Searching to Learn with Instructional Scaffolding

Câmara, A.; Roy, Nirmal; Maxwell, David; Hauff, Claudia

DOI
10.1145/3406522.3446012

Publication date
2021

Document Version
Final published version

Published in
CHIIR 2021

Citation (APA)
Câmara, A., Roy, N., Maxwell, D., & Hauff, C. (2021). Searching to Learn with Instructional Scaffolding. In *CHIIR 2021 : Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (pp. 209-218). ACM. https://doi.org/10.1145/3406522.3446012

Searching to Learn with Instructional Scaffolding

Arthur Câmara, Nirmal Roy, David Maxwell, Claudia Hauff
Delft University of Technology
Delft, The Netherlands
{a.barbosacamara,n.roy,d.m.maxwell,c.hauff}@tudelft.nl

ABSTRACT

Web search engines are today considered to be the primary tool to assist and empower *learners* in finding information relevant to their learning goals—be it learning something new, improving their existing skills, or just fulfilling a curiosity. While several approaches for improving search engines for the learning scenario have been proposed (e.g. a specific ranking function), *instructional scaffolding* (or simply *scaffolding*)—a traditional learning *support* strategy—has not been studied in the context of search as learning, despite being shown to be effective for improving learning in both digital and traditional learning contexts. When scaffolding is employed, instructors provide learners with support throughout their autonomous learning process. We hypothesize that the usage of scaffolding techniques within a search system can be an effective way to help learners achieve their learning objectives *whilst searching*. As such, this paper investigates the incorporation of *scaffolding* into a search system employing three different strategies (as well as a control condition): *(i)* $\texttt{AQE}_{\texttt{SC}}$, the automatic expansion of user queries with relevant subtopics; *(ii)* $\texttt{CURATED}_{\texttt{SC}}$, the presenting of a manually curated static list of relevant subtopics on the search engine result page; and *(iii)* $\texttt{FEEDBACK}_{\texttt{SC}}$, which projects real-time feedback about a user's exploration of the topic space on top of the $\texttt{CURATED}_{\texttt{SC}}$ visualization. To investigate the effectiveness of these approaches with respect to human learning, we conduct a user study ($N = 126$) where participants were tasked with searching and learning about topics such as *'genetically modified organisms'*. We find that *(i)* the introduction of the proposed scaffolding methods in the proposed topics does not significantly improve learning gains. However, *(ii)* it does significantly impact search behavior. Furthermore, *(iii)* immediate feedback of the participants' learning ($\texttt{FEEDBACK}_{\texttt{SC}}$) leads to undesirable user behavior, with participants seemingly focusing on the feedback gauges instead of learning.

ACM Reference Format:
Arthur Câmara, Nirmal Roy, David Maxwell, Claudia Hauff. 2021. Searching to Learn with Instructional Scaffolding. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3406522.3446012

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

CHIIR '21, March 14–19, 2021, Canberra, Australia.

ACM ISBN 978-1-4503-8055-3/21/03.
https://doi.org/10.1145/3406522.3446012

1 INTRODUCTION

Search and sensemaking are an intricate part of a user's learning process. For many learners today this is synonymous with accessing and ingesting information through web search engines [5, 28, 35]. Despite this, web search engines are not equipped to support users in the type of complex searches often required in learning situations[1] [15, 16, 24]. The process of what is now known as *Search as Learning (SAL)* [7] was first formally defined by Marchionini [24] as an iterative process, *mediated by a search system*, where learners purposefully engage by reading, scanning and processing a large number of documents with the ultimate goal of gaining knowledge about one specific learning objective. The finding, understanding, analyzing and evaluation [19, 45] of documents containing information relevant to answering this question is a time-consuming and cognitively demanding process.

Recently, a number of different research efforts have been devoted to the area of SAL, such as: *(i)* the influence of user characteristics and user strategies on learning while searching [13, 20, 22, 29, 30, 34]; *(ii)* the exploration of user behavior during learning-oriented search sessions [12, 13, 27, 34]; *(iii)* the prediction/observation of how knowledge changes over time and across different cognitive levels of learning [18, 21, 34, 49, 51]; *(iv)* the measuring of learning during searches [3, 4, 12, 46, 50]; and *(v)* the design of retrieval algorithms for learning-oriented search tasks and user interface components [40, 42, 43]. Despite the large number of prior works in the SAL field, only a small number have so far explored the *adaptation of the search system itself* to improve learning outcomes.

During a learning-oriented search session, realizing *what they do not know* about a topic is a key hurdle for learners to overcome. Previous work [51] has shown that learners, on average, are aware of only 40% of the different aspects pertaining to a topic before the search session commences. To counter this issue, the learning sciences provide us with the concept of *instructional scaffolding* for a classroom environment [6, 26, 32, 47]. Using scaffolding, an instructor or teacher provides *guidance* to learners through various means in order for them to achieve their learning goals. During the early stages of learning, these scaffolds provide plenty of structure and direction. Over time however, the responsibility of identifying core concepts about a topic shifts from the scaffolding to the learner. By the end of the learning process, the scaffold is withdrawn as no more guidance should be required.

When translating the idea of instructional scaffolding to digital learning, Hill and Hannafin [17] proposed a number of different scaffolding components. Of special interest to us are the so-called *conceptual scaffolds* (analogous to *topical outlines*), designed to "assist the learner in deciding what to consider or to prioritize what

[1] As a concrete complex search example, in our experiments we ask participants to learn about *radiocarbon dating considerations* (among other topics).

is important." In this paper, we explore to what extent conceptual scaffolds—which have been shown to be beneficial for human learning in digital learning environments—are beneficial for learning while searching.

To this end, we propose three different strategies of incorporating scaffolding into learners' search sessions: *(i)* $\texttt{AQE}_{\texttt{SC}}$, the *automatic expansion* of users' *queries* with relevant subtopics (i.e. key aspects of the topic to learn more about) as predefined by an expert; *(ii)* $\texttt{CURATED}_{\texttt{SC}}$, the presentation of a manually curated static list of relevant subtopics on the search engine result page, as also discussed recently by Smith and Rieh [38] (in contrast to $\texttt{AQE}_{\texttt{SC}}$ the learner here is explicitly aware of the subtopics related to the main topic); and *(iii)* $\texttt{FEEDBACK}_{\texttt{SC}}$, which projects real-time feedback about the users' exploration of the topic space on top of the $\texttt{CURATED}_{\texttt{SC}}$ visualization. This is inspired by recent works like ScentBar [44] and von Hoyer et al. [46], who posit that a better calibration of one's self-assessment of learning during search sessions can be achieved through the provision of automatically generated feedback that indicates learning progress.

We implemented these scaffolding variants on top of the SearchX framework [31], and conducted an inter-subject study, where 126 participants were randomly assigned to one of four conditions (the three variants introduced above, plus CONTROL, a standard search interface) to assess how conceptual scaffolds impact human learning while searching. By measuring the participants' knowledge before and after each learning-oriented search session, we were able to measure their *knowledge gain*. With this *Interactive Information Retrieval (IIR)* experiment, we aim to answer the following research questions:

RQ1 Is conceptual scaffolding beneficial to improve learners' knowledge gain compared to a standard search system setup?

RQ2 When scaffolding is introduced, to what extent does learners' search behavior change?

