoding our texts (participant input or expert knowledge) into sentence embedding vectors, we removed stopwords from the text using the English stopword list from the popular python package, Scikit-Learn [24]. Instead of using the raw cosine value of similarity between two sentence embedding vectors $\mathbf{u}$ and $\mathbf{v}$ as a measure of their semantic similarity, we used the angular similarity between the vectors, $\mathrm{sim}(\mathbf{u}, \mathbf{v})$ (Eqn. 1), as it provides better discrimination power than the raw cosine value.

$$sim(\mathbf{u}, \mathbf{v}) = \left(1 - \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}\right) / \pi\right) \tag{1}$$

First we calculated angular similarity between Pre-task entries and expert vocabulary $pre\_exp\_sim = sim(\mathbf{pre\_task}, \mathbf{expert})$, and angular similarity between Post-task entries and expert vocabulary $post\_exp\_sim = sim(\mathbf{post\_task}, \mathbf{expert})$,. Then we calculated our knowledge change measures by taking the difference (*KC_Sem_Diff*) between them (Eqn. 2), as well as their ratio (Eqn. 3) (*KC_Sem_Ratio*).

$$KC\_Sem\_Diff = post\_exp\_sim - pre\_exp\_sim \tag{2}$$

$$KC\_Sem\_Ratio = \frac{post\_exp\_sim}{pre\_exp\_sim} \tag{3}$$

4.2 Eye-tracking Measures (ET)

The eye-tracking variables we use reflect the process of reading. We calculated them on SERPs and CONTENT pages, and separately, on CONTENT pages relevant to a task, since we assume that participants learn most from reading such web pages. We label eye fixations as "reading", when a person is reading words sequentially in a horizontal line on a web page (in contrast with scanning text, when eyes are being fixated on isolated words) [4, 5], and we use only "reading fixations" in our measures.

We calculated total duration, and count of fixations on all pages of given a type, as well per page. We also calculated total length of reading sequences (in pixels), number of eye regressions (when eyes are moving back to fixate on a previously seen word), and the length of regressions (in pixels).

4.3 Search Interaction Measures (SI)

Search behavior was characterized by typical measures that included number of visited SERPs and CONTENT pages (we counted visits as well as revisits that were longer than 300ms and associated with at least two fixations), dwell time on pages by their type, number of queries, type of query reformulations [18], and commonality of words used in queries. This commonality was measured as the harmonic mean of the word usage-frequencies (obtained from the Google Web Trillion Word Corpus [9] which contains approximately 13.5 mln. unique words) of the words used in queries, with a lower number indicating that less common, or more specialized words were used in queries.

Search-effort is operationalized as a two part, multiple-component construct, composed of the above two groups of measures, SI and ET: (a) search interaction (i.e., visiting SERPs, clicking links and visiting content pages (result pages), entering queries); and (b) acquiring text from web pages (i.e., number and duration of reading fixations, length of reading sequences, number and length of eye regressions).

Table 1: Selected results showing differences in the search-interactions (SI) and Eye-tracking (ET) variables, for the HI and LO knowledge-change groups based on our proposed knowledge-change (KC) measures. More differences were observed in ET measures, and less for SI measures.

