Contrasting search as a learning activity with instructor-designed learning

Felipe Moraes
Delft University of Technology
Delft, The Netherlands
f.moraes@tudelft.nl

Sindunuraga Rikarno Putra
Delft University of Technology
Delft, The Netherlands
sindunuragarikarnoputra@student.tudelft.nl

Claudia Hauff
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

ABSTRACT

The field of *Search as Learning* addresses questions surrounding *human learning* during the search process. Existing research has largely focused on observing how users with learning-oriented information needs behave and interact with search engines. What is not yet quantified is the extent to which search is a viable learning activity compared to instructor-designed learning. Can a search session be as effective as a lecture video—our instructor-designed learning artefact—for learning? To answer this question, we designed a user study that pits instructor-designed learning (a short high-quality video lecture as commonly found in online learning platforms) against three instances of search, specifically (i) single-user search, (ii) search as a support tool for instructor-designed learning, and, (iii) collaborative search. We measured the learning gains of 151 study participants in a vocabulary learning task and report three main results: (i) lecture video watching yields up to 24% higher learning gains than single-user search, (ii) collaborative search for learning does not lead to increased learning, and (iii) lecture video watching *supported by search* leads up to a 41% improvement in learning gains over instructor-designed learning without a subsequent search phase.

ACM Reference format:
Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. 2018. Contrasting search as a learning activity with instructor-designed learning. In *Proceedings of 2018 ACM Conference on Information and Knowledge Management, Torino, Italy, October 22–26, 2018 (CIKM '18),* 10 pages.
https://doi.org/10.1145/3269206.3271676

1 INTRODUCTION

Search as Learning is a research area within information retrieval that considers questions surrounding *human* learning during the search process: how much or how little do users learn while they search and in what ways can search technology be adapted and optimised for human learning? In his seminal paper, Marchionini [27] remarked on the importance and complexity of what he called *learning searches* (i.e. search activities for the purpose of human learning), a subset of exploratory search: "*Learning searches involve multiple iterations and return sets of objects that require cognitive processing and interpretation. These objects [...] often require the information seeker to spend time scanning/viewing, comparing, and making qualitative judgements.*" At the Second Strategic Workshop on Information Retrieval in 2012 [1] search as learning was recognised as an important future research direction that will help "*people achieve higher levels of learning through [...] more sophisticated, integrative and diverse search environments*". This call for research has been taken up in recent years by a number of researchers in several directions, including optimising retrieval algorithms for human learning [38, 39], observing how users currently make use of search engines for learning-oriented information needs [9, 19], developing metrics to measure the amount of learning taking place during the search process [46], and arguing for more reflective search behaviour ("slow search") in contrast to the current demands for an instant—and increasingly proactive—search experience [41].

Search and sensemaking is an intricate part of the learning process, and for many learners today synonymous with accessing and ingesting information through Web search engines [8, 30, 40]. At the same time, Web search engines are not built to support users in the type of complex searches often required in learning situations [21, 25, 27]. But what effect does this lack of a learning-focused Web search engine design have on the ability of users to learn compared to a setting where they are provided with high-quality learning materials? In this paper we set out to answer this question by *measuring* how effective searching to learn is compared to (i) learning from—in our experiment: high-quality video—materials specifically designed for the purpose of learning, (ii) learning from video materials in combination with search, and, (iii) searching together with a partner to learn (i.e. collaborative search for learning).

The aim of our work is to *quantify* to what extent search as a learning activity is a viable alternative to what we call *instructor-designed learning*, that is, learning materials designed and created specifically for the purpose of learning. As not for every possible topic specifically designed learning materials exist, it is important to understand what effect that has on one's ability to learn. In addition, we are also interested in understanding whether the lack of learning materials can be compensated in the search setting by the presence of a second learner that has the same learning intent (i.e. collaborative search for learning).

Our work is guided by the following research questions:

RQ1 How effective (with respect to learning outcome) is searching to learn compared to instructor-designed learning?
RQ2 How effective (with respect to learning outcome) is instructor-designed learning *supported by search* in comparison to just instructor-designed learning?

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271676

RQ3 How effective is pair-wise collaborative search compared to single-user search for learning?

Specifically, in this work we conducted a user study with 151 participants and measured *vocabulary learning*, a particular instance of human learning (similar in spirit to [38, 39]), across five search and instructor-designed learning conditions. As high-quality instructor-designed learning materials we make use of lecture videos sourced from TED-Ed, Khan Academy and edX, popular online learning platforms. Our main findings can be summarised as follows:

- we find participants in the instructor-designed learning condition (watching high-quality lecture videos) to have 24% higher learning gains than participants in the searching to learn condition;
- collaborative search as learning does not result in increased learning gains;
- the *combination* of instructor-designed learning and searching to learn leads to significantly higher learning gains (an increase of up to 41%) than the instructor-designed learning condition without a subsequent search phase.

2 RELATED WORK

We now provide an overview of the areas related to our work: exploratory search, search with an educational intent and collaborative search.

