Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment

Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, Claudia Hauff
Delft University of Technology
Delft, The Netherlands
{n.roy,m.valletorre,u.k.gadiraju,d.m.maxwell,c.hauff}@tudelft.nl

ABSTRACT

Active reading strategies—such as content annotations (through the use of highlighting and note-taking, for example)—have been shown to yield improvements to a learner's knowledge and understanding of the topic being explored. This has been especially notable in long and complex learning endeavours. With web search engines nowadays used as the primary gateway for learners (or users) to find content that helps them realise their learning goals, they are often poorly equipped with the necessary tools to aid in sense-making, an important aspect of the *Search as Learning (SAL)* process. Within the *Information Retrieval (IR)* community, research efforts have explored ways to keep track of users' search context by providing a notepad-like interface for the collection of relevant articles, and aid them during the exploratory search process. However, these studies did not explicitly measure the effect that such tools have on knowledge and understanding during a complex, learning-oriented search task. In this paper, we address this research gap by carrying out an *Interactive IR* experiment with highlighting and note-taking tools built into the search interface. We conducted a crowdsourced between-subjects study (N = 115), where participants were assigned to one of four conditions: *(i)* **CONTROL** (a standard web search interface); *(ii)* **HIGH** (highlighting enabled); *(iii)* **NOTE** (note-taking enabled); and *(iv)* **HIGH+NOTE** (both highlighting and note-taking enabled). We assess participants' learning with a recall-oriented vocabulary learning task, and a cognitively more taxing essay writing task. We find that *(i)* active reading tools do not aid in the vocabulary learning task. However, *(ii)* participants in **HIGH** covered 34% more subtopics, and participants in **NOTE** covered 34% more facts in their essays when compared to **CONTROL**. Furthermore, *(iii)* we observed that incorporating active learning tools significantly changed the search behaviour of participants across a number of measures. This is the first work that sheds light on the effect of active reading tools on the SAL process, with important design implications for learning-oriented search systems.

This research has been supported by *DDS (Delft Data Science)* and *NWO* projects *SearchX* (639.022.722) and *Aspasia* (015.013.027).

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8055-3/21/03.
https://doi.org/10.1145/3406522.3446025

1 INTRODUCTION

The process of what is now known as *Search as Learning (SAL)* [12] was first formally defined by Marchionini [33] as an iterative process where learners purposefully engage with a search engine by reading, scanning and processing a large number of documents, *with the ultimate goal of gaining knowledge about one specific learning objective*. Learning via exploring, finding, analysing and evaluating documents [1, 35] containing information relevant to the desired learning objective is a time-consuming and cognitively demanding process. While learners often equate searching for information with searching the web [8, 37], web search engines are not equipped with tools to help users during the complex searches that are necessary in the context of learning [2, 22].

Outside of the web search scenario, *active reading strategies* such as annotating content (highlighting, note-taking, etc.) have been shown to have multiple benefits when engaging in long and complex learning tasks [40, 56]. These tools enable learners to limit their working memory load, as well as articulate and reformulate their thoughts. In turn, this can lead to substantial improvements in the understanding and retention of knowledge [25, 34]. Active reading strategies play a number of roles in the text comprehension process. *Highlighting* is used for text selection, and *note-taking* for organisation. Both have been shown to help with the learning process—especially in recall oriented tasks, like a *fill in the blanks test* [41, 56], or *multiple choice questions (MCQ)* [4, 50].

Despite the apparent benefits of active reading tools within a learning context, highlighting and note-taking tools are not found in contemporary web search engines. Efforts have however been made to develop *information organisational tools*. By providing a note-taking interface, they allowed users to keep track of their search context, collect relevant articles and improve sense-making during search [5, 15]. However, none of these works explicitly measured the effect these tools had on learning. *A/B* testing was not conducted either, meaning no comparison of benefits could be made against a control group.

This paper addresses the aforementioned research gap. More specifically, in this paper, we explore the impact that active reading tools—*integrated into the search interface*—have on learning-oriented search tasks, with respect to behavioural and learning outcomes. We implemented two active reading widgets—a highlighting tool and a note-taking tool—within an experimental search system. We conducted a between-subjects study (N = 115) where participants were assigned to one of four conditions, where the search interface contained (or lacked) the aforementioned tools: **CONTROL**, our control interface; **HIGH**, with text highlighting; **NOTE**, with note-taking; and **HIGH+NOTE**, including both tools. Participants

were assigned to one of two search topics, with their learning assessed over two tasks: a recall-oriented vocabulary learning *(receptive)* task [36, 44]; and a cognitively demanding essay writing *(critical)* task [31, 46]. As such, this user study aims to address the following two research questions.

RQ1 *To what extent do built-in highlighting and note-taking tools benefit users in learning oriented search tasks when compared to a conventional web search interface?*

RQ2 *How does the presence of active reading tools affect the search behaviour of users in learning oriented search tasks?*

Key findings. *(i)* The integration of active learning tools within the search interface does not aid in the receptive tasks. *(ii)* **HIGH** participants covered 34% more subtopics and **NOTE** participants covered 34% more facts in their essays compared to **CONTROL**. Providing both tools does not improve critical learning. *(iii)* The type of active learning tools has a significant impact on search behaviour. We found that participants with access to the tools queried less and viewed fewer documents. At the same time, **HIGH** and **HIGH+NOTE** participants spent more time reading documents; their **NOTE** counterparts spent considerable time writing notes.

2 RELATED WORK

The learning literature suggests that effective utilisation of active reading strategies (such as text highlighting, writing out keywords, note-taking and reflecting) helps to improve metacognitive monitoring of the learning process [16, 18, 38, 42, 50]. In turn, active reading strategies help to improve comprehension. Here, we outline prior works that have examined active learning strategies (pertaining specifically to text highlighting and note-taking), along with a wider discussion of recent, associated works in the SAL domain.

Text Highlighting. Important concepts, ideas and information within a passage of text are often explicitly marked (or *highlighted*) by a learner. This is one of the most common ways to self-regulate learning from text [23, 28, 56]. However, prior works have limitations. They typically examine text highlighting or other active learning tools on *printed text,* or a *single* digital document.

Leutner et al. [27] found that teaching learners to use active reading strategies like highlighting—together with lessons on self-regulation—was beneficial for learning. In contrast, Ponce and Mayer [41] found that providing highlighting functionality over a single document did improve the memorisation of highlighted terms, but *did not* lead to improved essay writing skills for their participants (when compared to a control condition, where no highlighting tool was present). Yue et al. [56] demonstrated that the highlighting of printed text improved the recall of keywords for a *fill in the blanks* task. The participants were able to answer more questions correctly from texts that they had highlighted when compared against texts without highlighting.

Ben-Yehudah and Eshet-Alkalai [4] compared text highlighting in both printed and on-screen text, and compared participants' learning here against a control setup (with no highlighting) for both mediums. They observed that highlighting helped in text comprehension (evaluated through a MCQ test), but only for printed text. The authors reasoned that under their setup, highlighting on the on-screen platform was not as *convenient* or *natural* when compared to highlighting on printed text. As a result, participants had to expend greater cognitive loads in the act of highlighting alone. The increase in cognitive load was therefore likely to harm the comprehension of the text. The authors also hypothesised that if highlighting for on-screen text were to become more convenient and natural for the learner, greater cognitive capabilities would be available for a deeper understanding and processing of the text. Liu et al. [30] observed that when used alone, text highlighting *may not be beneficial.* Externalising thoughts together with highlighting can however be effective. We draw on inspiration from these prior works, and examine the benefits (if any) that text highlighting provides to learners in a learning-oriented web search task.