Our main findings can be summarized as follows. *(i)* The proposed scaffolding methods are shown to not be significantly effective for increasing learners' knowledge gain, with gains ranging from 30% to a detrimental effect of 7%, when compared to the control condition. *(ii)* The type of scaffold has a significant impact on learners' search behavior. We also show the participants' queries to be heavily influenced by the scaffolding components. *(iii)* Participants in the $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{FEEDBACK}_{\texttt{SC}}$ conditions were more engaged with the platform, and issued more queries, viewed more documents, and spent more time searching. At the same time, the $\texttt{FEEDBACK}_{\texttt{SC}}$ cohort exhibited behavior indicating that they focused on the feedback gauge more than the actual learning process.

2 RELATED WORK

We now discuss the main findings in the SAL literature, which has been inspired by the observation that learners increasingly turn to the web (and thus web search engines) to support their learning needs [28, 33].

Influence of user characteristics and task characteristics on learning. O'Brien et al. [29] explored the impact of domain expertise (experts vs. non-experts) on learning, with expertise determined based on participants' self-reported frequency of searches for historical information. No significant differences in learning outcomes were found between the two groups. This is in contrast to Gadiraju et al. [13], where slightly higher knowledge gains for participants with less prior knowledge were observed. Roy et al. [34] investigated *when* during a search session learning takes place, and did observe differences between expert and non-expert learners, specially towards the end of the search session. Other than domain expertise, learners' cognitive abilities (such as working memory capacity and reading comprehension ability [30]) were found to be predictive of learning outcomes. In terms of user strategies for learning, Liu and Song [20] found learners who adapt their source selection to the type of task at hand (encyclopedia-style documents for receptive tasks and Q&A documents for critical tasks) to have better learning outcomes than learners who do not adapt their source selection strategy. Kalyani and Gadiraju [18] explored to what extent a search task's cognitive learning level (based on the revised Bloom's taxonomy [1]) impacts user behavior: as a trend, the higher the cognitive level of the search task (such as *remembering* vs. *applying*), the larger the amount of search interactions. Based on this insight, the authors proposed to train models in a supervised manner to automatically determine the complexity of a user's information needs. This in turn could lead to adaptive search systems that are optimized for learning.

Proxy measures of learning. The vast majority of the aforementioned works measure knowledge gain through knowledge tests: those are often multiple-choice tests, but manually annotated user summaries [20, 29, 50] and mind maps [21] have been explored as well. Knowledge tests are admitted before and after the search session in order to determine the knowledge gains throughout the session. However, in order to build search systems that are adaptive to users' learning needs, we require *scalable and easy to collect behavioral metrics* that are predictive of knowledge gain. Past works have explored which behaviors can be considered to be predictive of learning. Document dwell time was found to be indicative of learning [8, 12], as well as the number of *Search Engine Results Page (SERP)* clicks [8], the occurrence of contextually relevant imagery [42], and the diversity of the domains present among the top-ranked documents [12]. Pardi et al. [30] studied the relationship between the kind of documents users dwell on and their learning outcomes. They found *text-dominated* documents to be more effective for learning than video documents (this though is in contrast to the findings by Moraes et al. [27], who find high-quality video material to yield higher learning outcomes than searching). In contrast to the aforementioned studies that considered just a handful of predictors, Yu et al. [51] evaluated approximately 70 search-based features as predictors of learning; individually they were found to be only weakly correlated with knowledge gain, though some (document dwell time, query complexity) were more predictive than others. Although unarguably less scalable, a number of eye-tracking measures (such as the duration or reading fixations within documents) have also shown to be predictive of learning outcomes [3, 4]. Finally, we note that von Hoyer et al. [46] also found learners to be able to estimate their learning performance with increasing accuracy as the search session progresses.

Retrieval system adaptation. Few works have explored so far the adaptation of the retrieval system itself to support learning. Syed and Collins-Thompson [40] designed a retrieval algorithm specifically for vocabulary learning by ranking documents according

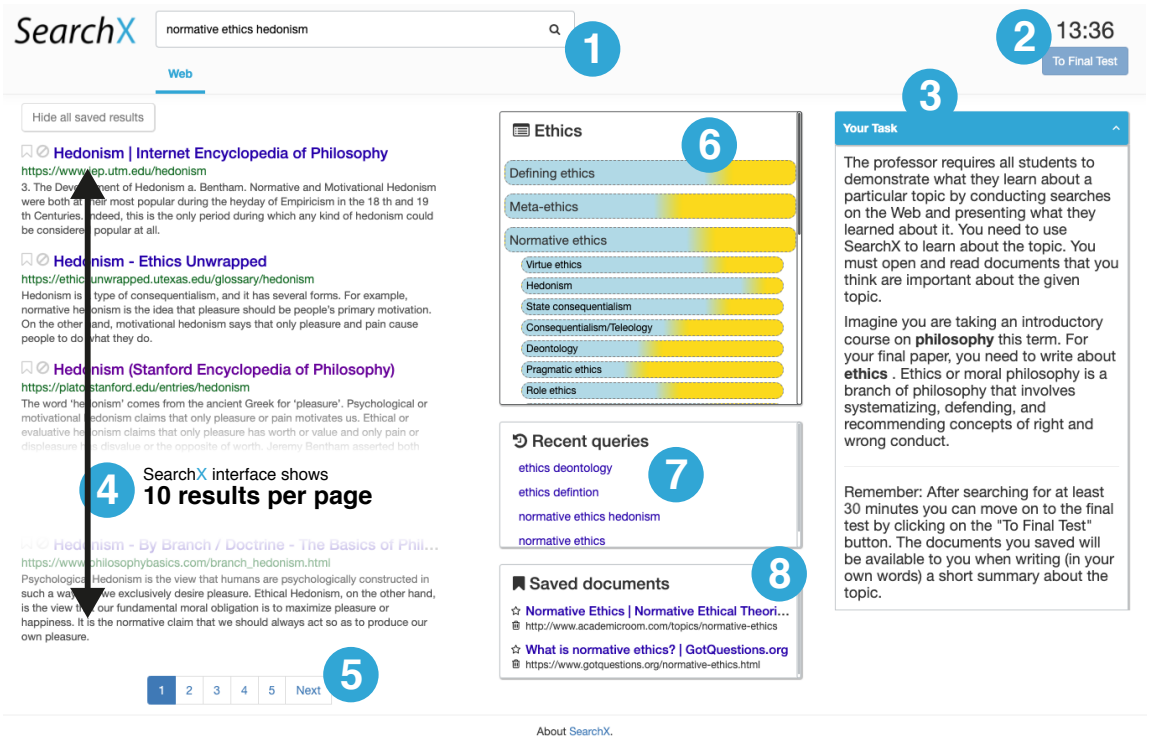

Figure 1: The SearchX interface: the eight annotated interface components are described in Section 3. Note that the scaffolding component (displaying the Ethics topic) shows the FEEDBACK$_{SC}$ scaffolding variant—complete with yellow progress gradients.

to their keyword density of the vocabulary items to learn. The user evaluation showed that at least for some topics, results with a higher keyword density leads to significantly higher learning gains (a follow-up study showed this to hold over a long period of time as well [42]). However, it should be noted that the user study fixed the documents to read for each topic, instead of allowing participants to search and adapting the retrieved results on the fly. Recently, Syed et al. [43] investigated whether automatic question generation can be utilized to improve learning outcomes *whilst reading* a document. Although the improvement in learning outcome was limited to learners with low levels of prior knowledge, it is not far fetched to imagine such an interface component to also be incorporated in a search system.