Category	Measure Name	*KC_Simple*			*KC_Sem_Ratio*			*KC_Sem_Diff*			Description
		LO: Mean (SD)	HI: Mean (SD)	M-W |z|	LO: Mean (SD)	HI: Mean (SD)	M-W |z|	LO: Mean (SD)	HI: Mean (SD)	M-W |z|	
Participant	*eHEALS*	36.56 (5.8)	38.25 (4.58)	1.40	36 (6.65)	38.83 (2.66)	1.50	36.72 (5.77)	38.08 (4.68)	1.04	eHealth Literacy Score of each participant
	WMC	4.78 (.94)	4.85 (1.1)	0.11	4.84 (1.04)	4.79 (1)	0.20	4.86 (1.02)	4.77 (1.03)	0.32	Working Memory Capacity (WMC) for remembering long and short words
Webpage (SI)	*pg_n*	24.72 (14.77)	28.62 (14.18)	1.07	26.8 (16.67)	26.45 (12.1)	0.46	26.52 (16.78)	26.75 (11.95)	0.71	Total number of webpages visited
	pg_serp_n	11.8 (6.25)	11.66 (7.87)	0.38	12.48 (7.93)	10.95 (5.98)	0.33	12.44 (7.93)	11 (5.99)	0.27	Number of SERPs
	pg_content_n	11.52 (9.17)	15.33 (9.37)	1.54	12.88 (9.9)	13.91 (8.97)	0.50	12.64 (10.05)	14.16 (8.76)	0.81	Number of CONTENT pages
	pg_content_rel_n	9.72 (13.67)	8.33 (5.25)	0.46	10.32 (13.85)	7.7 (4.45)	0.07	10.08 (13.93)	7.95 (4.34)	0.39	Number of CONTENT pages bookmarked (i.e. considered relevant)
Query (SI)	*query_n*	6.04 (3.36)	6.12 (3.62)	0.15	6.08 (3.76)	6.08 (3.18)	0.25	6.12 (3.78)	6.04 (3.15)	0.19	Number of queries entered
	qr_new_n	0.6 (1.25)	0.83 (.96)	1.49	.48 (.82)	.95 (1.33)	1.35	.4 (.76)	1.04 (1.33)	*1.95^*	Query reformulations: number of new queries
	search_effectiveness	1.63 (1.49)	2.1 (2.81)	0.06	1.77 (1.63)	1.95 (2.75)	0.31	1.74 (1.65)	1.98 (2.73)	0.17	Ratio of RELEVANT CONTENT pages visited to number of queries entered
	q_words_freq	9,203k (14,126k)	8,141k (12,601k)	0.78	10,325k (14,079k)	6,973k (12,442k)	*1.76^*	10,845k (14,087k)	6,431k (12,255k)	**2.16** *	Harmonic mean of usage-frequencies of words (from Google n-gram corpus [9]) used in queries. Lower number means less common (more specialized) words were used in queries.
Eye-tracking (ET)	*rseq_len*	107k (50k)	80k (37k)	**2.08** *	106k (50k)	80k (37k)	**2.06** *	106k (50k)	81k (37k)	**1.98** *	Total length of scan-paths obtained by joining 'reading' fixation points (in k-pixels)
	regr_n	199.52 (94.05)	156.58 (78.42)	1.63	202.28 (96.99)	153.7 (72.72)	*1.75^*	202.08 (97.24)	153.91 (72.53)	*1.71^*	Number of backward regressions
	regr_len	43k (20k)	31k (15k)	**2.04** *	42k (20k)	31k (15k)	*1.80^*	42k (20k)	31k (15k)	*1.78^*	Total length of regressions (in k-pixels)
	fix_dur_content_sum	475451 (170890)	439282 (212973)	1.32	490878 (166768)	423212 (212294)	**2.00** *	483675 (177089)	430715 (205695)	*1.78^*	Total duration of reading fixations on CONTENT pages, summed across all such pages (in ms)
	fix_dur_content_avg	25673 (22351)	15607 (9774)	**2.02** *	22908 (18316)	18486 (17615)	*1.80^*	22698 (18450)	18706 (17523)	1.54	Total duration of reading fixations on CONTENT pages, averaged across all such pages (in ms)
	fix_n_content_avg	92.05 (81.38)	56.57 (31.46)	*1.90^*	82.32 (67.59)	66.71 (60.6)	*1.74^*	82.29 (67.61)	66.75 (60.58)	*1.72^*	Number of reading fixations per CONTENT page, averaged across pages visited
	pRR_serp	.29 (.1)	.23 (.11)	**2.10** *	.28 (.1)	.23 (.12)	1.50	.28 (.1)	.23 (.12)	1.60	Probability of continuing to read on SERPs, averaged across SERPs visited
	pRR_content	.3 (.06)	.34 (.12)	0.54	.3 (.07)	.34 (.11)	0.64	.3 (.07)	.34 (.11)	1.02	Probability of continuing to read on CONTENT pages, averaged across pages visited
Workload	*NASA_TLX*	2.98 (.7)	3.25 (.7)	1.18	2.87 (.62)	3.36 (.71)	**2.16** *	2.88 (.62)	3.36 (.72)	**2.11** *	NASA Task Load Index average score

For Mann Whitney (M-W) statistics, (green cells with *) indicates $p < .05$, and (yellow cells with ^) indicates approaching .05 significance ($.05 \leq p < .1$).