Note-Taking. The externalisation of thoughts can be achieved through careful note-taking as learners read and comprehend information presented to them. Through qualitative interviews, Capra et al. [11] found that users in exploratory search tasks reported note-taking as one of the most common activities during the search session. However, the effects of note-taking on knowledge gain or learning was not explored. Liu et al. [30] observed that for video learning, users showed higher learning gains (compared to a control group) using their active reading tool over video transcripts—which offered text highlighting, note-taking and questioning functionality. Camporro and Marquardt [10] conducted a study to understand user preferences between paper and on-screen note-taking, where on-screen notes were written on a tablet device. A majority of participants were found to prefer on-screen note-taking, so long as it did not increase their cognitive load by distracting them from listening to presentations. In contrast to the works that have considered note-taking in the context of a *single document or video lecture*, we explore in this paper the benefits of note-taking within learning-oriented search tasks that spans *multiple webpages*.

Search as Learning. Previous research within the SAL domain has focused on: *(i)* understanding user behaviours when undertaking a learning-oriented search task [9, 17, 24, 29, 32, 36]; *(ii)* exploring different types of users and their behaviours (e.g., novices vs. experts) [6, 17, 39, 44]; and *(iii)* the optimisation of retrieval functions for learning [47–49].

Liu and Song [29] observed that learners who adapted their source selection strategies (e.g., reading encyclopedia documents from *Wikipedia* for *receptive learning tasks*, like vocabulary learning; or reading Q&A documents from platforms such as *Stack Overflow* for *critical learning tasks*, such as analysing an issue or solving a problem) showed better learning outcomes when compared to learners who did not adapt these strategies. Kalyani and Gadiraju [24] also explored the effects of cognitive complexities for learning tasks (such as *remembering* vs. *applying* knowledge) on search behaviours, and observed that more cognitively taxing tasks led to a higher number of interactions with the search interface.

Characteristics of users have also been shown to influence the amount of learning that takes place during a search session. Gadiraju et al. [21] observed that participants with little prior knowledge achieved higher learning gains than learners with at least some knowledge *a priori*. In contrast, O'Brien et al. [39] found no difference in learning outcomes (measured by essay quality) between domain experts and non-experts. Liu et al. [31] reported that participants in their study underwent knowledge changes during different stages of a search session. However, the changes did not depend

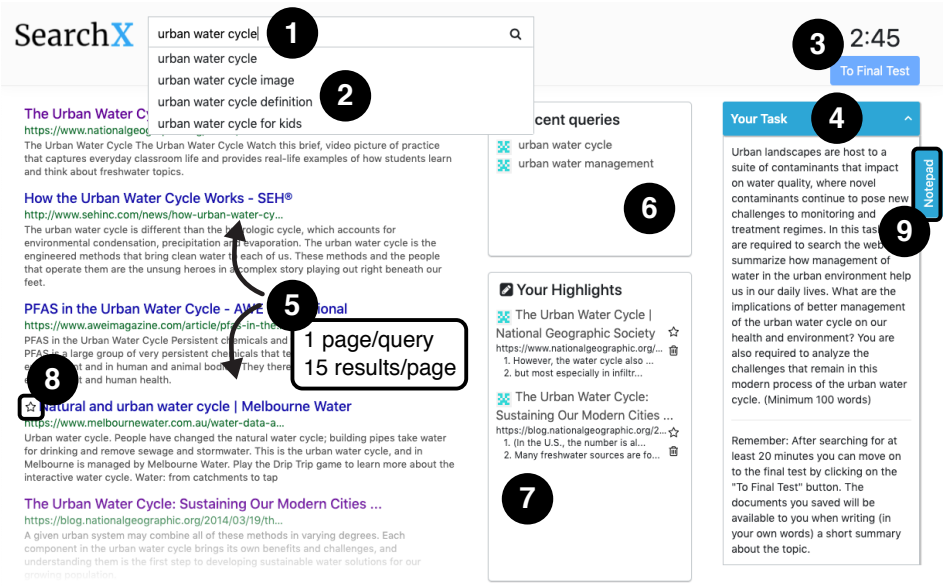

Figure 1: The SearchX interface as used for this study with annotations—refer to §3.1. This screenshot is an amalgamation of what would have been seen over all experimental conditions; refer to §4.1 for details.

on their prior knowledge about a topic. More recently, Roy et al. [44] examined *when* learning occurs during a search session. They observed a difference between participants with higher and lower level prior knowledge levels, with the former showing higher learning gains towards the end of the search session. In this paper, we are interested in observing the benefits of active reading tools over two different learning tasks—a *low-level, receptive* vocabulary learning task, and a *high-level, critical* essay writing task. Since search and user characteristics have been shown to affect a user's behaviours and learning outcomes [21, 24, 29], we explore how *the inclusion of active reading tools* affect search and learning behaviours during a learning-oriented search task.

3 HIGHLIGHTING AND NOTE-TAKING

To carry out our study, we used `SearchX` [43], a modular open-source retrieval framework that provides out-of-the-box support for crowdsourced interactive IR experiments. The standard interface provides a series of *widgets*, which, when taken together, comprise the look and feel of a contemporary web search engine's *Search Engine Results Page (SERP)*. Figure 1 shows the interface we used for our study. Figure 2 shows the two widgets we implemented: *text highlighting* and *note-taking*.[1]

3.1 SearchX Interface

The `SearchX` interface comprises of a SERP akin to a contemporary web search engine. There are additional *widgets* which are provided to aid users during the searching and learning process.

Starting with ❶, users are able to enter and submit queries using a standard query box. In addition, we also provide *Query Autocompletion (QAC)* ❷ to assist users in formulating their queries. To the top right of the interface ❸ is a clock indicating the elapsed time that has passed since the search task began, along with a button `To Final Test`. This button becomes enabled after 20 minutes of the search session and when clicked, moves the participant to the next stage in the experiment. The task description is provided at ❹ to allow participants to re-familiarise themselves with the task at hand. Results are presented underneath the query box ❺. Using the conventional and familiar *link/URL/snippet* layout, up to 15 results are presented. Clicking links (blue denoting unread; purple denoting previously examined) will open the document viewer widget, as shown in Figure 2a. In our experiment, pagination is *not* included as studies have shown that users often do not move to the second page of results or beyond [21, 36]. We also include a widget that lists queries that the participant previously issued during the session ❻. The widget lists queries in chronological order, with the most recent query placed at the top. In addition, we also provide a widget that lists previously made highlights ❼; it presents all documents that contain at least one highlight, as well as the corresponding highlights. Note that if highlighting is disabled for an experimental condition, this widget will simply list documents that participants decide to *save*. That is, participants will instead *save a list of documents* that are deemed to be useful to them in addressing the task. This is achieved by *starring* a document, as shown at ❽. This is in contrast to when highlighting is present, where participants will curate a *list of highlights that are created over each document examined* (here, starring a document is unavailable). Lastly, we provide note-taking functionality with the `Notepad` button ❾—see §3.4.

[1] All source code, tasks, and descriptions are available online at `https://github.com/roynirmal/searchx-front-highlighting` and `https://github.com/roynirmal/searchx-back-highlighting`.