Visualization of search progress. Lastly, we want to point to the work on ScentBar by Umemoto et al. [44] which—though unrelated to SAL—inspired one of our scaffolding variants (FEEDBACK$_{SC}$): it is a query suggestion interface that visualizes to what extent information relevant to the information need remains unexplored. A user study on a number of intrinsically diverse tasks showed that users were indeed better able to determine when to stop searching for relevant information when the amount of missed information was made visible to them.

3 INSTRUCTIONAL SCAFFOLDING IN SEARCHX

We implemented our scaffolding variants as part of SearchX [31], a modular, open-source search framework which provides quality control features for crowdsourcing experiments and fine-grained search logs[2]. Figure 1 showcases the user interface we designed for our experiments. The eight main components are listed here. (1) denotes the query box (without query auto-completion). (2) represents the countdown timer to help our participants gauge the remaining minimum task time. (3) highlights the task description. We show (4) ten search results per page (each document can be saved to the Saved documents component for later usage, or hidden from future SERPs). Pagination is enabled (5). (6) shows the scaffolding component, with the FEEDBACK$_{SC}$ variant illustrated here (complete with yellow progress gradients). (7) shows the list of all issued queries so far in the search session, and (8) shows the list of all documents *saved* so far in the search session. It should be noted that interface components (6), (7) and (8) provide scrollbars to scroll through content in each component. In the remainder of this section, we focus our discussion on our scaffolding variants, after introducing the approach behind our topical outlines.

3.1 Topical Outlines

A key ingredient of all our scaffolding strategies are the topical outlines for each learning topic (cf. Figure 1, where the scaffolding component shows part of the outline for the topic Ethics). Effective outlines are typically hierarchical in nature [14, 17], and follow a *specific order* (ideally one that is best suited to master the topic). By providing such structure, we can point a learner toward a list of *subtopics*—or topical aspects—that are important to the main topic.

[2]Behaviors logged include document dwell time, clicked documents, mouse hovers, document snippets shown on screen, bookmarked documents, etc.

Such outlines can either be created by instructors [2, 36] or automatically (this is known as *outline generation* [52]). The latter is desirable as it is scalable and not dependent on the availability of a domain expert—this is however a nontrivial challenge. For this reason, we rely upon manually created outlines for this study. More specifically, we used the heading structure of the corresponding Wikipedia article for each of our topics, as provided by the TREC CAR 2017 dataset [11][3]. This can be considered as employing a *crowd of experts* [23] for creating the outline. A concrete example outline from Wikipedia for the *subprime mortgage crisis* topic is shown in Figure 2. Each outline was manually cleaned; we only consider subtopics in our outline up to two levels deep (we refer to those levels as $L1$ and $L2$, cf. Figure 2) and we remove generic subtopics that occur across a range of topics (such as *References*).

3.2 Variant AQE_{SC}

Scaffolding can be incorporated in different ways within a search system. It can be incorporated in the frontend (as we explore with $\text{CURATED}_{\text{SC}}$ and $\text{FEEDBACK}_{\text{SC}}$), or the backend. In the backend, we can either modify the retrieval function (as proposed by Syed and Collins-Thompson [41, 42]), or reformulate the *to-be-submitted* queries. We chose the latter setup, as this is agnostic to the employed search engine (*Bing* in the present study, via the *Bing Search API*). More specifically, we reformulated each user query by appending the topic name (e.g. *subprime mortgage crisis*) and one of the $L1$ subtopics (e.g. *causes*) before submitting it to the search backend. Which subtopic we appended was dependent upon the *time* the query was submitted during the search session. Each $L1$ subtopic was considered *active* an equal amount of time. For example, for a search session estimated to last 30 minutes[4], a topic with six $L1$ subtopics will have each subtopic active for five minutes. We chose to only include $L1$ topics here, as: *(i)* the inclusion of $L2$ topics (of which there are usually two or three times as many) would lead to too many topical changes in a short period of time; and *(ii)* the returned search results would often be overly specific. We kept the order of the subtopics as present in the topical outline intact. The search interface the study participants see in this variant is as shown in Figure 1, but without the 6 scaffolding component. Finally, we note that the CONTROL variant has the same user interface as AQE_{SC}, but no automatic query expansion is employed. Additionally, the participants had no visual indication that their queries were modified.

3.3 Variant $\text{CURATED}_{\text{SC}}$

As already mentioned, the next two scaffolding techniques are focused on changes to the frontend. Here, we explore to what extent making learners *explicitly aware* of the topical outline impacts their search behaviors and knowledge gains. The first variant, $\text{CURATED}_{\text{SC}}$, is as seen in Figure 1, though *without* the yellow progress gradient (i.e. component 6 is static, with solid blue backgrounds throughout). The scaffolding component contains the topic name (here: Ethics) and a list of $L1$ and $L2$ subtopics in order. As previously mentioned, the component has fixed dimensions, but can be scrolled

Figure 2: Hierarchical topic structure for the topic *Subprime Mortgage Crisis*. Topic and structure derived from TREC CAR 2017 [11]. Note that third level subtopics (and deeper) and footnotes/references are excluded (illustrated in the figure by use of ~~strikethroughs~~).

at anytime. While the task description does not point explicitly to the component (as seen on the right of Figure 1), we do introduce the component in an interactive tutorial before the start of the search session as follows:

> *This is a list of important subtopics. Each sub-topic can itself be broad enough to be split into several sub-topics. Explore the subtopics as much as you can.*

The intuition behind this scaffolding choice is that learners that are pursuing a given list of curated subtopics should achieve higher knowledge gain than those searching without this guidance.

3.4 Variant $\text{FEEDBACK}_{\text{SC}}$

While $\text{CURATED}_{\text{SC}}$ presents a static component to the learner which does not change during the search session, in $\text{FEEDBACK}_{\text{SC}}$ we provide feedback about the learners' progress throughout the search session. To do this, we estimate the exploration of each subtopic, and display this information as a progress bar as shown in Figure 1, inspired by Umemoto et al. [44]. In contrast to their approach, we cannot precompute the match of each document in the corpus to each subtopic (as we are using the open web, rather than a static corpus). The computation of how a list of viewed documents contributes to the progress of each subtopic is therefore nontrivial. This needs to happen in (near) real-time to avoid a noticeable lag.

Each time 10 search results (documents) are retrieved from the Bing Search API for a given query, we compute, for each document/subtopic pair the semantic similarity between them. To this end, we tokenize both document and subtopic[5], and extract their sentence embedding using a pre-trained BERT-base model [10][6]. We then compute the cosine similarity between both embeddings, and that score, between 0 and 1, is used to increase the progress bar for the respective subtopic *if the user views the respective document.*

[3]We note that topical outlines can also be extracted from text books or online courses; we picked Wikipedia here as source of our outlines since these are naturally hierarchical and readily available in the TREC CAR dataset (outlined in more detail in Section 4).
[4]As we set a minimum task time of 30 minutes in our study this is a reasonable setup.

[5]For tokenization we employ https://github.com/huggingface/tokenizers.
[6]Here, we follow the recommendations proposed by the authors of BERT of averaging all token embeddings from the second-to-last layer: https://github.com/google-research/bert/issues/71.