5 RESULTS

In the context of our research questions, the ET and SI measures are our dependent variables, The three KC measures constructed from Pre- and Post-task responses are our independent variables. We calculated these variables separately for each task, because we argue that knowledge of disparate topics are independent of each other, and that task-topics can interact with participants' knowledge, and motivate their cognition differently. So we consider a user-task pair to be our unit of analysis. Since 30 participants performed two tasks each, there were 60 user-task pairs. Due to technical difficulties during the study (computer crash, noisy eye-tracking data, etc.) some data had to be discarded, and usable data is available for 49 user-task pairs (26 for Task 1 and 23 for Task 2). The analysis reported in this section are performed on these 49 units of analysis.

For each of the three KC measures, we partitioned the user-tasks into a LO group and a HI group, based on median-split of the particular KC score. Thus, we had three LO groups and three HI groups in total, corresponding to each KC measure.

Since our knowledge change measures are new, we do not fully know their properties. So we checked in how many instances the same user-task pair was placed in different groups (LO vs HI) by the different KC measures. We saw that the LO and HI groups were nearly identical for the two semantic similarity measures (*KC_Sem_Diff* and *KC_Sem_Ratio*) and differed only in 2/49 cases, while they were slightly different when semantic similarity measures were compared with the simple measure (*KC_Simple*) (9/49 cases). We conclude that our KC measures assess changes in topic vocabulary in similar way.

Since the measures were not normally distributed, we used non-parametric Mann-Whitney U tests to compare differences between the groups reflected in the ET and SI measures described in previous section. The results and test statistics are reported in Table 1, where the significant results are marked.

There were no differences between the eHEALS and WMC scores for the LO and HI groups, and there were no differences between the groups for most search interaction measures. LO and HI users visited about the same number of SERPs, and CONTENT pages, and marked about the same number of the latter as relevant. There was also no difference in time spent on both types of pages (total and average), no difference in the number of queries entered, and no difference in search effectiveness (number of relevant pages found per query). However, the LO group entered fewer new queries in reformulations (significant difference for *KC_Sim_Diff*, and approaching significance for *KC_Sim_Ratio*), and they used more common (or less specialized) words in their queries (*q_words_freq*). The commonality of words was measured using word-usage frequency in English, from the Google Web Trillion Word Corpus [9] (Sec. 4.3).

On CONTENT pages, LO group fixated more, had higher duration of reading fixations, and higher duration of fixations per page. In particular, on relevant CONTENT pages, the LO group had longer overall reading sequences and more eye regressions in reading. But there were no such differences in reading SERPs. On the other hand, LO users reported lower workload than HI users.

We note that these differences generally have similar pattern for our simple knowledge change measure (*KC_Simple*), as well as for our more sophisticated measure based on semantic similarity (*KC_Sem_Diff* and *KC_Sem_Ratio*).

In addition to examining differences between groups based on knowledge change measures, we separately examined differences between the two groups split based on the semantic similarity of their Post-task entries to expert vocabulary (*post_exp_sim*). This measurement, taken at one point in time after each task, represents assessment of similarity of users' vocabulary to that of experts. The results are in Table 2. The LO group visited fewer CONTENT pages, but about the same number of SERPs. LO found fewer relevant CONTENT pages, but entered about the same number of queries and used much fewer new queries in their reformulations: thus the LO group had lower search effectiveness. This group also reported lower mental workload on search tasks (approaching significance). The LO tended to continue reading on SERPs (*pRR_serp*), but had lower probability of reading on CONTENT pages (*pRR_content*). In contrast the HI group had higher probability to continue reading CONTENT pages but lower on SERPs. LO group had longer total fixation duration and more fixations on SERPs, but no such difference was found for CONTENT pages.