2.1 Exploratory search

Exploratory search tasks are often complex, open-ended and multi-faceted [45]. They tend to span several sessions and require next to finding, the analysis and evaluation of the retrieved information. Marchionini's overview of exploratory search challenges and opportunities [27] marked the beginning of a long series of related workshops and evaluation campaigns that continue to this day [7, 44]. Several works have characterised users' search behaviours in this setting. Recently, Athukorala et al. [4] investigated to what extent simple lookup tasks differ from exploratory search tasks with respect to easily measurable behaviours such as the initial query length, the time spent on analysing the first SERP, the scroll depth and task completion time. Later, Athukorala et al. [5] leveraged their positive findings (these tasks do indeed differ in several behaviours) and proposed a robust predictor that determines based on the first traces of a search session whether the session will end up being of an exploratory nature.

Besides analysing users' exploratory search behaviours, a number of studies have focused on developing user interfaces and algorithms to support complex information needs, e.g. [21, 25, 35]. Golovchinsky et al. [21] proposed several interface elements to better support multi-session search, with a heavy focus on visualising the query history and query patterns, while Ruotsalo et al. [35] presented an interactive intent modeling interface to simplify the process of moving the exploration into one direction or another. On the algorithmic side, Hassan et al. [25] explored an automated approach (based on query logs) towards decomposing complex search tasks into relevant subtasks, a step of the search process that, in current Web search engines, is largely left to the user.

Our work is in line with prior search behaviour observation studies: we create different learning conditions and then observe and analyse our participants' behaviours in a relatively common Web search setup. One particular type of exploratory search are learning searches [27], which in recent years have been explored under the search as learning heading [13] as we discuss next.

2.2 Search as Learning

Information scientists have observed that learners of all ages increasingly turn to search engines to support their learning [20, 30, 34]. At the same time, concerns have been raised about the lack of individuals' "*critical and analytical skills to assess the information they find on the Web* [34]."

Several works have explored data-driven methodologies to determine the impact of (developing) expertise on search behaviour [19, 43] and subsequently to exploit measurable behavioural traces (log traces, eye-tracking traces) as proxies of domain knowledge [12, 47]. Relying on users' log traces and features derived from them (e.g. query complexity, diversity of domains on the SERP, document display time) enables the use of a large user population (e.g. more than 700K search sessions in [19]); at the same time though, these heuristics can only be considered to be crude proxies of learning gain metrics (i.e. the difference between the knowledge at the end and the start of the search session) and they require large-scale log traces to overcome the variance of the user population. Instead of relying on search behaviour proxies, some works have measured learning directly through the explicit assessment (e.g. through multiple-choice tests, the writing of a summary) of domain knowledge before and after the search as learning session [14, 15, 38, 46]—this of course is only viable in a lab setting with a limited set of users. In this paper, we follow the latter line of prior works, conducting a user study and measuring learning gains by assessing our participants before and after the learning session.

The main setup of our study is inspired by [14, 38, 39]. Collins-Thompson et al. [14] conducted a user study to investigate whether certain search strategies (single-query, multi-query, and intrinsic-diversified search results) are conducive to learning. They measured learning outcomes via manually assessed open-ended questions as well as self-reports and found both to correlate highly. Syed et al. [38, 39] introduced a document ranking model optimised for learning (instead of relevance as standard ranking models) and showed it to be more beneficial than standard retrieval algorithms with respect to learning outcomes. This finding though is based on a rather artificial study setup: the study participants were provided with a fixed list of ranked documents (produced by variants of the document ranker) on a given topic that they were required to read, before answering knowledge assessment questions—the user study explicitly avoided the use of an actual search engine and the associated typical search behaviour (issuing several queries before clicking a document, skipping over documents in the ranked list, etc.). In the work we present here, we investigate a more realistic setup, with topics drawn from online learning platforms and search sessions that require our participants to search the Web as they would usually do. Importantly, we compare the effectiveness of learning not just within search variants but also with respect to instructor-designed learning material.

2.3 Collaborative Search

In addition to single-user search variants, we also explore collaborative search (i.e. multiple users collaborating in the search process) in our study. The inclusion of this variant stems from the fact that collaborative searches for highly complex information needs, as may be encountered during learning, can yield significantly better results with respect to material coverage and knowledge gain when conducted in collaboration [28, 36, 37].

A number of collaborative search systems have been proposed in the past [2, 6, 11, 17, 23, 29, 31], though few of those systems are still accessible and functioning today. They all have been designed with a number of goals in mind, the most essential ones being (i) *awareness* of each others' actions (e.g. through a shared query history), (ii) enabling the *division of labour* (e.g. through algorithmic approaches [36] or a chat to explicitly divide the work), and (iii) *knowledge sharing* so that the collaborators do not duplicate their work (e.g. through shared bookmarks).

Lastly we note that systems can support different types of collaborations. Golovchinsky et al. [22] identified four dimensions: *intent* (explicit or implicit collaboration), *depth* (algorithmic changes to support collaborative search vs. user interface changes), *concurrency* (synchronous vs. asynchronous) and *location* (remote vs. co-location). In our work, we designed our collaborative search system to be used in an explicit collaboration, with changes restricted to the user interface level and remote users collaborating in a synchronous manner. These choices are not only governed by our user study setup, but also the fact that those are the most common characteristics of existing collaborative search systems.