As this pairwise operation is expensive to do in near real-time (e.g. for the topic *'noise induced hearing loss'* with 27 subtopics, we have to compute 270 document/subtopic similarities each time), we employ two additional filters that can be computed quickly: *(i)* we remove documents with fewer than 50 tokens from consideration (there is little to learn in those cases), as well as *(ii)* documents which contain less than 20% of the unique terms in the section of the Wikipedia article for the subtopic.

Thus, the similarity score of document D_i for subtopic t_j can be computed as follows:

$$S(D_i, t_j) = \begin{cases} \frac{\phi(D_i)\cdot\phi(t_j)}{||\phi(D_i)||\times||\phi(t_j)||}, & \text{if } |D_i| > 50 \wedge \frac{|D_i \cap t_j|}{|t_j|} > 0.2 \\ 0, & \text{otherwise} \end{cases}$$

where $\phi(\cdot)$ is the embedding operation described before. Each viewed document can thus contribute to the progress score of multiple subtopics. We a consider subtopic's progress bar completely *'filled up'* when the aggregate similarity score reaches 10. This is a constant that is determined based on the search session length, and the number of subtopics present.

4 USER STUDY SETUP

Having outlined our scaffolding variants, we now consider the overall study setup, including a discussion on our choice of topics, the metrics we employ to measure learning gain, our study participants, and the workflow we followed.

4.1 Topics

We used a subset of the 117 training topics from the TREC CAR 2017 [11] dataset. This dataset is a set of outlines, extracted from Wikipedia headings, with the original goal being to find relevant passages for each of these headings. This structure makes this dataset a good match for this task, since it already provides the required hierarchical topical outlines.

We extracted the 100 topics whose topical outlines have at least two hierarchy levels, and then filtered those to an initial set of 48 by discarding topics that lack complexity. Of those, we picked 10 topics based on their difficulty and complexity, judged by 17 STEM graduate students[7]. Finally, we removed 3 topics: *'Norepinephrine'*, as the Wikipedia page of the topic was mostly comprised of images; *'research in lithium ion batteries'*, which contains a much larger number of subtopics (almost 50) than our other topics; and *'theory of mind'*, which showed to be too easy, as almost no study participant was assigned to it (cf. Section 4.3 for how users were assigned to each topic). In the end, we worked with the remaining 7 topics, which are listed in Table 2. Each of the topics selected has between 11 and 27 subtopics. The choice for the most difficult topics was made so that we could maximize the potential learning of the participant during the experiment, and that any knowledge gained would be clearly apparent.

In order to measure users' learning gains, we followed the established approach of resorting to a pre- and a post-test of important concepts related to a topic [13, 27, 34, 40, 51] (i.e. users are queried about their knowledge of the concepts *before and after* the search session). In line with previous works, we resorted to a vocabulary knowledge test as the mean to evaluate domain knowledge. To this end, two of the authors manually selected 10 concepts per topic (listed in Table 1) from the corresponding Wikipedia article—after an initial list of 100 candidate unigram/bigram concepts were automatically extracted using the highest IDF scores, computed on the TREC CAR 2017 corpus (a subset from Wikipedia), post stopword removal. When choosing the concepts, we aimed to pick the most representative terms for each topic by analyzing the respective Wikipedia article. Some unigrams and bigrams were further combined when needed for context (e.g. `inquiry commission` → `financial crisis inquiry commission`) and stopwords were also re-introduced when needed (e.g. `overstimulation hair cells` → `overstimulation of hair cells`).

Table 1: Overview of the 10 concepts per topic utilized in the pre- and post tests. Highlighted are the easiest and most difficult two concepts per topic: marked in orange (yellow) are the two concepts of each topic with, on average, the lowest (highest) post-test knowledge scores.

Business cycle	economic cycles, distribution cycles, swing cycle, wage cycle, marxist model, endogenous causes, friedman, capital profitability, model recession, austrian school
Ethics	anarchist ethics, descriptive ethics, normative ethics, relational ethics, virtue ethics, ethical resistance, consequentialism, epicurean ethics, ethics feasible, ethics spheres
Genetically modified organism	transgenic, genomes, selective breeding, microinjection enzyme, chromosome, plasmid, myxoma, kanamycin, severe combined immunodeficiency, Leber's congenital amaurosis
Irritable bowel syndrome	bifidobacteria infantis, mesalazin, bile acid malabsorption, selective serotonin reuptake inhibitors, gut-brain axis, antidepressants, laxatives, probiotics, celiac disease, epithelial barrier
Noise-induced hearing loss	acoustic trauma, discomfort threshold, cochlear damage, audiogram, overstimulation of hair cells, noise conditioning, excitotoxicity, OSHA, sensorineural hearing loss, tinnitus, threshold shift
Radiocarbon dating considerations	carbon exchange reservoir, isotopic fractionation, polarity excursion, carbonate, geomagnetic reversals, mass spectrometry, upwelling, radiocarbon, neutrons, photosynthesis pathways
Subprime mortgage crisis	mortgage, subprime, financial crisis inquiry commission, securities, ben bernanke, investment banks, housing bubble, lehman brothers, foreclosures, default

4.2 Metrics

We evaluate the knowledge gain of a concept by utilizing the *Vocabulary Knowledge Scale (VKS)* [48] across four levels (in line with [27, 34]):

(1) *I don't remember having seen this term/phase 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 ...*

This means that in both the pre- and post-tests, study participants were asked to rate themselves on their knowledge levels of each concept. Note that a self-assessment of *(3)* or *(4)* requires participants to write down a definition of the concept in their own words, which in turn allows us to grade the quality and reliability of the self assessment. It's also worth mentioning that the participants were not aware, at the start of the experiment, that the same questions would be asked again in the post-test, as this could influence their search behavior.

In order to compute the learning gain, we assign a score of `0` to both knowledge levels *(1)* and *(2)*. Since level *(3)* indicates the participant is not certain about a concept's meaning, we assign it a

[7] Each assessor received all 48 topics, in a randomized order, and was asked to select the 10 that appeared most difficult to them for learning about. Finally, the 10 topics selected most often were chosen as our topic set.

Table 2: Overview of the topics used in our study, with associated statistics. Two-way ANOVA tests revealed no significant differences in average number of queries between topics ($F(6, 99) = 2.01, p = 0.07$), or between the average number of bookmarks ($F(6, 99) = 0.41, p = 0.87$).