We did not analyze the content of the notes taken by the participants, as we saw that most of the notes were directly copy-pasted from the webpages they visited, rather than their own assimilation of the information they found.

Table 2: Selected results showing differences in some search-interactions (SI) and Eye-tracking (ET) variables, for the HI and LO groups based on similarity of post-task to expert knowledge.

Measure Name	***post_exp_sim***		
	LO: Mean (SD)	**HI: Mean (SD)**	**M-W |z|**
pg_n	24.44 (15.5)	28.91 (13.24)	*1.65^*
pg_serp_n	12.48 (7.73)	10.95 (6.25)	0.55
pg_content_n	10.4 (8.6)	16.5 (9.29)	**2.71***
pg_content_rel_n	8.16 (13.64)	9.95 (5.29)	**2.68***
qr_new_n	.28 (.79)	1.16 (1.23)	**3.29***
search_effectiveness	1.38 (1.47)	2.36 (2.75)	**2.19***
fix_dur_serp_avg	3356.92 (1408.21)	2681.13 (1377.82)	*1.72^*
fix_n_serp_avg	12.51 (5.4)	9.94 (5.25)	*1.78^*
pRR_serp	.3 (.11)	.22 (.1)	**2.39***
pRR_content	.27 (.06)	.37 (.1)	**3.37***
NASA_TLX	2.98 (.73)	3.25 (.67)	1.35

For Mann Whitney (M-W) statistics, (*) indicates $p < .05$, and (ˆ) indicates approaching .05 significance ($.05 \leq p < .1$).

6 DISCUSSION

The observed differences between LO and HI groups do not seem to be related to differences in their health literacy or working memory capacity. The measured knowledge change was based on lower-level vocabulary change and was assessed only during a relatively short duration of experimental session. This may be a reason why we did not observe effects of literacy in the general search-task-area (health literacy) on knowledge change. It is also plausible that our participant sample was too uniform to observe effects of their general mental capacity, such as working memory capacity (WMC).

The difference in reading behavior between LO and HI groups on CONTENT pages is in an unexpected direction. The LO group generally spent more time on reading CONTENT pages, and moved their eyes backward in reading sequences more often, and by a longer distance than the HI group. Thus, as reflected in ET measures, the LO group put more effort into reading on the tasks, yet our KC measures indicate that they learned less. This is a likely indication that Lo group had more difficulty in acquiring information, and that in spite of investing more effort, they learned less.

Queries entered by HI group contained more specialized vocabulary than LO group. This evidence, taken together with ET measures, show that difficulty experienced by LO group in acquiring information may be indicative of lower general verbal skill of LO group. We have not measured this individual difference in the current study.

The higher mental workload reported by HI group may be, in part, a result of higher effort needed to produce more specialized queries. Prior work showed that query production is associated with higher levels of cognitive load in web search [13].

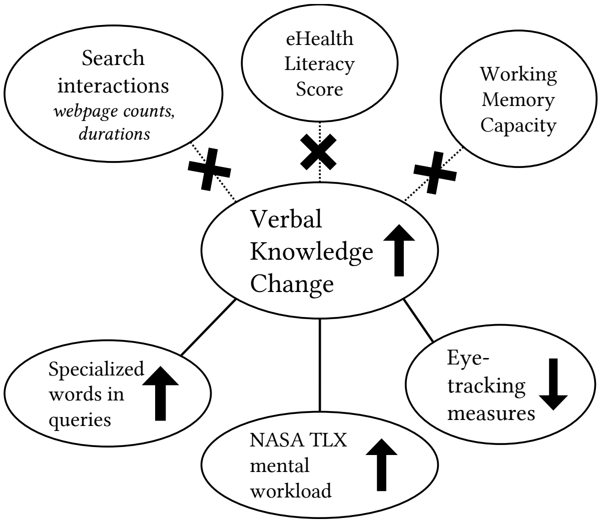


Figure 3: Summary of our findings. Arrows represent directions of relationship between measurement categories, while crosses represent no relationship.