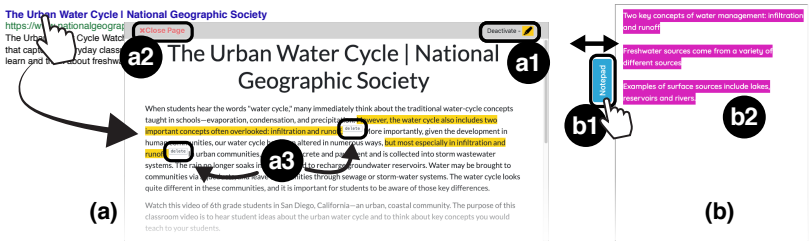

Figure 2: Examples of the two new widgets introduced to SearchX for this study. *(a)* On the left is the document view, complete with text highlighting capabilities. *(b)* On the right is the note-taking widget, which is visible when Notepad is clicked. Note that these features were not available to all participants of the study; refer to §4.1 for more information.

3.2 SearchX Logging

The SearchX system generates fine-grained search logs, allowing us to capture a number of key behavioural measures.[2] The system also provides a number of quality control features. As an example, participants who switched out of the search interface more than three times were automatically disqualified. This was to ensure that participants did not unduly become distracted or end up using alternative search engines to complete their task. It was also employed to ensure that participants would use our system, rather than simply running down the clock while being engaged with some other activity on screen.

3.3 Text Highlighting

Encapsulated within the *document widget*, as shown in Figure 2a, is the highlighting tool. When presented with a SERP, a participant identifies a document that they wish to examine in more detail. By clicking the link associated with the document, the document widget then appears *on top of the SERP*, with the title and document content shown within the popup that appears. Participants may then begin to highlight portions of text within the document; the highlighter is enabled by default. The participant clicks and drags over the text they wish to highlight, and let go of the mouse or trackpad they are using when they have selected what they wish to highlight. Highlights are automatically saved by the system and made available in the 7 `Your Highlights` widget.

Highlights can also be deleted; this is demonstrated by the small `delete` button that appears at the end of the highlight in question, as shown by a3 in Figure 2a. The highlighting feature can also be disabled by clicking the button at a1 in Figure 2. The document widget can be closed by clicking `Close Page` at a2, which will then return the participant to the SERP.

Note that the document widget was specifically created to aid participants in highlighting text within a document. By extracting the text from the markup of the page in question, and presenting it within a plain popup (with black text on white), the complex styling of contemporary web pages is avoided, making highlighting easier to achieve and more impactful to the user.

[2]Logs include the list of snippets shown on screen, any documents that were examined, dwell times, mouse hovers, etc.

3.4 Note-Taking

In addition to the text highlighting tool, we have also implemented a note-taking widget, stylised as `Notepad`. With experimental conditions that permit it, the note-taking widget is available initially as a non-intrusive 'tab'-style button, as shown in Figure 2b at b1. When the participant clicks on this button, the note-taking widget appears to the right of the viewport, *floating above* all other elements of the SERP. This means that the widget is visible in any state, regardless of whether the document widget is present or not.

Once open, a participant can write whatever notes they wish as they read through snippets and documents. Text can be copied and pasted from snippets and documents into the note-taking widget. It is important to note that the highlighting widget and note-taking widget are *not* linked together. It was decided not to do this to grant the participants freedom in how they took notes (if any), rather than to introduce restrictions into the note-taking process. All notes are automatically saved as they are typed, and are present for the entirety of the task (i.e., they do not pertain to a specific document).

4 USER STUDY DESIGN

In this section we describe our study design, outlining our setup and experimental conditions.

4.1 Experimental Conditions

To investigate how highlighting and note-taking functionality influences users during a learning orientated search task, we consider four experimental conditions:

`CONTROL` The standard `SearchX` search interface is provided *without* highlighting or note-taking capabilities. As outlined in §3.1, users are able to save documents and 7 becomes the `Saved Documents` widget.

`NOTE` In addition to the `Saved Documents` widget as for **`CONTROL`**, the note-taking widget is enabled.

`HIGH` In this condition the highlighting widget is enabled (i.e., 7 as shown in Figure 1).

`HIGH+NOTE` Both the highlighting and note-taking widgets are enabled.

4.2 Procedure

Our study flow is illustrated in Figure 3. It is inspired by recent studies in the SAL domain [21, 36, 44]. Independent of the experimental condition, participants were first asked to answer seven questions

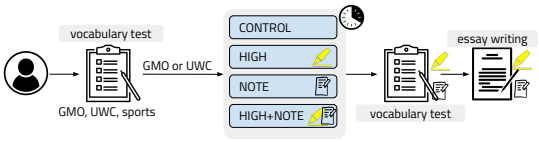

Figure 3: Overview of the study's workflow.

(designed to prime them towards learning-oriented searches, highlighting and note-taking). We asked them to reflect the last time they made such searches and their opinion regarding the benefits of active reading tools. Next, they completed two vocabulary knowledge tests, each one covering 10 vocabulary questions on a particular topic (we also include a third topic as a participant engagement check, outlined in §4.3). The topic they know the least about is then chosen as the topic to learn more about during the search session. We randomly assign each participant to one of our four experimental conditions. Participants have to stay in the search phase for at least twenty minutes (hence the timer at ❸ in Figure 1), as this has been shown to be a reasonable time for people to accrue knowledge [17, 24, 36]. After the minimum search time has passed, participants can continue to the post-test, which consists of a vocabulary test on their topic (we ask *the same* 10 questions as in the pre-test) as well as an essay writing assignment (with a minimum required length of 100 words). During the post-test phase, the participants can access their saved documents, highlights and notes (though editing is now prohibited). Gauging participants' learning across receptive and critical learning tasks (§4.3) provides us with a more comprehensive understanding of how much a participant has *learned* about a particular topic than only a recall-oriented vocabulary learning task as conducted in [36, 44]. Lastly, we asked our participants seven reflective questions to gauge the perceived usefulness and ease of using our tools. These questions were restricted to participants that received one or both of those widgets.

4.3 Topics

In line with prior studies [31, 46], we construct two learning-oriented tasks in our experiment: one *receptive learning task* and one *critical learning task*. Receptive learning is defined as understanding, remembering and reproducing what is taught [26]. Concretely, we ask participants to provide definitions (if they can) of ten vocabulary terms relevant to the topic at hand. In contrast, critical learning includes criticising and evaluating ideas from multiple perspectives [26]. In our study, we ask our participants to analyse challenges and provide their own view of the topic. Overall, the two learning tasks encompass the lower level and higher level cognitive process dimensions of Anderson and Krathwohl's taxonomy [52].

The two topics we used for this study along with the ten vocabulary terms that participants were asked to define in the pre- and post-test, and sub-topics corresponding to the topics were chosen from [14]. They are presented below.[3] We also included the topic of *sports*, as an engagement check for all participants in the pre-test: if participants exhibited the same or less prior knowledge on *sports* compared to the other two topics, they were rejected. This is in line with [36, 44] as we expect participants to have reasonably high knowledge regarding the vocabulary terms for the sports topic.

[3]Note that these subtopics were used in our manual evaluation of the users' essay content, but were not explicitly conveyed to them.