	[1]Business cycle	[2]Ethics	[3]Genetically modified organisms	[4]Irritable bowel syndrome	[4]Noise induced hearing loss	[6]Radiocarbon dating considerations	[7]Subprime mortgage crisis
Level 1 subtopics	4	6	5	10	8	4	8
Level 2 subtopics	15	12	6	15	19	8	19
Study participants	16	20	15	15	19	21	20
Participants for CONTROL	3	3	4	4	5	6	5
Participants for AQE_{SC}	3	5	3	3	3	5	6
Participants for $CURATED_{SC}$	4	5	4	3	4	6	7
Participants for $FEEDBACK_{SC}$	6	7	4	5	7	5	2
Average number of queries	11.1(±6.4)	11.4(±8.0)	6.6(±3.9)	10.1(±8.8)	7.9(±6.6)	7.8(±5.5)	5.8(±3.4)
Median number of queries	9.5	9.5	6.0	7.0	7.0	6.5	5.0
Average number of bookmarks	6.9(±6.4)	8.8(±6.1)	6.1(±3.5)	7.7(±6.9)	10.1(±11.0)	10.0(±22.6)	5.5(±5.9)
Median number of bookmarks	4.5	7.0	6.0	5.0	5.0	5.0	4.0

score of 1. Choosing level *(4)* indicates the participant is confident in their assessment, and we assign it a score of 2.[8]

In line with [9, 27, 37, 41, 42], we utilize *Realized Potential Learning (RPL)* as our main learning gain metric. RPL depends on the *Absolute Learning Gains (ALG)* which is measured in terms of either the number of new concepts learned (indicated by a score change of 0 to 1 or 0 to 2 from pre-test to post-test), *or* the number of concepts they became more confident at (indicated by a score change of 1 to 2 from pre-test to post-test). RPL normalizes ALG by the *maximum possible learning potential (MLG)*, which is 2 if the pre-test score is 0 or 1 if the pre-test score is 1. And thus, for n concepts:

$$ALG = \frac{1}{n}\sum_{i=1}^{n} max(0, vks^{post}(v_i) - vks^{pre}(v_i))$$

$$MLG = \frac{1}{n}\sum_{i=1}^{n} 2 - vks^{pre}(v_i)$$

$$RPL = \frac{ALG}{MLG}.$$

Here, $vks^X(v_i)$ is our assigned score of concept v_i (0, 1 or 2), X is either *pre* or *post* and $n = 10$. Intuitively, RPL measures the percentage of knowledge gained from the total possible knowledge to be gained. To provide the reader with some intuition, we provide concrete examples of how pre/post-test scores translate into RPL in Figure 3.

We note that *ALG* and *RPL* are not the only possible metrics to measure learning. Instead of treating each concept in the same manner, *difficulty weighted learning gains* can be computed too (as for instance done by Syed and Collins-Thompson [42], where vocabulary items such as `earth` and `temperature` were mixed with more technical vocabulary items). Based on the manner we selected our concepts, we do not believe this to be necessary as they are similarly difficult. Some prior works have also manually annotated participants' summaries or mind maps to derive qualitative and quantitative metrics [21, 29, 50]. We leave the analyses of the user summaries we collected in this manner for future work.

[8]We note that this scoring scheme is equivalent to the *fine-grained setup* employed by Moraes et al. [27].

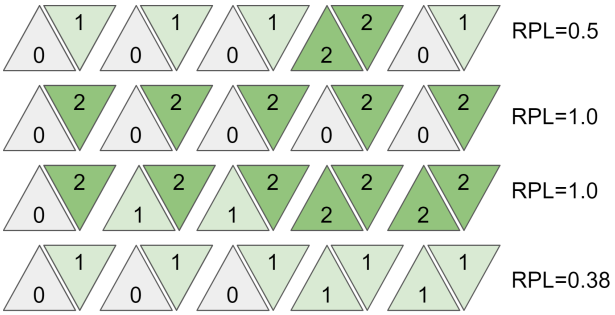

Figure 3: RPL examples: △ represents vks^{pre} and ▽ represents vks^{post}. Here, $n = 5$.

4.3 Study Workflow

The flow of our user study is presented in Figure 4; it is implemented within our SearchX instance. When a participant enters the study, two of our seven topics are randomly selected. In addition to this (and to weed out non-complying crowd workers), we add the topic *'sports'* to the pre-test as we expect reasonable participants to demonstrate high knowledge levels on this topic. The pre-test thus consists of 30 VKS questions in total. We rejected crowd workers that score lower on *'sports'* than the other two topics. The topic they know the least about is then chosen as the one to learn about during the search session. We introduced the simulated learning task as shown in Figure 1, item 3. The minimum task time was set to thirty minutes. We also filtered any web document returned from the Bing Search API that either came from a Wikipedia domain, or domains that are known clones of Wikipedia[9]. Wikipedia and its clones were excluded as we drew our topical outlines from the relevant Wikipedia article – the said Wikipedia article would therefore be the best to read. While for a large portion of topics Wikipedia is a great tool for learning, we cannot expect good Wikipedia pages for all topics, especially niche or highly specific topics. Therefore, we believe that the formulation outlined in this section is still a reasonable search task.

[9]We blacklisted a total of 72 domains. All subtopics were submitted to the Bing Search API, with the top 10 results returned. Each result was then inspected to determine whether it came from a Wikipedia clone in our blacklist.

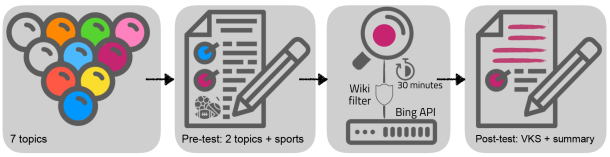

Figure 4: Overview of the flow of our user study.

During the search session, participants could search, view and bookmark documents. In the post-test, we asked them again about their knowledge on the ten concepts for their topic. In addition, we asked them to write a summary (100 words minimum)[10] about the topic. We note here that the knowledge tests require understanding, but no application or synthesis (i.e. higher-level cognitive processes of learning [19]) of the materials—here we make the bookmarked documents available to our participants.

4.4 Study Participants

We conducted our study on the *Prolific Academic platform*[11] across three days. In order to ensure responses of high quality, we required our participants to have at least 15 previous submissions, an approval rate of 90+% and be native English speakers. The study took about an hour to complete and participants were reimbursed with £6 per hour. 144 participants completed our study. We had to reject 18 participants (leading to $N = 126$ valid participants) as they had completed more than three browser tab changes (we enforced this rule to ensure our participants actively using our search system instead of running down the timer). Of the valid participants, 65 were male, 59 female (2 withheld gender information) with a median age of 27 (minimum 18, maximum 63). 44 participants reported as highest formal education level a high school degree, 47 reported a Bachelor's degree and 20 had a Master's degree. The remaining 15 participants indicated other educational levels.

In Table 2, we report the number of participants per topic. The maximum number of participants were assigned to the topic *'radiocarbon dating considerations'* (21), while the minimum were assigned to *'genetically modified organisms'* and *'irritable bowel syndrome'* (15). The table also contains statistics on the number of queries and bookmarks per topic, indicating that our study participants actively engaged in the search session. The median number of queries ranges from 5 to 9.5, with the median number of bookmarks ranging from 4 to 7 respectively across the topics.

At the end of the data collection, we had collected answers for 1260 VKS questions and 126 essays. In order to determine the quality of the VKS self-assessments, we sampled 100 concept definitions written by our participants: 50 for knowledge levels *(3)* and *(4)* respectively. Two annotators labeled them as *correct*, *partially correct*[12] and *incorrect*[13]. We find that 25.2% of the vocabulary scores self assessed as *(3)* were correct; 65.9% were partially correct; and the remaining 8.9% were incorrect. Among the definitions self assessed as *(4)*, 64.8% were correct, 28.9% were partially correct and the remaining 6.3% were incorrect. Based on these numbers, we consider the self-assessment to be largely reliable. Thus, we report RPL based on self-assessed vocabulary knowledge levels.