Our findings in this work are in line with what was observed in past efforts to measure knowledge change. For instance, Yu et al. [32] observed that the total, average and maximum time spent on webpages have the highest predictive power for measuring knowledge change. Similarly, we also report significant differences in (a) *fix_dur_content_sum* between HI and LO groups of *KC_Sem_Ratio*, and (b) *fix_dur_content_avg* between HI and LO groups of *KC_Simple*. However, interestingly for us, the direction of difference in our study is opposite to intuition, i.e. users with higher knowledge gain spent less time on reading in CONTENT pages. It is important to remember however, that KC measure of Yu et al. [32] was based on a series of factual questions related to the topic of the search task, whereas ours is a topic-independent measure. Yu et al. [32] further observed that count of unique terms used in queries was the only query-related feature that showed predictive power. This is also corroborated by our finding that *q_word_freq* differs significantly for HI and LO groups based on *KC_Sem_Diff*, and is tending towards significance for *KC_Sem_Ratio* groups. In other words, people with increased knowledge change used less frequent words to perform their searches, which indicates that they were using specialized vocabulary. Similar to Yu et al's [32] random forest model showing that counts and percentages of webpages and SERPs visited are very weak predictors of knowledge change, our results also show no significant differences in number of visited SERPs and CONTENT pages between people who learned more and learned less.

The one-point measure of semantic similarity between participant's post-task free-recall and expert vocabulary (*post_exp_sim*) showed interesting relationships between LO/HI groups, and the SI and ET measures. The LO group were SERP readers who opened fewer CONTENT pages, found fewer relevant pages, and acquired less vocabulary. They entered about the same number of queries and thus visited about the same number of SERPs as HI group. But the LO group took more time to read the SERPs rather than investing more time in reading CONTENT pages, which is what HI group did. The LO group was apparently unable to identify more relevant results in-spite of investing effort in reading SERPs. The result was that words and phrases produced by HI group in free-recall were semantically closer to expert vocabulary than free-recall entries produced by LO group. Thus we could plausibly speculate that reading CONTENT pages rather than SERPs positively affects gaining more specialized vocabulary.

Fig. 3 highlights the findings of our study: users who scored higher on our knowledge change measure used less frequent / uncommon words in their queries, did lesser amount of reading on webpages, and reported higher mental workload, than those who scored less. No significant differences were found between groups w.r.t. search interactions, online health literacy, and working memory capacity.

7 CONCLUSION

Our new measures of knowledge change are related to differences in reading and querying behavior. But they uncovered some unexpected relationships, and plausibly tapped into interaction effects of search behavior, and reading with individual differences (such as, verbal ability), which were not measured.

Limitations of our work include using only two search tasks that were of similar nature (limited to health related topics), performing data analysis at the task level, a relatively uniform group of participants, and a short-time frame of experimental session. In future we plan to use a wider range of tasks, more diverse participant samples, additional individual difference tests, such as assessment of verbal skills, and conduct multiple-session study so that learning could be measured over a longer period of time.

ACKNOWLEDGEMENTS

We thank Dr. Andrzej Kahl, our medical doctor consultant for helping to review and clean the crowd responses during expert knowledge corpus creation; and Yinglong Zhang, for contributing to experimental data collection, and crowdsourcing of expert vocabulary. We also thank An Thanh Nguyen for suggesting the Universal Sentence Encoder from TensorFlow Hub.

The research described in this paper was partially funded by IMLS Award #RE-04-11-0062-11 to Jacek Gwizdka. We also gratefully acknowledge the ACM SIGIR Student Travel Grant awarded to Nilavra Bhattacharya.

REFERENCES

[1] Lorin W. Anderson, David R. Krathwohl, and Benjamin Samuel Bloom. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Longman. 08901.

[2] P. Borlund. 2003. The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research* 8, 3 (2003), paper no. 152.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, and others. 2018. Universal Sentence Encoder. *arXiv preprint arXiv:1803.11175* (2018).

[4] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. 2013. Inferring User Knowledge Level from Eye Movement Patterns. *Information Processing & Management* 49, 5 (Sept. 2013), 1075–1091.