3 SEARCH SYSTEM DESIGN

We developed our search system `SearchX` [33] as an extension and update of the `pienapple search` (PS) framework [10]. It includes the following functionalities (novel additions with respect to PS are underlined), some of which are specifically geared towards search experiments with crowd-workers:

- Search back-end connects to the Bing API to serve high-quality Web search results;
- Bookmarking as interface element to enable easy marking and access to relevant material;
- Single-user search and synchronous collaborative search of two or more users (included collaborative interface elements are a chat, shared bookmarks and a shared query history);
- Search verticals (web, images, videos and news);
- Extensive logging for subsequent data analysis;
- Integrated diagnostic (pre/post) tests;
- Interactive step-by-step user interface guide;
- Crowd-worker compliance settings.

We implemented `SearchX` in JavaScript and based on `node.js` and `React`; it is open-sourced at http://felipemoraes.github.io/searchx. Figure 1 shows our system's user interface when used in the collaborative search setup.

4 EXPERIMENTAL DESIGN

We set up our study as a *vocabulary learning task* which requires study participants to recall and produce the meaning of domain-specific terms. This task enables us to measure the *learning gain*—the dependent variable in our study—effectively and efficiently as the difference between the vocabulary knowledge in a pre- and post-test. Importantly, this task can be executed within a short time frame—such as a single search session—permitting us to recruit crowd-workers for our study as also previously done in [38, 39]. Learning tasks with more cognitively complex activities such as *create* or *design*, in contrast, require longitudinal studies (e.g. [9]) and considerable more assessment efforts to judge the artefacts created during learning (e.g. summaries [46]).

We now describe how we selected the topics for our study, then discuss metrics to measure vocabulary learning and finally present the five different experimental conditions we evaluated in our work.

4.1 Search as Learning Topics

One particular setting where we envision search as learning to play an important role is online learning—video lectures are widespread today and a vital component of the increasingly popular Massive Open Online Courses (MOOCs). Choosing high-quality lecture videos on very specific topics that were designed (often by instructional designers in the case of MOOCs) for learning makes the search challenge hard—enabling us to get a realistic answer to our research questions. Initially we chose three large-scale sources of lecture video content: Khan Academy[1], edX[2] and TED-Ed[3]. From both TED-Ed and Khan Academy we selected ten of the most popular videos (more than half a million views each); on the edX platform we first selected ten of TU Delft's STEM MOOCs (hypothesising that those cover more difficult materials than some other types) at undergraduate level and then selected a lecture video from within the first two course weeks that was no longer than 15 minutes. The selected candidate videos cover a range of topics including *dystopia*, *stoicism*, *magnetism*, *photosynthesis*, *radioactive decay* and *climate change*. Two authors of this paper manually created a vocabulary list for each of the in total thirty selected videos—a term entered the vocabulary list if (i) it was mentioned in the video at least once and (ii) it does not frequently occur outside of the domain-specific context as judged by the two annoators. This resulted in vocabulary lists with a median size of 30 items (minimum 23, maximum 73).

As such large lists were not feasible to be used in our actual study, we filtered the videos and vocabulary items by their difficulty and only retained the ten videos and their respective ten *most difficult* vocabulary items. Here, we employed the *amount of unfamiliar terminology* in a video as a proxy of video difficulty. In order to ascertain the difficulty of the videos and vocabulary respectively we asked three staff members of our institute (all with a PhD in computer science) to label all of the vocabulary items with a score between 1 (akin to *unknown term*) and 4 (akin to *I know the meaning*)[4]. The labelers only received the vocabulary list, not

[1] https://www.khanacademy.org/
[2] https://www.edx.org/
[3] https://ed.ted.com/
[4] Concretely, we employed the Vocabulary Knowledge Scale as outlined in Section 4.2, but did not require our labelers to actually write down the meaning of the items identified as knowledge levels (3) or (4) due to the sheer size of the vocabulary list.