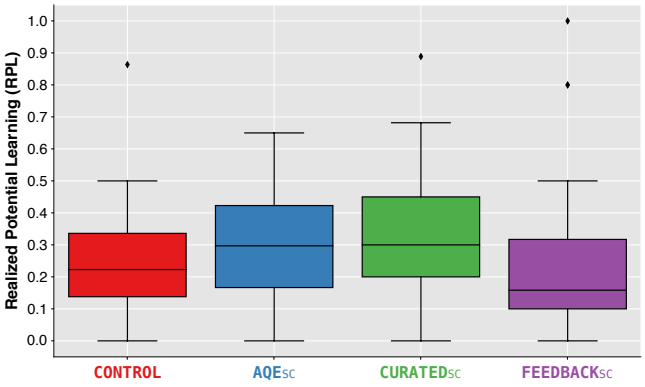

Figure 5: RPL over the four different conditions.

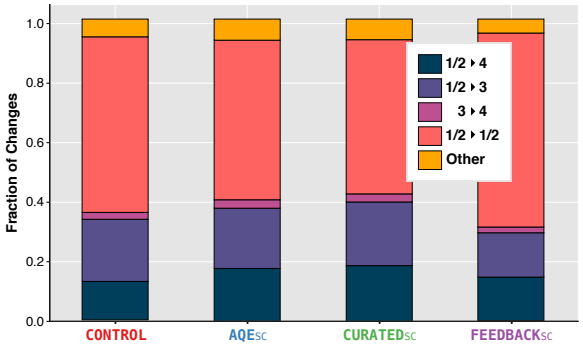

Figure 6: Fraction of change in vocabulary tests.

5 RESULTS

We now turn to addressing our research questions. In terms of statistical tests reported within this section, we performed two-way ANOVA tests (with two factors: the intervention type and topic), followed by a post-hoc two-way Tukey HSD pairwise test in case of significance ($\mathbf{p < 0.05}$).[14]

5.1 RQ1: Impact of Scaffolding on Learning

In Figures 5 and 6, we present the RPL across the four conditions (each one with between 28 and 36 participants, and an average search session duration[15] of more than 36 minutes, cf. Table 3), and a more fine-grained presentation of the knowledge changes.

Recall that RPL provides us insights into the amount of learning that has taken place with respect to the maximum possible amount of learning (which may differ per participant; some participants may have no prior knowledge of any of the ten concepts, while others have a good understanding of 2-3 concepts already). For the CONTROL condition, the mean RPL is 0.26. Participants in both AQE$_{SC}$

[10] Specifically, we phrased this as: *"Your professor also asks you to write a summary of what you learned about the topic you searched about. This summary should be enough that your fellow students that read it can get a first idea of what the topic is, without having to search for it themselves. Please write your summary here (at least 100 words)."*
[11] https://www.prolific.co/
[12] Partially correct definition example of *tinnitus* (i.e., noise induced hearing loss topic): *"hearing loud sounds in one's ears."*
[13] Incorrect definition example of *genomes* (genetically modified organism topic): *"the amount of chromosomes."*
[14] For further investigations, An anonymized version of the data is available at https://github.com/ArthurCamara/searchx-scaffolding
[15] We compute the search session duration as the time between the first query issued and the last viewed document closing.

Table 3: Mean (± standard deviations) of RPL and search behavior metrics across all participants in each condition. † indicates two-way Anova significance, while ^{C},$^{\mathcal{A}}$,$^{\mathcal{U}}$,$^{\mathcal{F}}$ indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) increases vs. CONTROL, AQE$_{SC}$, CURATED$_{SC}$ and FEEDBACK$_{SC}$ respectively.

		CONTROL	AQE$_{SC}$	CURATED$_{SC}$	FEEDBACK$_{SC}$
I.	**Number of participants**	30	28	33	36
II.	**Search session duration (minutes)**	$36m33s(\pm12m15s)$	$39m59s(\pm10m59s)$	$41m31s(\pm13m6s)$	$38m15s(\pm12m46s)$
III.	**RPL**	$0.26(\pm0.18)$	$0.30(\pm0.16)$	$0.31(\pm0.20)$	$0.24(\pm0.24)$
IV.	**Number of queries†**	$5.13(\pm2.61)^{\mathcal{UF}}$	$5.29(\pm2.98)^{\mathcal{UF}}$	$11.09(\pm6.99)^{\mathcal{CA}}$	$11.86(\pm7.60)^{\mathcal{CA}}$
V.	**Fraction of query terms coming from topical outline†**	$0.26(\pm0.28)^{\mathcal{UF}}$	$0.33(\pm0.31)^{\mathcal{UF}}$	$0.58(\pm0.34)^{\mathcal{CA}}$	$0.58(\pm0.29)^{\mathcal{CA}}$
VI.	**Fraction of topical outline terms used for querying†**	$0.04(\pm0.04)^{\mathcal{UF}}$	$0.05(\pm0.04)^{\mathcal{UF}}$	$0.32(\pm0.23)^{\mathcal{CA}}$	$0.34(\pm0.24)^{\mathcal{CA}}$
VII.	**Average time between queries (minutes)**	$5m57s(\pm5m26s)$	$6m31s(\pm8m31s)$	$3m31s(\pm2m45s)$	$3m52s(\pm4m40s)$
VIII.	**Average time between document close and next document load (secs.)**	$60.15(\pm27.17)$	$68.06(\pm33.44)$	$74.42(\pm45.14)$	$57.32(\pm39.13)$
IX.	**Average document dwell time (secs.)**	$76.77(\pm51.14)$	$100.61(\pm61.59)$	$92.15(\pm97.60)$	$55.33(\pm51.04)$
X.	**Number of unique documents viewed†**	$14.77(\pm8.85)$	$10.96(\pm4.08)^{\mathcal{F}}$	$14.09(\pm7.95)$	$18.50(\pm9.56)^{\mathcal{A}}$
XI.	**Number of unique document snippets viewed†**	$97.47(\pm47.37)^{\mathcal{F}}$	$81.07(\pm44.58)^{\mathcal{UF}}$	$136.42(\pm76.97)^{\mathcal{A}}$	$152.44(\pm84.23)^{\mathcal{CA}}$

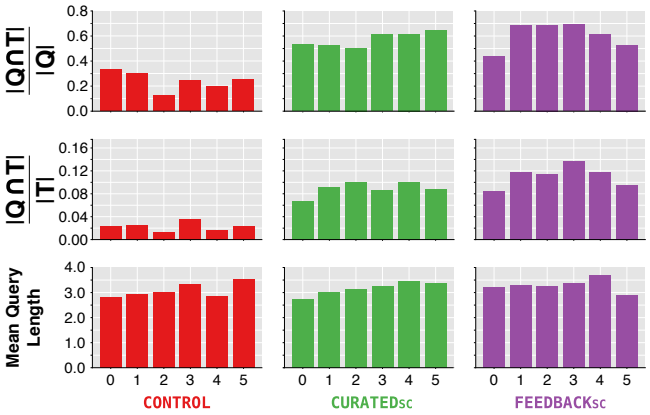

Figure 7: For each row: *(top)* the fraction of query terms taken from topic outlines; *(middle)* the fraction of topic outline terms used for querying; and *(bottom)* the mean query length, over 5 minute blocks (x axes) of the 30 minute search session, considering: CONTROL *(left)*; CURATED$_{SC}$ *(center)*; and FEEDBACK$_{SC}$ *(right)*. Here, we consider the first query instance as the start of the first interval.

and CURATED$_{SC}$ on average report higher learning gains with an RPL of 0.3 and 0.31 respectively. To evaluate the impact of AQE$_{SC}$ in the set of retrieved documents, we collected the SERPs of both the original user-formulated and automatically reformulated queries, and found that, among the top 10 retrieved documents, an overlap on average of 1.5 documents. That indicates that AQE$_{SC}$ had a great impact on how the SERP was presented.