Urban Water Cycle (UWC): Urban landscapes are host to a suite of contaminants that impact water quality, where novel contaminants continue to pose new challenges to monitoring and treatment regimes. In this task, you are required to search the web and summarise how management of water in the urban environment can help us in our daily lives. What are the implications of better management of the urban water cycle on our health and environment? You are also required to analyse the challenges that remain in this modern process of the urban water cycle. (Minimum 100 words).
Vocabulary terms: Lesoto Highlands, eutrophication, Endocrine disrupting compounds, typhoid fever, coagulation, activated carbon filtration, membrane filtration, cholera, Legionella bacteria, recontamination.
3 subtopics: benefits of water management (WM) on health, benefits of WM on environment, remaining challenges

Genetically Modified Organisms (GMO): Genetically Modified Organisms (GMOs) have become controversial as their benefits for both food producers and consumers are accompanied by potential biomedical risks and environmental side effects. Imagine in a 'Biotechnology' course, you chose the topic of GMOs. You intend to introduce the benefits of GMOs on modern society to your class. At the same time, you analyse why GMOS can become a potential risk on health, the economy, and society at large – and finally give your conclusion on whether we should progress our research on GMOs and their commercial use. In order to complete this presentation, you need to search for relevant information and prepare an essay for yourself. (Minimum 100 words).
Vocabulary terms: transgenic, genomes, selective breeding, microinjection enzyme, chromosome, plasmid, myxoma, kanamycin, severe combined immunodeficiency, Leber's congenital amaurosis.
5 subtopics: benefits of GMOs, risk on health, risk on environment, risk on economy, own conclusion

Sports: Imagine you are taking an introductory course on Sports. For your term paper, you have decided to write about *Sports Development and Coaching*.
Vocabulary terms: olympics, weight lifting, karate, martial arts, aerobics, athletes, soccer, baseball, snowboarding, hockey.

4.4 SearchX Setup

Search Results and QAC Suggestions. The search session is facilitated by the *Bing Search API*. The Bing Search API was used not only for the retrieval of search results (up to 15 per query), but also for QAC suggestions. QAC suggestions were retrieved on a per-keystroke basis, after at least three characters were present within the query box. Snippets were used in the search interface as-is from Bing.

Document Prefetching. As shown in Figure 2a, the document widget presents web pages in a heavily altered format. Page-specific styling is removed to yield the content in black text on a white background, complete with images associated with the page (but excluding images within stylesheets, for example). This is done to: *(i)* make it easier for participants to highlight text (without complex page layouts); *(ii)* observe their highlights; and *(iii)* reduce the likelihood of distractions. As such, presenting the document in a timely manner presented a major technical challenge.

To parse content before before being viewed, we *prefetched* the documents in the results list returned by the Bing Search API for each query issued. Web pages were accessed and crawled, and stored in a cache. As some pages may have been unavailable (through server downtime, for example), pre-warming the cache with results from previously issued queries was undertaken to minimise the risk of prolonged (5 seconds or more) delays in returning results to participants. Queries for the same topics were selected from the study by [14], with the top 50 results saved to the cache. By completing this step, 60,000 documents were prefetched.

Removal of Wikipedia(-like) Pages. One identified risk was the inclusion of *Wikipedia* and *Wiki*-style pages that comprehensively

Table 1: The number of participants exploring each topic in our study, together with related statistics. Two-way ANOVA tests revealed no significant differences in average number of queries between topics ($F(1, 107) = 1.83$, $p = 0.07$). ± indicates the standard deviation.

	GMO	UWC
Overall	**71**	**44**
⇒ `CONTROL`	21	11
⇒ `HIGH`	19	10
⇒ `NOTE`	17	12
⇒ `HIGH+NOTE`	14	11
Average number of queries	4.61(±2.97)	5.51(±2.38)
Median number of queries	4	5

would outline the topics given. By reading a single page, a participant could then find acceptable answers for all posed questions; this would render the need to search and examine additional pages redundant. As such, a large number of Wikipedia articles (and documents from known Wikipedia clones) were removed from the search results, such as the Wikipedia article on GMOs[4]. We used a curated list of known Wikipedia clones[5], and excluded these domains from the presented results. In all, 72 Wikipedia clone domains were excluded from the presented results.

4.5 Participants

Since insights from crowdsourced experiments are comparable to lab-based ones [20, 57], we recruited participants for our study using the crowdsourcing platform *Prolific*[6]. The platform has been shown to be an effective choice for relatively complex and time-consuming interactive information retrieval experiments [54]. The study was undertaken over a two day period in the autumn of 2020. To ensure reliable and high-quality responses, we required our participants to have: *(i)* successfully completed 100 prior submissions on the Prolific platform; *(ii)* possess an approval rate of 90% or higher; and *(iii)* have native proficiency in English. Including the minimum search time of 20 minutes, the complete study took approximately forty five minutes to complete. For their time, participants were compensated at the rate of GBP£7.50 per hour.

We computed the required sample size in a power analysis for a *Between-Subjects ANOVA* using the software *G*Power* [19], resulting in the sample size of 120 participants. In all, 131 participants completed our study; 16 submissions were rejected based on our quality control criteria.[7] This led to the headline figure of $N = 115$. Of the valid participants, 64 identified as male, and 48 identified as female—with 3 withholding their gender identity. In terms of age, participants reported a median age of 33 (youngest 18; oldest 72). A total of 37 participants reported the highest formal education level as a *high school degree/diploma*. 48 reported a *Bachelor's degree*, with 11 possessing a *Master's degree*. The remaining 19 participants reported other education levels.

[4]https://en.wikipedia.org/wiki/Genetically_modified_organism
[5]This curated list is provided by Câmara et al. [14].
[6]https://www.prolific.co/
[7]Quality control criteria included counting browser blurring events (discussed in §3.2); participants should issue at least two queries, view two documents, and finish the post-test with a reasonable essay (as deemed through a manual evaluation).

Table 1 reports the number of participants per topic, over each of the four conditions trialled. Of the 115 participants, 71 were assigned to the **GMO** topic, with the remaining 44 to the **UWC** topic. Remember that topics were assigned to participants based on their pre-task surveys (participants received the topic they had the least knowledge about), leading to a skewing towards the **GMO** topic. The table also contains basic statistics on the number of queries issued which is comparable to that reported in previous studies [24, 36] and shows that participants were fairly active on our platform; refer to §5 for more information.

4.6 Measuring Learning

Realised Potential Learning (RPL). Our vocabulary learning task is evaluated via the *Vocabulary Knowledge Scale* [51] which the participants use to rate their knowledge in line with prior work [13, 36, 45, 47, 48].

(1) *I don't remember having seen this term/phrase before.*
(2) *I have seen this term/phrase before, but I don't think I know what it means.*
(3) *I have seen this term/phrase before and I think it means ...*
(4) *I know this term/phrase. It means ...*

Importantly, the self-assessment of *(3)* or *(4)* requires participants to write down a definition of the vocabulary term in their own words. Having collected the participants' knowledge ratings, we compute RPL $\in [0, 1]$ for each participant, which denotes what fraction of knowledge (amongst all knowledge) they could have gained (i.e., rating all terms with *(4)*) with respect to what they actually gained. We follow earlier works and assign a score $s^X(t_i)$ (where X is either *pre* or *post*) of 0 to knowledge levels (1) and (2) for term t_i, a score of 1 to knowledge level 3 and a score of 2 to knowledge level 4. We first compute the *Absolute Learning Gain (ALG)* across all n vocabulary terms as follows:

$$ALG = \frac{1}{n} \sum_{i=1}^{n} max(0, s^{post}(t_i) - s^{pre}(t_i)). \quad (1)$$

Note the $max()$ function ensures that knowledge of a vocabulary term cannot drop. Given the short time-frame (20 minutes) of the search session, this is a realistic assumption. RPL then normalises ALG by the maximum possible learning potential:

$$RPL = \frac{ALG}{\frac{1}{n}\sum_{i=1}^{n} 2 - s^{pre}(t_i)}. \quad (2)$$

T-Depth, F-Fact and Readability. For the critical learning task, we determine participants' knowledge expressed in their essays by following the work of Wilson and Wilson [53], who proposed and compared a number of measures for this very task. Concretely, we employ *F-Fact*, which counts the number of individual facts present in a summary, and *T-Depth*, which rates to what extent each subtopic is covered in the summary on a scale of $0 - 3$ (from *not covered at all* to *covered with great focus*), as both of these measures were shown to be good indicators of learning. Both of these measures require a manual annotation effort. A concrete example of how we annotated facts and subtopic coverage in participants' summaries is provided in Table 2. Three annotators split the 115 essays among them. There were 18 essays which were analysed by

Table 2: Example annotation of facts and subtopics and the computation of F-Fact and T-Depth. Note that sentences demonstrating knowledge of the topic are colour coded—each colour pertains to an individual subtopic (see the *T-Depth* column).