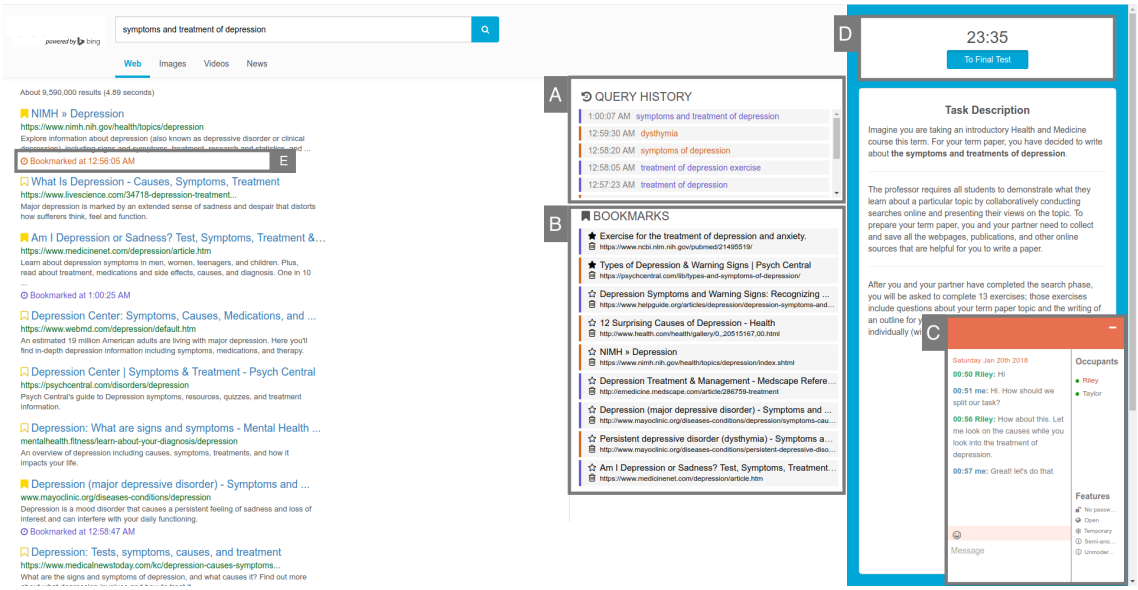

Figure 1: Overview of our search system user interface in the collaborative setup. Visible are next to the standard Web search interface the [A] query history widget, [B] bookmarking widget, [C] chat, [D] task timer and [E] time of bookmarking. In the single-user interface widgets [A], [C] and [E] are missing; the bookmarking widget is no longer shared among users. The two colour schemes in the collaborative widgets indicate which collaborator added the bookmark and/or query and at what time.

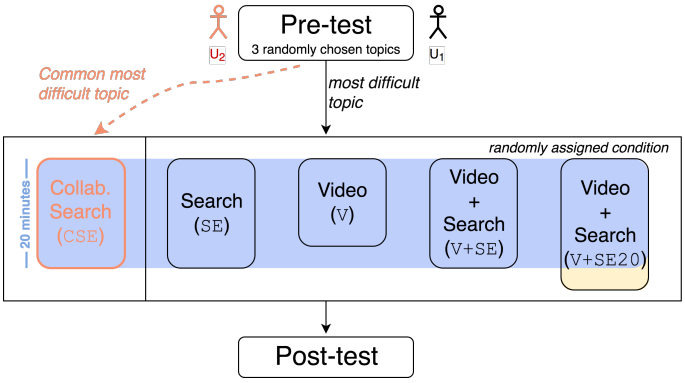

Figure 2: Study design overview: four single-user conditions and one collaborative (pairwise) condition.

the corresponding video. This resulted in a *vocabulary knowledge score* per video, which is simply the average score of all vocabulary items across the three labelers. We then ranked the videos according to their average score and selected the ten videos with the *lowest* scores, i.e. those with the largest amount of unknown terminology according to our labelers. Similarly, we also ranked each video's vocabulary items according to the average score across the three labelers and retained the ten least known ones. The final list of videos (identified by their topic), as well as a selection of the retained vocabulary items are listed in Table 2. The vocabulary items shown are highly domain-specific, a setup that contrasts with [38, 39] where participants' vocabulary knowledge was also tested on less domains-specific vocabulary such as "temperature" and "earth". The majority of videos in Table 2 are from edX; the average video length is 7.3 minutes, a common length of MOOC lecture videos [24].

This topic/video selection process ensures that our study participants are likely to find at least one of our topics unfamiliar with a high potential for vocabulary learning. That this is indeed the case, is visible in Figure 4—on average more than half of the tested vocabulary terms were unknown (knowledge levels 1 or 2) to the study participants.

4.2 Assessing Vocabulary Knowledge

We employ the *Vocabulary Knowledge Scale* (VKS) test [18, 42] as it has been shown to be a reliable indicator of vocabulary knowledge. The VKS tests the incremental stages of word learning [16] with the following statements:

(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 ___.*

We employ statements (1) to (4) to test the vocabulary knowledge for each of our vocabulary items[5]; the latter two statements require our study participants to recall and reproduce the meaning of the vocabulary item. Choosing statement (3) indicates uncertainty about the meaning's accuracy, statement (4) indicates certainty on the correctness of the provided meaning. In order to investigate to what extent this self-assessment is correct among the crowd-workers that participated in our study (§4.5 provides more information on them), we randomly sampled 100 of the meanings written by our

[5]Note, that the VKS test also contains a fifth statement geared towards second language learners. As in our study we only include native English speakers, we ignore it here.

participants across all vocabulary items—fifty from participants self-reporting levels (3) and (4) respectively. We manually labelled the statements as either *incorrect*[6], *partially correct*[7] or *correct*[8]. The results in Table 1 show that 88% of the statements self-assessed at knowledge level (4) are either correct or somewhat correct. At level (3), this holds for 68%. These results indicate that the self-assessment scores are robust and thus we use them without further manual labelling of the more than 3000 assessed vocabulary items. This is in line with the study conducted in [14], where users' perceived learning outcomes (i.e. the self-assessment) matched closely the actual learning outcomes (i.e. the produced definitions). Finally, it is worth pointing out that this setup is more difficult to tackle (as it requires the *production* of a definition) than closed multiple-choice questions (which require the *recognition* of a definition) as employed to test vocabulary learning in prior work [38, 39].