Participants in the FEEDBACK$_{SC}$ condition had the lowest average RPL (0.24) as well as the highest standard deviation. This finding seems counter-intuitive, as the extra feedback available was hypothesized to be beneficial to the learning experience (as also envisioned, among others, by von Hoyer et al. [46]). We provide a further investigation of possible reasons for this finding in Section 5.2.

To further analyze how the conditions differ, we provide a detailed breakdown of the knowledge state transitions in Figure 6. We are particularly interested in the transitions from states 1/2 (where very little is known about a concept), to state 4 (where the concept is completely understood). The percentage of concepts for which this holds is largest among CURATED$_{SC}$ participants; similarly, the lack of knowledge increase (i.e. the transition $1/2 \rightarrow 1/2$) is smallest for this cohort. This result implies that the CURATED$_{SC}$ cohort, on average, was most confident in their knowledge increase.

Overall, we conclude that there is a lack of evidence to support the conclusion that scaffolding increases participants' learning gains, despite the positive trends we observe for AQE$_{SC}$ and CURATED$_{SC}$. We found no significant difference ($F(3, 99) = 0.75$, $p = 0.522$) between the four scaffolding conditions, which means that we cannot reject the null hypothesis that there is no learning gain difference among them. It is thus not as simple as introducing an outline or providing instantaneous feedback to yield reliable and large learning gains across a range of participants and across a range of topics.

5.2 RQ2: Search Behavior Analyses

Besides learning gains, we are also interested in the search behaviors of our participants. To answer our second research question, *When scaffolding is introduced, to what extent does learners' search behavior change?*, we report a number of search behavior metrics (mean and standard deviations) in Table 3.

5.2.1 Influence of visual scaffolds on querying. Our participants in the CURATED$_{SC}$ and FEEDBACK$_{SC}$ conditions issued significantly more queries (on average more than twice as many) than participants in the CONTROL and AQE$_{SC}$ conditions (in line with [44]). As a consequence, the average time between queries in those two conditions is much lower (less than four minutes on average, vs. more than six minutes on average) than in CONTROL and AQE$_{SC}$. We hypothesize that the readily available cues of what to query for enticed our participants to issue more queries, as they are aware of the various topical aspects. To validate this hypothesis—and in order to explore to what extent the participants in CURATED$_{SC}$ and FEEDBACK$_{SC}$ made use of these visual cues—we determined: *(i)* the percentage of unique query terms drawn from the topical outline; and *(ii)*, the percentage of unique terms in the topical outline present in at least one submitted query. To this end, we converted the queries ($\mathcal{Q}$) and topical outlines ($\mathcal{T}$) into bags-of-words with normalization (stopword removal, capitalization, etc.), and computed $\frac{|\mathcal{Q}\cap\mathcal{T}|}{|\mathcal{Q}|}$ as well as $\frac{|\mathcal{Q}\cap\mathcal{T}|}{|\mathcal{T}|}$. The results in Table 3 (rows V & VI) show

clearly that the presence of the outline influences the querying behavior significantly: more than half the query terms are *'borrowed'* from the topical outline in $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{FEEDBACK}_{\texttt{SC}}$, while this is the case for 33% and 26% on average for $\texttt{AQE}_{\texttt{SC}}$ and CONTROL respectively where participants had no access to the outline. In addition, when considering the coverage of the topical outline by query term, we see once again that a much larger percentage of outline terms were queried at least once ($> 30\%$ on average for $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{FEEDBACK}_{\texttt{SC}}$ vs. $\leq 5\%$ on average for the other two conditions) by participants in the variants with access to the outline. In the top two rows of plots in Figure 7, we break down this comparison of query terms and topical outline terms further by splitting our search sessions into five minute intervals, and computing $\frac{|Q \cap \mathcal{T}|}{|Q|}$ and $\frac{|Q \cap \mathcal{T}|}{|T|}$ separately for each interval. We find that participants in the CONTROL condition were not picking up more topical outline terms over time (despite the fact that they have read more documents on the topic by each passing interval). However, we do see a slight increase over time for $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{FEEDBACK}_{\texttt{SC}}$, which then drops again in the later stages of the search session.

5.2.2 Too much feedback considered harmful. Previous works [13, 27, 34, 51] have shown the number of queries issued to be a good proxy for learning. In our work, this finding holds for $\texttt{CURATED}_{\texttt{SC}}$, though not for $\texttt{FEEDBACK}_{\texttt{SC}}$: on average, a similarly high numbers of queries were submitted, but the learning gains for $\texttt{FEEDBACK}_{\texttt{SC}}$ are low. For completeness, the bottom row of plots in Figure 7 shows mean query lengths across time: as the recorded search sessions progressed, queries tended to become longer.

We hypothesize that $\texttt{FEEDBACK}_{\texttt{SC}}$, with its additional feedback to the participants, is counterproductive to their learning efforts due to *the effects of gamification*. That is to say, instead of focusing on learning, participants are focused on trying to *'fill up'* the progress bar. This leads to less self-reflection whilst reading documents as participants' focus is now on the progress of the scaffolding bar. Consequently, this causes a decrease in the learning gain.

To empirically evaluate this hypothesis, we can look at the average document dwell time (Table 3, row IX): it is on average 55 seconds in the $\texttt{FEEDBACK}_{\texttt{SC}}$ variant, which is significantly lower than that of the $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{AQE}_{\texttt{SC}}$ variants (with an average document dwell time of 92 seconds and 100 seconds, respectively). At the same time, $\texttt{FEEDBACK}_{\texttt{SC}}$ participants viewed on average the most documents, and the most document snippets (Table 3, rows X and XI). In addition, Figure 7 (middle row) shows that, as the search session progresses, participants from $\texttt{FEEDBACK}_{\texttt{SC}}$ tend to use more terms from the outlines than their $\texttt{CURATED}_{\texttt{SC}}$ counterparts.

To explain the large gap between the results of $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{FEEDBACK}_{\texttt{SC}}$, Swinnen et al. [39] in a psychology study showed that learners who are presented with frequent feedback on their learning progress tend to learn less than others that do not. It is hypothesized that this is because this frequent feedback may impair their ability to *reflect* on what they have learned. Similarly, Mayer et al. [25] corroborate these findings in the setting of multimedia learning, demonstrating that too much extra information can distract learners from their core learning material. We believe that a similar effect may be in play here.

6 CONCLUSIONS

In this work, we have explored three strategies to introduce instructional scaffolding into a web search system with the goal of improving a learner's knowledge gain during the search process. These strategies were: *(i)* automatic query rewriting ($\texttt{AQE}_{\texttt{SC}}$) which is agnostic to the search backend; *(ii)* a curated static topical outline ($\texttt{CURATED}_{\texttt{SC}}$); and *(iii)* a curated topical outline with instant feedback on the exploration of the topic space ($\texttt{FEEDBACK}_{\texttt{SC}}$).

We conducted a user study with 126 participants and aimed to answer the following research questions:

RQ1 Is scaffolding effective to increase learning outcomes?
RQ2 How does the introduction of scaffolding change behaviors?