Essay	F-Fact	T-Depth
GMOs, or GE (genetic engineering) technology provides a **number of potential benefits to farmers.**	1	Benefits of GMO = 3
GE crops are bred to **answer some of the pest**, **disease**, and **weed challenges producers**, by **adding resistance** or **other traits to the crops** .	5	
For instance, **some crops have been modified for resistance to particular diseases** or **pest pressure**, while **others are herbicide resistant** .	3	Risk on health = 1
The argument is essentially that **GE crops allow for more efficient use of land**, with **greater yields** on **less acres** (and **with higher profit margins**) .	4	
There has been some **controversy from consumers over the safety of eating GE crops**, and **whether they can increase levels of food allergies** or **affect human health** .	3	Risk on environment = 2
There is also **concern about the modified genes mixing with gene pools in the wild**, **potentially contaminating other non-GE seeds** or **animals** .	3	
I'm not entirely opposed to GE technology, but I think that it's a crude tool that largely benefits big agribusiness at the cost of farmers and consumers.	0	Risks on economy = 1
Additionally, **GE creates the potential for insects** and **weeds to develop resistance to current effective controls** , **which creates a sort of arms race of GE tech to stay ahead of the resistance** .	4	
(I could go on for literal hours here... but it wouldn't be based on the research I was doing)	0	Conclusions = 1
Metric Score	**23**	$(3+1+2+1+1)/5 = 1.6$

all annotators; observing a Pearson correlation of 0.78 ($p = 0.002$) for *T-Depth* sores, and a correlation of 0.76 ($p = 0.002$) for *F-Fact* scores which indicates high inter-annotator agreement. Lastly, as neither of those measures is concerned with the readability of participants' essays, we also computed the Flesch-Kincaid[8] readability scores. A high score indicates that the text is fairly easy to read, whereas a lower score indicates that the text is fairly complex and can be best understood by university graduates.

5 RESULTS

First, we need to assess the reliability of the participants self assessment regarding their vocabulary knowledge. We randomly sampled 50 answers for knowledge levels *(3)* and *(4)*; labelling them as *correct*, *partially correct*[9], or *incorrect*[10]. We found that for knowledge level *(3)*, 38% of the answers were correct, 56% were partially correct and remaining 6% were incorrect. Out of the answers self-assessed as (4) 72% were correct, 26% were partially correct and remaining 2% were incorrect. Based on these numbers we argue that the self-assessments of the participants are largely reliable. On average participants marked 2.2(±1.8) answers as knowledge levels (3) or (4). This indicates that the participants still needed to learn fair bit of the topics for our tasks.

We now turn our attention to presenting the results of our study in line with our research questions. Measures were analysed considering both the conditions and the topics used; two-way ANOVAs were conducted using these as factors; main effects were examined with $\alpha = 0.05$. TukeyHSD pairwise tests were used for post-hoc analysis. Note that ± values in the tables and corresponding narrative both indicate the **standard deviation**.

5.1 Highlighting, Note-Taking and Learning

Our first research question, **RQ1**, considers *how beneficial the highlighting and note-taking widgets are for learning-oriented search tasks, when compared to a standard web search interface.* Table 3 presents an overview of our learning measures (amongst other behavioural measures) across our four experimental conditions. We report the RPL (**III**), T-Depth essay scores (**IV**), F-Fact essay scores (**V**), and Flesch essay scores (**V**). We first examine the effects of highlighting and note-taking on vocabulary learning.

Our analysis shows that mean RPL scores varied between 0.11 (**CONTROL**) and 0.15 (**NOTE**), all with similar levels of variance. Indeed, our ANOVA analysis yielded no significant differences between the four conditions. The reported mean RPL figures showed that participants gained less than 20% of the knowledge that *could* have been acquired when considering the results of their receptive learning surveys. This finding shows that although highlighting tools have been shown to improve receptive knowledge while learning from a single document [4, 41, 50, 56], they do not aid receptive learning to a similar extent in complex search sessions. Further analysis showed a very small fraction of vocabulary terms that were present in the recorded text highlights (**XVII**) or notes (**XXI**).

T-Depth, F-Fact and Flesch essay scores, that pertain to evaluating critical learning ability, are presented on rows **IV**, **V** and **VI** respectively in Table 3. Looking first at the T-Depth essay scores, we see a general trend showing that for conditions where additional tools were available (**HIGH** at 1.64 ± 0.59, **NOTE** at 1.40 ± 0.61, and **HIGH+NOTE** at 1.48 ± 0.67), more subtopics were covered by participants in sufficient detail than when compared to those assigned to **CONTROL** at 1.22 ± 0.43. Post-hoc analysis yielded a significant

[8] We use `textstat` for computing the Flesch readability score.
[9] An example of partially correct answer from **UWC** topic: *an illness* for the vocabulary term *typhoid fever.*
[10] An example of incorrect answer from **GMO** topic: *Relating to plasma* for the vocabulary term *plasmid.*

Table 3: Mean (± standard deviations) of RPL and search behaviour metrics across all participants in each condition. A dagger (†) denotes two-way ANOVA significance, while ^{C},$^{\mathcal{H}}$,$^{\mathcal{N}}$,$^{\mathcal{B}}$ indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) over the four conditions CONTROL, HIGH, NOTE and HIGH+NOTE respectively.