Table 1: Labelling of 100 sampled VKS level 3/4 statements.

	Correct	Partially Correct	Incorrect
VKS level 3	42%	26%	32%
VKS level 4	76%	12%	12%

4.3 Learning Metrics

As [39], we report *absolute learning gain* (*ALG*) and *realised potential learning* (*RPL*), enabling us to directly compare our study results to prior works. *ALG* is the aggregated difference in knowledge observed in the post- and pre-test across all vocabulary items $v_1, ...v_m$. Here, $vks^X(v_i)$ is the knowledge score assigned to v_i; X is the test (pre or post). As knowledge state changes from level (1) to (2) between pre- and post-test are natural (after the pre-test, each item has been seen at least once), we collapse the two lowest levels and assign both a score of 0. Items at knowledge levels (3) and (4) are treated in two ways: (i) in the binary setup we treat items at both levels in the same manner and assign a score of one; (ii) in the more fine-grained setup we assign scores of 1 and 2 respectively. The advantage of the binary setup is a more intuitive explanation of the *ALG*/*RPL* metrics as we will see later. We also assume that knowledge does not degrade between the pre- and post-test. *ALG* is then computed as follows:

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

The *RPL* metric normalises *ALG* by the maximum possible learning gain (*MLG*) of each item (either 1 in the binary case or 2 in the fine-grained setup):

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

$$RPL = \begin{cases} \frac{ALG}{MLG}, & \text{if } MLG > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

[6]Incorrect example: *superposition* (Qubit topic) described as "this has to do with the linear system".
[7]Partially correct example: *Bra* (Qubit topic) described as "vector".
[8]Correct example: *propofol* (Anesthesia topic) described as "an inhalation anesthetic used to induce sleepiness".

To compute the metric for a particular condition, we average the metric across all participants in that condition. We determine statistical significance through the non-parametric Kruskal-Wallis rank test which allows for the comparison of more than two groups.

4.4 Experimental Conditions

Figure 2 provides an overview of the study design we employed. Across all conditions, every participant first conducts a *pre-test* for which randomly three of our ten final topics are selected; for each of those topics all ten vocabulary items are assessed as described in §4.2. The participant is then assigned the topic $T_{difficult}$ for which she reported the lowest average knowledge levels. In case of a tie between topics, we randomly pick one. After that, the participant is randomly assigned to one of the conditions. The experiment ends with the post-test, in which the participant is again assessed on her vocabulary knowledge—this time only for items of $T_{difficult}$. In addition, the post-test also requires the participant to write a short summary on the topic as well as an outline[9]. In the collaborative search experiment, the two collaborators independently perform the pre-test and the post-test and collaborate during the collaborative search phase. For collaborating users we slightly extended the pre-test phase: we provided examples of collaborative searches and added seven questions on their past collaborative Web search experiences—taken from a large survey on collaborative Web search [28]—in order to reinforce the collaborative nature of the upcoming task.

Our participants are randomly assigned to one of five conditions:

Video (V) In this condition, a participant is given access to the lecture video and can watch it at her own pace (the common video player functions pause, rewind and skip are enabled).

Search (SE) Here, the participant is provided with the single-user search interface and instructed to search on the assigned topic *for at least 20 minutes*.

Video+Search (V+SE) The participant first views the video as in the **V** condition and afterwards is provided with the single-user search interface and asked to search on the assigned topic. The minimum time for this task (across both video watching and searching) is 20 minutes.

Video+Search (V+SE20) This condition is similar to **V+SE**, the only difference is that now the participant is instructed to spend 20 minutes searching after having viewed the video.

Collaborative Search (CSE) Two participants search together using the collaborative version of our search system for at least 20 minutes.

For all conditions involving a search phase, we employed the task template in Figure 3, adapted to our use case from previous studies [14, 26]. In the video-only condition (**V**) we instruct our participants to watch the video without any mention of search. Note, that the task description above does not explicitly state the nature of the post-test (ten of the thirteen "exercises" are our vocabulary learning questions), instead the focus is on acquiring on overview of the specific topic.

Apart from the video-only condition, all other conditions have a minimum task time; the participants are provided with a visible timer, and can complete the post-test as soon as the required time

[9]While collected, we leave the analyses of the outline and summary to future work.

Table 2: Overview of topics/conditions. Conditions: [SE] search only; [V] video only; [V+SE20] video followed by 20 minutes of search; [V+SE] video and search totalling at least 20 minutes; [CSE] collaborative search. The three right-most columns contain three examples (items at difficulty rank 1, 5 and 10) of the ten vocabulary items within a video.