[5] Michael J. Cole, Jacek Gwizdka, Chang Liu, Ralf Bierig, Nicholas J. Belkin, and Xiangmin Zhang. 2011. Task and User Effects on Reading Patterns in Information Search. 23, 4 (2011), 346–362. https://doi.org/10.1016/j.intcom.2011.04.007 0016.

[6] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 163–172.

[7] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 223–232.

[8] Greg Francis, Ian Neath, and Daniel R. VanHorn. 2008. *CogLab on a CD*. Wadsworth/Thomson Learning. 00019.

[9] Alex Franz and Thorsten Brants. 2006. All Our N-Gram Are Belong to You. *Google Machine Translation Team* 20 (2006).

[10] Luanne Freund, Rick Kopak, and Heather O'Brien. 2016. The Effects of Textual Environment on Reading Comprehension: Implications for Searching as Learning. *Journal of Information Science* 42, 1 (Feb. 2016), 79–93. 00001.

[11] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 2–11. https://doi.org/10.1145/3176349.3176381

[12] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching As Learning: Exploring Search Behavior and Learning Outcomes in Learning-Related Tasks. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 22–31. https://doi.org/10.1145/3176349.3176386

[13] Jacek Gwizdka. 2010. Distribution of Cognitive Load in Web Search. *Journal of the American Society for Information Science and Technology* 61, 11 (2010), 2167–2187. https://doi.org/10.1002/asi.21385

[14] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2015. Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies* 8, 1 (2015), 1–254.

[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[16] Bernard J. Jansen, Danielle Booth, and Brian Smith. 2009. Using the Taxonomy of Cognitive Learning to Model Online Searching. *Information Processing & Management* 45, 6 (Nov. 2009), 643–663. https://doi.org/10.1016/j.ipm.2009.05.004

[17] Kurt Kraiger, J. Kevin Ford, and Eduardo Salas. 1993. Application of Cognitive, Skill-Based, and Affective Theories of Learning Outcomes to New Methods of Training Evaluation. *Journal of Applied Psychology* 78, 2 (1993), 311–328.

[18] Chang Liu, Jacek Gwizdka, Jingjing Liu, Tao Xu, and Nicholas J. Belkin. 2010. Analysis and Evaluation of Query Reformulations in Different Task Types. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47 (ASIS&T '10)*. American Society for Information Science, Silver Springs, MD, USA, 17:1–17:10. https://doi.org/10.1002/meet.14504701214

[19] Gary Marchionini. 1997. *Information Seeking in Electronic Environments*. Cambridge University Press. Cited by 1717.

[20] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (April 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[22] Delia. Neuman. 2011. *Learning in Information-Rich Environments: I-LEARN and the Construction of Knowledge in the 21st Century*. Springer, New York. 00011.

[23] Cameron D Norman and Harvey A Skinner. 2006. eHEALS: the eHealth literacy scale. *Journal of medical Internet research* 8, 4 (2006).

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[26] Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. 1996. Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. 213–220.

[27] Karthik Raman, Paul N. Bennett, and Kevyn Collins-Thompson. 2013. Toward Whole-Session Relevance: Exploring Intrinsic Diversity in Web Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 463–472. https://doi.org/10.1145/2484028.2484089 Cited by 0000.

[28] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *Journal of Information Science* 42, 1 (2016), 7–18.

[29] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*. ACM Press, Barcelona, Spain, 132.

[30] Barbara M. Wildemuth. 2004. The Effects of Domain Knowledge on Search Tactic Formulation. *Journal of the American Society for Information Science and Technology* 55, 3 (Feb. 2004), 246–258.

[31] Ran Yu, Ujwal Gadiraju, and Stefan Dietze. 2018. Detecting, Understanding and Supporting Everyday Learning in Web Search. *arXiv:1806.11046 [cs]* (June 2018). arXiv:cs/1806.11046

[32] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting User Knowledge Gain in Informational Search Sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 75–84. https://doi.org/10.1145/3209978.3210064

[33] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. 2011. Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 1225–1226.