Measure			CONTROL	HIGH	NOTE	HIGH+NOTE
	I.	Number of participants	32	29	29	25
	II.	Search session duration (minutes)	$23m40s(\pm10m26s)$	$27m53s(\pm9m50s)$	$20m3s(\pm6m46s)$	$29m17s(\pm15m15s)$
Essay Scores	III.	RPL	$0.11(\pm0.10)$	$0.14(\pm0.21)$	$0.15(\pm0.15)$	$0.11(\pm0.10)$
	IV.	*T-Depth* scores of essays†	$1.22(\pm0.43)^{\mathcal{H}}$	$1.64(\pm0.59)^{C}$	$1.40(\pm0.61)$	$1.48(\pm0.67)$
	V.	*F-Fact* scores of essays†	$14.56(\pm10.36)^{\mathcal{N}}$	$16.55(\pm5.51)$	$19.59(\pm8.53)^{C}$	$15.92(\pm8.06)$
	VI.	Flesch scores of essays†	$32.19(\pm39.78)$	$21.40(\pm62.71)$	$15.86(\pm61.54)^{\mathcal{B}}$	$46.43(\pm16.75)^{\mathcal{N}}$
	VII.	Mean #. of essay terms	$181.56(\pm76.35)$	$200.83(\pm85.61)$	$225.86(\pm112.53)$	$193.00(\pm87.94)$
	VIII.	Number of queries†	$5.81(\pm3.54)^{\mathcal{H},\mathcal{B}}$	$4.63(\pm2.68)^{C}$	$4.93(\pm2.53)$	$4.28(\pm1.74)^{C}$
Times	IX.	Average time between queries (seconds)	$281.95(\pm271.50)$	$307.22(\pm265.15)$	$254.20(\pm148.67)$	$289.49(\pm206.90)$
	X.	Average time between documents (secs.)	$96.60(\pm49.19)$	$68.63(\pm174.82)$	$121.30(\pm65.72)$	$114.25(\pm174.20)$
	XI.	Average document dwell time (secs.)†	$339.50(\pm34.74)^{\mathcal{H},\mathcal{N},\mathcal{B}}$	$522.13(\pm38.86)^{C,\mathcal{N}}$	$\$166.03\ (\pm26.16)^{C,\mathcal{H},\mathcal{B}}$	$470.64(\pm50.80)^{C,\mathcal{N}}$
Docs.	XII.	Number of unique documents viewed†	$12.16(\pm5.95)^{\mathcal{H},\mathcal{N},\mathcal{B}}$	$8.07(\pm3.95)^{C}$	$8.24(\pm3.57)^{C}$	$8.60(\pm3.50)^{C}$
	XIII.	Number of unique document snippets viewed†	$99.72(\pm57.34)^{\mathcal{H},\mathcal{B}}$	$72.93(\pm35.19)^{C}$	$87.03(\pm38.37)$	$71.00(\pm25.25)^{C}$
Highlighting	XIV.	Number of highlight additions	–	$53.53(\pm38.64)$	–	$50.56(\pm43.99)$
	XV.	Number of highlight deletions	–	$7.20(\pm17.56)$	–	$4.36(\pm11.25)$
	XVI.	Number of words highlighted per highlight action	–	$29.48(\pm17.37)$	–	$30.30(\pm7.87)$
	XVII.	Fraction of vocabulary terms present in highlights	–	$0.04(\pm0.06)$	–	$0.06(\pm0.08)$
	XVIII.	Fraction of essay terms present in highlights	–	$0.38(\pm0.16)$	–	$0.44(\pm0.21)$
Notes	XIX.	Percentage of user who took notes	–	–	86.2%	52%
	XX.	Number of words in note-pad	–	–	$1014.17(\pm2475.23)$	$379.04(\pm907.39)$
	XXI.	Fraction of vocabulary terms present in notes	–	–	$0.04(\pm0.09)$	$0.00(\pm0.02)$
	XXII.	Fraction of essay terms present in notes†	–	–	$0.37(\pm0.25)^{\mathcal{B}}$	$0.20(\pm0.28)^{\mathcal{N}}$
Perception	XXIII.	Ease of highlighting tool *(1 (easy) - 5 (difficult))*	–	$1.48(\pm0.91)$	–	$1.88(\pm1.17)$
	XXIV.	Usefulness of highlighting tool *(1 (not useful) - 5 (useful))*	–	$3.93(\pm1.33)$	–	$3.80(\pm1.38)$
	XXV.	Ease of notepad tool *(1 (easy) - 5 (difficult))*	–	–	$2.07(\pm1.28)$	$2.24(\pm1.39)$
	XXVI.	Usefulness of notepad tool *(1 (not useful) - 5 (useful))*	–	–	$3.76(\pm1.27)$	$3.08(\pm1.55)$

difference between the **CONTROL** and **HIGH** conditions ($F(3, 107) = 2.72, p = 0.04$). Significant differences were also found between conditions **CONTROL** (14.56 ± 10.36) and **NOTE** (19.59 ± 8.53) when looking at the F-Fact scores ($F(3, 107) = 2.68, p = 0.04$). We observed higher mean F-Fact scores corresponding to **HIGH**, **NOTE**, and **HIGH+NOTE** when compared to **CONTROL** (although this difference was not statistically significant for **HIGH** and **HIGH+NOTE**). This suggests that participants in other conditions discussed a greater number of facts in their essays when compared to their counterparts in **CONTROL**.

Turning our attention to the readability of the participant's essays, we observe that the Flesch readability scores (row **VI**, Table 2) also offer significant differences between conditions. Essays written by participants subject to **CONTROL** on average were easier to read than **HIGH** and **NOTE**. Additionally, essays written with the **NOTE** condition were significantly more complex to read than those written by **HIGH+NOTE** ($F(3, 107) = 2.64, p = 0.04$). We should also note that we observed negative Flesch scores for essays of 14 participants across all conditions. This typically happens when participants do not write complete sentences (e.g. bullet points) which renders the pieces of text to be more difficult to read.

From the above, we can see that highlighting and note-taking functionality aid different aspects of essay writing, with the former helping with subtopic coverage, and the latter with fact coverage. However, using both in tandem (**HIGH+NOTE**) does not lead to any significant learning outcome improvements compared to **CONTROL**. Our results contradict those found by Ponce and Mayer [41], who did not observe any significant differences in essay quality amongst participants with and without highlighting capabilities on the systems they used. However, it is important to note here that in the aforementioned study the participants had access only to a *single document*, and essays were evaluated using different measures (a presence of nine pre-defined items in the essays).

5.2 Highlighting, Note-Taking and Search Behaviour of Users

RQ2 considered *how active reading tools altered the search behaviour of participants.* For this question, we observe that the participants having access to one or both active reading tools issued fewer queries than those in **CONTROL**, and significantly so for **HIGH** and **HIGH+NOTE**. Previous studies [21, 36, 44, 55] have shown that participants issuing more queries observe higher knowledge gains in the receptive vocabulary learning task. The observations in our study might explain the lack of significant difference in RPL scores. However, despite issuing fewer queries (4.63 ± 2.68 vs. 5.81 ± 3.54, row **VIII**, Table 3), **HIGH** participants cover significantly more subtopics in their essays than their **CONTROL** counterparts ($F(3, 107) = 2.68, p = 0.04$). Looking deeper, we observe that participants in **HIGH** spend significantly more time reading documents than those in **CONTROL** (row **XI**) ($F(3, 107) = 5.63, p = 0.001$). This suggests that the highlighting tool facilitates user reflection more while reading a particular document, thereby internalising concepts more effectively than participants in **CONTROL**. The higher document dwell times for **HIGH** and **HIGH+NOTE** participants are in line with findings by Ben-Yehudah and Eshet-Alkalai [4], where

highlighting was shown to increase reading time of documents. Comparing the highlighting behaviour of the two groups in Figure 4, we observe a similar trend. Most of the highlighting activities are performed at the beginning of the search session. Later, highlights decrease to below 5 on average. This is coupled with the fact that fewer participants are involved in highlighting activity.