	Participants per Condition						Video	Avg.	Vocabulary item difficulty rank		
Topic	SE	V	V+SE	V+SE20	CSE	Source	length	VKS	1 [most difficult]	5	10 [least difficult]
Radioactive decay	4	5	7	5	10	edX	6m53s	2.72	*Auger electron*	*K-shell electron*	*electron capture decay*
Qubit	5	3	5	2	2	edX	12m24s	2.81	*Ket*	*superposition*	*quantum information*
Water quality aspects	2	6	2	4	0	edX	10m45s	2.88	*trihalomethanes*	*bacteriophages*	*blue baby syndrome*
Religions	0	1	1	0	0	TEDEd	11m09s	2.91	*dharma*	*compendium*	*pilgrimage*
Sedimentary rocks	3	0	0	2	6	edX	5m03s	2.92	*feldspars*	*mud flats*	*sedimentary rocks*
Anesthesia	4	3	6	2	2	TEDEd	4m55s	2.94	*sevoflurane*	*diethyl ether*	*opium poppy*
Glycolysis	5	4	3	6	20	Khan	13m29s	2.97	*krebs cycle*	*electron transport chain*	*cellular respiration*
Urban water cycle	1	1	0	0	2	edX	7m40s	3.01	*Lesoto Highlands*	*coagulation*	*recontamination*
Depression	0	0	1	0	0	TEDEd	4m28s	3.02	*norepinephrine*	*transcranial*	*cholesterol*
Industrial biotech	2	2	0	4	8	edX	5m48s	3.02	*tobacco mosaic virus*	*prokaryotic*	*fungi*
#Participants total	26	25	25	25	50						

Imagine you are taking an introductory **[general topic, e.g. *Health and Medicine*]** course this term. For your term paper, you have decided to write about **[specific topic covered in the video e.g., *the symptoms and treatments of depression*]**.
The professor requires all students to watch a course video about a particular topic. Then, the students have to demonstrate what they learn about a particular topic by collaboratively conducting searches online and presenting their views on the topic. To prepare your term paper, you and your partner need to collect and save all the web-pages, publications, and other online sources that are helpful for you to write a paper. After you and your partner have completed watching the course video and the search phase, you will be asked to complete 13 exercises; those exercises include questions about your term paper topic and the writing of an outline for your term paper. Those exercises are solved individually (without your partner).

Figure 3: Task template for all conditions containing a search phase. The underlined green phrases were only added in the CSE condition; shown in dashed orange are the instructions only added for the V+SE and V+SE20 conditions.

on the task is reached. We settled on a twenty minute task time to provide participants with sufficient time to search and learn while keeping the study time feasible for crowd-workers. We added three compliance steps in our study design: (i) we included a *sports* topic (with well-known vocabulary items such as *football*, *winner*, etc.) in the pre-test and excluded workers who chose knowledge levels 1/2 here; (ii) we disabled copy & paste and recorded all tab changes in the pre- and post-tests and alerted participants to the fact that more than three tab changes lead to non-payment (to avoid participants searching the Web for answers to the questions); we limited the tab changes in the video watching period to three changes as well; and (iii) we required participants to adhere to a minimum word count in the open questions of the post-test.

We arrived at this design after a number of small pilot studies on the crowdsourcing platform CrowdFlower[10]. As CrowdFlower is mostly suitable for short tasks, we performed the actual experiments on the Prolific Academic platform[11], which has been shown to be a more reliable source of workers for cognitively demanding tasks than CrowdFlower [32].

4.5 Study Participants

Over the course of 27 days, a total of 151 study participants completed our experiment successfully across the five conditions on Prolific. Their median age was 31 (minimum: 18, maximum 66). 62.7% of our participants were female; most participants are from the UK (70.9%), the remaining participants are from Australia, the USA and Canada. Their academic backgrounds varied: 43.0% reported a high-school diploma as highest academic degree, 35% an undergraduate degree and the remaining 22% a graduate degree. We paid our study participants £5.00 per hour for the experiment. The median time they spent in our experiment (including the pre- and post-tests) was 49 minutes.

As the **CSE** condition is set up as a synchronous collaborative search task (i.e. two study participants have to be online at the same time), we added a waiting period for at most 10 minutes at the end of the pre-test; if within that time, no other participant completed the pre-test with the same topics, we released the participant from the task and paid £1.25 for a completion of the pre-test and the waiting period only.

Finally, we note that next to the 151 valid ones, we rejected 20 submissions—these participants did not adhere to our compliance standards (such as at most three tab changes). We continued the crowdsourcing task until we reached at least 25 participants/pairs for each condition. The relatively low number of rejections despite the complexity and length of the task indicates Prolific to be a suitable platform for this type of user study.

5 RESULTS

We now discuss the results, organised according to the three research questions.