Answering, **RQ1**, we do not find sufficient evidence to corroborate that any of the proposed methods significantly impacts learning outcome. However, we open a new research venue, showing that scaffolding significantly changes user behavior on a number of metrics. This is shown by our analysis answering **RQ2**, where we show that explicit scaffolding (namely $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{FEEDBACK}_{\texttt{SC}}$) significantly alters users behavior in a number of important search metrics, like dwell time, number of queries issued and number of clicks. This is important, and should lead to further investigation on how we can use this behavior difference to better support learners.

Additionally, we have speculated that the discrepancy in behavior between $\texttt{CURATED}_{\texttt{SC}}$ and $\texttt{FEEDBACK}_{\texttt{SC}}$, albeit not significant, may be due to a gamification effect: instead of focusing on the task at hand (learning), participants are more focused on making progress on filling up their progress bars, and in the process lose sight of their goal. This is corroborated by the difference in dwell time, as the $\texttt{FEEDBACK}_{\texttt{SC}}$ condition led participants to skim the documents more than in other conditions (i.e. that condition had the lowest document dwell time) while spending more time on the SERP (highest number of document snippets viewed). Finally, we found that participants in the two conditions receiving the topical outline submitted more queries with many more query terms matching the terms in the topical outline.

From these results, there are several lines of future work to follow. Firstly, a better scaffolding component is needed: what type of interface/feedback to learners respond to best? In order to make this approach deployable in practice, we need to be able to *automatically generate* hierarchical outlines for any learning-oriented information need instead of relying on manually curated outlines. Those outlines should preferably be personalized, depending on users' domain expertise and other user characteristics. While exploration into (non-personalized) automatic outline generation [52] is relatively new, it remains unclear whether such slightly noisy outlines are beneficial for users' learning outcomes. In addition, it remains to be seen to what extent the changes in user behavior hold across time (as for instance explored by Syed and Collins-Thompson [42]), and whether users remain engaged over time when a scaffolding component is permanently introduced on the search interface. Finally, we need to consider that we measured learning with a vocabulary knowledge task, which covers only the lowest cognitive levels of learning [1]. Is scaffolding beneficial for learners that face learning tasks that target higher cognitive levels of learning [18]?

REFERENCES

[1] L. Anderson-Inman and L. Zeitz. Computer-based concept mapping: Active studying for active learners. *The Computer Teacher.(Aug./Sept.)*, 1993.

[2] B.R. Belland. Instructional scaffolding: foundations and evolving definition. In *Instructional Scaffolding in STEM Education*, pages 17–53. Springer, 2017.

[3] 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.

[4] 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.

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

[6] A.L. Brown and A.S. Palincsar. Guided, cooperative learning and individual knowledge acquisition. Technical report, U. Illinois Urbana-Champaign, 1986.

[7] K. Collins-Thompson, P. Hansen, and C. Hauff. Search as Learning. *Dagstuhl Reports*, 7(2):135–162, 2017. ISSN 2192-5283.

[8] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 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*, pages 163–172. ACM, 2016.

[9] 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.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, volume 1, pages 4171–4186, 2019.

[11] L. Dietz, M. Verma, F. Radlinski, and N. Craswell. Trec complex answer retrieval overview. TREC, 2017.

[12] 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.

[13] 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.

[14] S.M. Glynn and F.J. Di Vesta. Outline and hierarchical organization as aids for study and retrieval. *J. Educational Psychology*, 69(2):89, 1977.

[15] 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.

[16] A. Hassan 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.

[17] J.R. Hill and M.J. Hannafin. Teaching and learning in digital environments: The resurgence of resource-based learning. *Educational Tech. Research & Development*, 49(3):37–52, 9 2001.

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

[19] D.R. Krathwohl and L.W. Anderson. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2009.

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

[21] 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.

[22] Y. Lu and I-H. Hsiao. Personalized information seeking assistant (pisa): from programming information seeking to learning. *IRJ*, 20(5):433–455, 2017.

[23] B. Luyt. The inclusivity of wikipedia and the drawing of expert boundaries: An examination of talk pages and reference lists. *JASIST*, 63(9):1868–1878, 2012.

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

[25] 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. Experimental Psychology: Applied*, 14(4):329, 2008.

[26] D.C. Merrill, B.J. Reiser, M. Ranney, and J.Gregory. Trafton. Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *J. Learning Sciences*, 2(3):277–305, 1992.

[27] 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.

[28] D. Nicholas, I. Rowlands, D. Clark, and P. Williams. Google generation ii: web behaviour experiments with the bbc. In *Aslib proceedings*, volume 63, pages 28–45, 2011.

[29] 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.

[30] G. Pardi, J. von Hoyer, P. Holtz, and Y. Kammerer. The role of cognitive abilities and time spent on texts and videos in a multimodal searching as learning task. In *Proc. 5th ACM CHIIR*, pages 378–382, 2020.

[31] S.R. Putra, F. Moraes, and C. Hauff. Searchx: Empowering collaborative search research. In *Proc. 41st ACM SIGIR*, pages 1265–1268, 2018.

[32] B. Rogoff. Adult assistance of children's learning. *The contexts of school-based literacy*, pages 27–40, 1986.

[33] I. Rowlands, D. Nicholas, P. Williams, P. Huntington, M. Fieldhouse, B. Gunter, R. Withey, H.R. Jamali, T. Dobrowolski, and C. Tenopir. The google generation: the information behaviour of the researcher of the future. In *Aslib proceedings*, volume 60, pages 290–310, 2008.

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

[35] N. Selwyn. An investigation of differences in undergraduates' academic use of the internet. *Active learning in higher education*, 9(1):11–22, 2008.

[36] P. Sharma and M. J. Hannafin. Scaffolding in technology-enhanced learning environments. *Interactive learning environments*, 15(1):27–46, 2007.

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

[38] C.L. Smith and S.Y. Rieh. Knowledge-context in search systems: Toward information-literate actions. In *Proc. 4th ACM CHIIR*, pages 55–62, 2019.

[39] S.P. Swinnen, R.A. Schmidt, D.E. Nicholson, and D.C. Shapiro. Information feedback for skill acquisition: Instantaneous knowledge of results degrades learning. *J. Experimental Psychology: Learning, Memory, & Cognition*, 16(4):706, 1990.

[40] R. Syed and K. Collins-Thompson. Optimizing search results for human learning goals. *IRJ*, 20(5):506–523, 2017.

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

[42] 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.

[43] 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.

[44] K. Umemoto, T. Yamamoto, and K. Tanaka. Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In *Proc. 39th ACM SIGIR*, pages 405–414, 2016.

[45] K. Urgo, J. Arguello, and R. Capra. Anderson and krathwohl's two-dimensional taxonomy applied to task creation and learning assessment. In *Proc. 5th ACM ICTIR*, pages 117–124, 2019.

[46] J. von Hoyer, G. Pardi, Y. Kammerer, and P. Holtz. Metacognitive judgments in searching as learning (sal) tasks: Insights on (mis-) calibration, multimedia usage, and confidence. In *Proc. 1st ACM SALMM*, pages 3–10, 2019.

[47] L.S. Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard Univ. Press, 1980.

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

[49] B.M. Wildemuth. The effects of domain knowledge on search tactic formulation. *JASIST*, 55(3):246–258, 2004.

[50] 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.

[51] 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.

[52] R. Zhang, J. Guo, Y. Fan, Y. Lan, and X. Cheng. Outline generation: Understanding the inherent content structure of documents. In *Proc 42th ACM SIGIR*, pages 745–754, 2019.