Document dwell time is however significantly lower for participants in **NOTE** (166.03 ± 26.16 secs.) when compared to all other conditions (e.g., 522.13 ± 38.86 secs. for **HIGH**). Although not significant, participants on average in **NOTE** spent more time on the SERP between reading two consecutive documents. This together with the lower number of snippets viewed indicates that participants in **NOTE** take notes after reading a particular document. We also observe that a significantly large portion of essay terms come from the notes of **NOTE** participants compared with **HIGH+NOTE** participants. This can explain the significantly more complex essays (indicated by the Flesch scores) written by **NOTE** participants when compared to that of their **HIGH+NOTE** counterparts. **NOTE** participants also wrote the longest essays on average—albeit not significantly so. Moreover, when compared to **HIGH+NOTE**, **NOTE** participants took more notes (row **XX**, there was no significant difference due to the high variance). From Figure 4, we see that **NOTE** participants take more notes towards the beginning—with **HIGH+NOTE** towards the end. For the latter, this also coincides with the time period where they are highlighting less. Thus, spending time on taking notes can be a contributing factor for participants acquiring more knowledge—and consequently using this in their essays.

We also observe that only 52% of the **HIGH+NOTE** participants engage in note-taking activities, compared to 87% of **NOTE** participants. This might indicate that participants in general prefer highlighting over note-taking given a choice. Our findings collectively suggest that providing both active reading tools might not be optimal for all users. Considering rows **XXIII** - **XXVI**, we see that on average (albeit not significantly), the highlighting and note-taking tools were considered more useful and easy to use in the standalone interfaces compared to **HIGH+NOTE**. Individually, the highlighting tool was perceived to be easier and more useful than the note-taking tool.

6 CONCLUSIONS AND FUTURE WORK

In this work, we have explored the effect of providing two active reading tools (highlighting and note-taking) with the goal of benefiting learners in learning-oriented web search tasks. To this end, we conducted a between-subjects user study, where $N = 115$.

We observed that neither the highlighting nor the note-taking tool helped participants in the receptive vocabulary learning tasks. However, participants having access to the highlighting tool only (**HIGH**) covered significantly more subtopics (34%) in their critical task essays compared to the control group (**CONTROL**). On the other hand, those with access only to the note-taking tool (**NOTE**) covered significantly more facts (34%) in their essays than the control group. Having access to both tools (**HIGH+NOTE**) did not lead to any significant learning gains in either the receptive or the critical tasks. Perhaps this is because both tools together add to the cognitive demand of the participants, which is evident from the fact that 52% of participants in condition **HIGH+NOTE** did not use the note-taking tool. This study therefore adds to a body of literature indicating that if we want people to perform better, *we need to find ways to reduce the cognitive load in search interfaces.* Our study also shows that having access to active reading tools significantly changes user behaviour when considering measures, such as the number of queries issued, the document dwell time, and the number of documents viewed. More specifically, we observe that having access to the highlighting tool leads to participants submitting fewer queries, and spending more time examining documents.[11] On the other hand, note-taking leads to participants spending *less* time reading documents, and taking more notes.

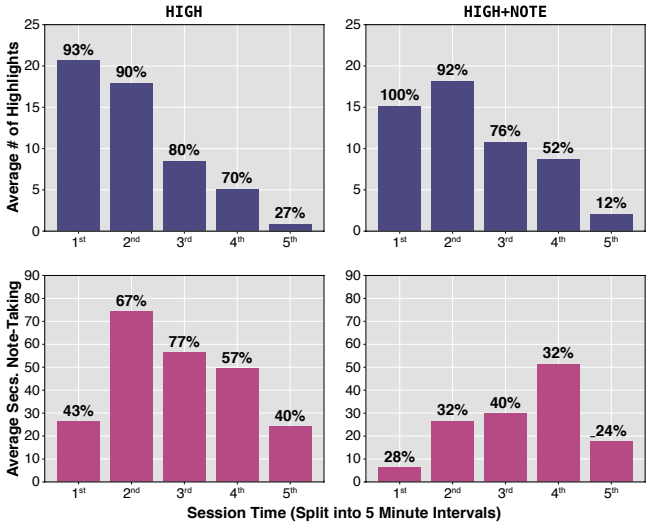

Figure 4: Average #highlights and average #seconds of note-taking activity in each five minute interval. The number on top of each bar shows the % of participants with 1+ highlights or 1+ seconds note-taking activity in that interval.

The findings from this study open up a number of possible future research directions. Extensions to this study could expand the work on examining how search behaviours can act as proxies for predictive measures of learning during search [6, 7, 17, 31, 44, 53] and to what extent user characteristics like their pre-knowledge or education level influence their highlighting or note-taking strategies and consequently their learning outcomes. In addition, advances in this area could lead to the development of an adaptive search system. More pertinent to this study however is identifying what highlighting and note-taking strategies exist between our participant cohort—and how these strategies affect learning outcomes. For example, Yue et al. [56] observed differences in efficiency of highlighting keywords between *heavy* and *light* highlighters on printed text. Would similar observations hold in a SAL context? Finally, further analysis of behavioural log data could provide insights into the document understanding process. For example, would recorded highlights and notes indicate more relevant/interesting sections of a given document, and if so, could retrieval algorithms be manipulated to promote documents that contain these *'hotspots'*? Findings could also eventually lead to the comparison of manual and automatic tools for active reading, and automatic thought externalisation.

[11] These findings are reflected by *Search Economic Theory (SET)* [3] that indicates with similar time limits, as the number of queries issued drops, more documents will be examined (or longer will be spent on them).

REFERENCES

[1] L.W. Anderson, B.S. Bloom, et al. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman,, 2001.

[2] A.H. Awadallah, R.W. White, P. Pantel, S.T. Dumais, and Y.-M. Wang. Supporting complex search tasks. In *Proc. 23rd ACM CIKM*, pages 829–838, 2014.

[3] L. Azzopardi. Modelling interaction with economic models of search. In *Proc. 37th ACM SIGIR*, SIGIR '14, page 3–12, 2014.

[4] G. Ben-Yehudah and Y. Eshet-Alkalai. The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *J. Edu. Multimedia & Hypermedia*, 27(2):153–178, 2018.

[5] K. Bharat. Searchpad: Explicit capture of search context to support web search. *Computer Networks*, 33(1-6):493–501, 2000.

[6] N. Bhattacharya and J. Gwizdka. Relating eye-tracking measures with changes in knowledge on search tasks. In *Proc. 10th ACM ETRA*, pages 1–5, 2018.

[7] N. Bhattacharya and J. Gwizdka. Measuring learning during search: differences in interactions, eye-gaze, and semantic similarity to expert knowledge. In *Proc. 4th ACM CHIIR*, pages 63–71, 2019.

[8] J.P. Biddix, C.J. Chung, and H.W. Park. Convenience or credibility? a study of college student online research behaviors. *Internet & Higher Education*, 14(3): 175–182, 2011.

[9] M. Bron, J. Van Gorp, F. Nack, L.B. Baltussen, and M. de Rijke. Aggregated search interface preferences in multi-session search tasks. In *Proc. 36th ACM SIGIR*, pages 123–132, 2013.

[10] M.F. Camporro and N. Marquardt. Live sketchnoting across platforms: Exploring the potential and limitations of analogue and digital tools. In *Proc. 38th ACM CHI*, pages 1–12, 2020.

[11] R. Capra, G. Marchionini, J. Velasco-Martin, and K. Muller. Tools-at-hand and learning in multi-session, collaborative search. In *Proc. 28th ACM CHI*, pages 951–960, 2010.

[12] K. Collins-Thompson, P. Hansen, and C. Hauff. Search as learning (dagstuhl seminar 17092). In *Dagstuhl reports*, volume 7, 2017.

[13] H.G. Colt, M. Davoudi, S. Murgu, and N.Z. Rohani. Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical endoscopy*, 25(1): 207–216, 2011.