5.1 Search vs. Instructor-Designed Learning

In **RQ1** we investigate whether search as learning is as *effective* as instructor designed learning, that is, as effective as watching the lecture video. We thus focus on comparing conditions **SE** and

[10] https://www.crowdflower.com/
[11] https://www.prolific.ac/

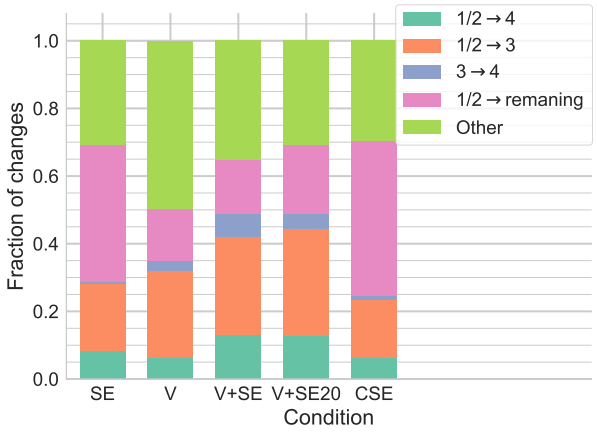

Figure 4: Overview of vocabulary knowledge state changes aggregated across all topics for each condition.

V. We computed the absolute learning gain and realised potential learning for each study participant in the binary setup[12]; the results, averaged across participants of a condition, are shown in Table 3. In the **V** condition, the average *ALG* is 0.32, that is, on average the participants increased their knowledge on three out of ten vocabulary items from knowledge levels 1/2 to levels 3/4. The interpretation of *RPL* is equally intuitive: for the **SE** condition for example, this metric is 0.3, indicating that on average the participants reached knowledge levels of 3/4 for thirty percent of the terms that were unknown to them.

When comparing **SE** with **V**, although we do not observe a statistically significant difference between the two (recall that the sample size overall is not very large), the results show a trend: instructor-designed learning leads to a 14% (measured in *ALG*) and a 24% (*RPL*) increase in learning gains respectively. Practically, the change in *RPL* from 0.3 to 0.37 means that participants in the **V** condition reached knowledge levels 3/4 for "almost" one more vocabulary item than participants in **SE**.

In Figure 4 we zoom in on the knowledge state changes between the pre- and post-test and report the fraction of the most important types of changes. Across all conditions, participants in the **SE** condition have the largest percentage (40.38%) of vocabulary items that remain at knowledge levels 1/2 in the post-test. As expected, in the **V** and the two video+search conditions **V+SE** and **V+SE20** this percentage is considerably lower (15.6%, 20.4% and 16% respectively), as all tested vocabulary items are mentioned in the video.

One expected difference between the **SE** and **V** conditions is the amount of time it takes to complete the task. As seen in Table 2 our selected videos have a length between five and fourteen minutes. As in any standard Web video player, our participants are free to pause, re-wind and skip ahead. In the **SE** condition, we require our participants to spend at least twenty minutes within our search system. We next examine the *actual* time spent on the respective interface(s). For **V**, this is the time difference between the first video-play event and the last video-stop event. For **SE** we consider this

[12]We found the same statistical differences in the more fine-grained setup (which distinguishes knowledge levels 3 and 4); here we report the binary case as it is more intuitive to interpret.

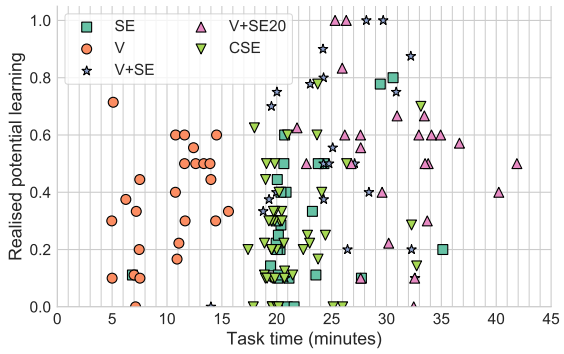

Figure 5: Active task time. Each point is a study participant.

to be the time difference between the end of the interactive guide through the search interface and the time of the last submitted query, viewed document or mouse hover over a snippet (whatever came last). For the mixed video and search conditions we add up the times spent on those two interfaces. Figure 5 shows the relation between *RPL* and the active task time. Participants in the video-only condition spent between five and sixteen minutes (median 10m55s), while most of our **SE** participants spent the required minimum amount of time *actively* searching. Since we pre-set a minimum task time we have to leave an investigation into the *minimum* time required to achieve the same knowledge gain as participants in the video-only condition for future work. What we can say though is that on average **SE** participants require ten more minutes to achieve a comparable learning gain. This is clearly an *upper bound* as we did not investigate a reduction in search time.

Lastly, in Table 4 we list some key characteristics of the search session behaviours across conditions. Participants' search behaviour in **SE**, **V+SE** and **V+SE20** was very similar—participants are unlikely to noticeably change their search behaviour considerably within a single session. Prior work [19] has shown that within-session learning is possible, however, a large user population is required in order to observe the small changes in behaviour reliably. Most of our participants submitted between six and ten queries during the search session (four example sessions of our participants are shown in Table 5) that were slightly longer (3-4 terms) than the typical Web search queries (2-3 terms). On average participants in the **SE** condition clicked on 10 links per session, mostly within the Web vertical. They also bookmarked slightly more documents than they clicked (on average 12.5) and spent on average nearly 7 minutes reading the clicked documents. These numbers indicate that our crowd-workers engaged with our search system and the task at hand as intended.