[14] A. Câmara, N. Roy, D. Maxwell, and C. Hauff. Searching to learn with instructional scaffolding. In *Proc. 6th ACM CHIIR*, 2021.

[15] D. Donato, F. Bonchi, T. Chi, and Y. Maarek. Do you want to take notes? identifying research missions in yahoo! search pad. In *Proc. 19th WWW*, pages 321–330, 2010.

[16] J. Dunlosky, K.A. Rawson, E.J. Marsh, M.J. Nathan, and D.T. Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Sci. in the Public Interest*, 14(1):4–58, 2013.

[17] C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proc. 7th ACM WSDM*, pages 223–232, 2014.

[18] W. Fass and G.M. Schumacher. Effects of motivation, subject activity, and readability on the retention of prose materials. *J. Educational Psychology*, 70(5):803, 1978.

[19] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175—191, May 2007.

[20] U. Gadiraju, S. Möller, M. Nöllenburg, D. Saupe, S. Egger-Lampl, D. Archambault, and B. Fisher. Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, pages 6–26. Springer, 2017.

[21] U. Gadiraju, R. Yu, S. Dietze, and P. Holtz. Analyzing knowledge gain of users in informational search sessions on the web. In *Proc. 3rd ACM CHIIR*, pages 2–11, 2018.

[22] G. Golovchinsky, A. Diriye, and T. Dunnigan. The future is in the past: designing for exploratory search. In *Proc. 4th IIiX*, pages 52–61, 2012.

[23] D.D. Johnson and B. von Hoff Johnson. Highlighting vocabulary in inferential comprehension instruction. *J. Reading*, 29(7):622–625, 1986.

[24] R. Kalyani and U. Gadiraju. Understanding user search behavior across varying cognitive levels. In *Proc. 30th ACM HT*, pages 123–132, 2019.

[25] D. Kirsh. Thinking with external representations. *AI & society*, 25(4):441–454, 2010.

[26] H. Lee, J. Lee, K. Makara, B. J Fishman, and Y. Hong. Does higher education foster critical and creative learners? an exploration of two universities in south korea and the usa. *Higher Education Research & Development*, 34(1):131–146, 2015.

[27] D. Leutner, C. Leopold, and V. den Elzen-Rump. Self-regulated learning with a text-highlighting strategy. *J. Psychology*, 215(3):174–182, 2007.

[28] D. Leutner, C. Leopold, and E. Sumfleth. Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior*, 25(2):284–289, 2009.

[29] C. Liu and X. Song. How do information source selection strategies influence users' learning outcomes'. In *Proc. 3rd ACM CHIIR*, pages 257–260, 2018.

[30] C. Liu, C.-L. Yang, J.J. Williams, and H.-C. Wang. Notestruct: Scaffolding note-taking while learning from online videos. In *Proc. 37th ACM CHI*, pages 1–6, 2019.

[31] H. Liu, C. Liu, and N.J. Belkin. Investigation of users' knowledge change process in learning-related search tasks. *Proc. ASIS&T*, 56(1):166–175, 2019.

[32] J. Liu, N.J. Belkin, X. Zhang, and X. Yuan. Examining users' knowledge change in the task completion process. *IP&M*, 49(5):1058–1074, 2013.

[33] G. Marchionini. Exploratory search: from finding to understanding. *Comm. ACM*, 49(4):41–46, 2006.

[34] C.C. Marshall. Annotation: from paper books to the digital library. In *Proc. 2nd ACM Conf. Digital Libraries*, pages 131–140, 1997.

[35] R.E. Mayer, E. Griffith, I.TN. Jurkowitz, and D. Rothman. Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. *J. Exp. Psychology: Applied*, 14(4):329, 2008.

[36] F. Moraes, S.R. Putra, and C. Hauff. Contrasting search as a learning activity with instructor-designed learning. In *Proc. 27th ACM CIKM*, pages 167–176, 2018.

[37] D. Nicholas, I. Rowlands, D. Clark, and P. Williams. Google generation ii: web behaviour experiments with the bbc. *ASLIB proceedings*, 63(1):28–45, 2011.

[38] S.L. Nist and M.C. Hogrebe. The role of underlining and annotating in remembering textual information. *Literacy Research & Instruction*, 27(1):12–25, 1987.

[39] H.L. O'Brien, A. Kampen, A.W. Cole, and K. Brennan. The role of domain knowledge in search as learning. In *Proc. 5th ACM CHIIR*, pages 313–317, 2020.

[40] D. Persico and K. Steffens. Self-regulated learning in technology enhanced learning environments. In *Tech. Enhanced Learning*, pages 115–126. 2017.

[41] H.R. Ponce and R.E. Mayer. An eye movement analysis of highlighting and graphic organizer study aids for learning from expository text. *Computers in human behavior*, 41:21–32, 2014.

[42] M. Pressley. What should comprehension instruction be the instruction of? *Handbook of reading research*, 3:545–561, 2000.

[43] S.R. Putra, K. Grashoff, F. Moraes, and C. Hauff. On the development of a collaborative search system. In *DESIRES*, pages 76–82, 2018.

[44] N. Roy, F. Moraes, and C. Hauff. Exploring users' learning gains within search sessions. In *Proc. 5th ACM CHIIR*, page 432–436, 2020.

[45] J.L. Shefelbine. Student factors related to variability in learning word meanings from context. *J. Reading Behavior*, 22(1):71–97, 1990.

[46] X. Song, C Liu, and H. Liu. Characterizing and exploring users' task completion process at different stages in learning related tasks. *Proceedings of the Association for Information Science and Technology*, 55(1):460–469, 2018.

[47] R. Syed and K. Collins-Thompson. Retrieval algorithms optimized for human learning. In *Proc. 40th ACM SIGIR*, pages 555–564, 2017.

[48] R. Syed and K. Collins-Thompson. Exploring document retrieval features associated with improved short-and long-term vocabulary learning outcomes. In *Proc. 3rd ACM CHIIR*, pages 191–200, 2018.

[49] R. Syed, K. Collins-Thompson, P.N. Bennett, M. Teng, S. Williams, W.W. Tay, and S. Iqbal. Improving learning outcomes with gaze tracking and automatic question generation. In *Proc. 29th WWW*, pages 1693–1703, 2020.

[50] S. Wang, D.S. Unal, and E. Walker. Minddot: Supporting effective cognitive behaviors in concept map-based learning environments. In *Proc. 38th ACM CHI*, pages 1–14, 2019.

[51] M. Wesche and T.S. Paribakht. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Lang. Review*, 53(1):13–40, 1996.

[52] L.O. Wilson. Anderson and krathwohl–bloom's taxonomy revised. *Understanding the New Version of Bloom's Taxonomy*, 2016.

[53] M.J. Wilson and M.L. Wilson. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *JASIST*, 64(2): 291–306, 2013.

[54] L. Xu, X. Zhou, and U. Gadiraju. How does team composition affect knowledge gain of users in collaborative web search? In *Proc. of the 31st ACM HT*, pages 91–100, 2020.

[55] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, and S. Dietze. Predicting user knowledge gain in informational search sessions. In *Proc. 41st ACM SIGIR*, pages 75–84, 2018.

[56] C.L. Yue, B.C. Storm, N. Kornell, and E.L. Bjork. Highlighting and its relation to distributed study and students' metacognitive beliefs. *Edu. Psy. Review*, 27(1): 69–78, 2015.

[57] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J. M. Jose, and L. Azzopardi. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval*, 16(2):267–305, 2013.