Table 3: Learning effectiveness metrics. Superscript ^{X} indicates a statistically significantly higher metric than condition X (Kruskal-Wallis, † p-value < 0.05, ‡ p-value < 0.01).

	SE	V	V+SE	V+SE20	CSE
ALG	0.281	0.320 $^{\text{CSE}\dagger}$	0.420 $^{\text{SE}\dagger\ \text{CSE}\ddagger}$	0.444 $^{\text{SE}\ddagger\ \text{V}\ddagger\ \text{CSE}\ddagger}$	0.234
RPL	0.296	0.368 $^{\text{CSE}\ddagger}$	0.501 $^{\text{SE}\dagger\ \text{CSE}\ddagger}$	0.518 $^{\text{SE}\ddagger\ \text{V}\dagger\ \text{CSE}\ddagger}$	0.254

Table 4: Basic search behaviour characteristics across the search conditions shown in Average (Standard Deviation). The clicks column lists the clicks on documents in the Web vertical as well as the aggregated clicks on all other verticals. For the Max. Clicked Rank column, we average the maximum click rank of each participant.

	Search Session Length [in minutes]	#Queries	Query Length [in words]	#Clicks Web vertical/ Other	Max. Clicked Rank	#Bookmarks	Reading Time [in minutes]
SE	21m58s (4m59s)	7.50 (6.12)	3.53 (2.14)	10.42 (6.96)/0.38 (1.16)	12.40 (11.67)	12.50 (8.03)	6m53s (5m08s)
V+SE	17m42s (11m13s)	8.16 (5.87)	3.12 (1.53)	12.44 (8.79)/0.56 (2.80)	10.04 (6.51)	10.12 (8.38)	7m23s (5m18s)
V+SE20	20m56s (2m36s)	7.48 (5.75)	3.31 (2.47)	13.20 (14.80)/0.68 (1.70)	10.78 (7.21)	14.72 (15.05)	6m37s (4m43s)
CSE	23m13s (12m52)	6.22 (3.86)	3.32 (2.06)	12.00 (6.89)/0.24 (0.55)	8.40 (8.57)	7.62 (7.88)	6m09s (3m32s)

Table 5: Four example search sessions logged in our experiment.

Water quality aspects	**Depression**	**Qubit**	**Anesthesia**
chemical processes relevant to ensure safe drinking water→threats to drinking water→chemical water treatment→safe drinking water supply→chemical water treatment	symptoms of depression→addressing depression→natural remedies for depression→natural remedies for depression websitetype:.org→natural remedies for depression .org→recognizing depresion symptoms→treatment for depression→causes of depression→supporting someone with depression	qubit→qubit transpose→quantum bits basics→quantum bits calculating length→quantum bits unitary vector→quantum bits amplitude→quantum bits amplitude vector→quantum bits basic state→quantum bits terminology→quantum bits calculating the transpose→quantum bits notation→quantum bits filetype:pdf	how anesthesia works → regional anesthesia → inhalational anesthesia → intravenous anesthesia → what does anesthesia do to your body

5.2 Instructor-Designed Learning with(out) Search Support

To address **RQ2**, we now explore whether a search phase immediately following the instructor-designed learning phase has a significant impact on the learning gain. The results in Table 3 indicate that this is indeed the case: both metrics *ALG* and *RPL* increase for **V+SE** and **V+SE20** compared to condition **V**. Both video and search conditions lead to significantly higher learning gains (absolute and potential) than the search-only condition; **V+SE20** significantly outperforms **V** as well. While in the video-only condition participants are able to increase their knowledge for slightly more than a third of vocabulary terms not known to them (RPL_V=0.37), in **V+SE** as well as **V+SE20** this is the case for more than half of the previously unknown (RPL_{V+SE}=0.5, RPL_{V+SE20}=0.52) vocabulary items.

Table 6: Location of vocabulary items (VIs)

VKS change	**Measure**	**SE**	**V+SE**	**V+SE20**	**CSE**
1/2 →1/2	%clicked docs with VIs	9.68	1.68	2.01	18.12
	%snippets with VIs	0.12	0.01	0.06	0.19
1/2 → 3	%clicked docs with VIs	7.89	7.31	6.78	14.16
	%snippets with VIs	0.08	0.12	0.18	0.25
1/2 → 4	%clicked docs with VIs	7.26	4.52	2.46	6.59
	%snippets with VIs	0.11	0.09	0.05	0.15

Figure 4 shows two interesting insights on the knowledge state changes: first, **V+SE**/**V+SE20** participants are more certain about their learning than participants in the video-only condition (with the number of vocabulary state changes from 1/2→ 4 doubling); secondly, participants are able to confirm their partial knowledge to a higher degree—with most knowledge state change transitions of the type 3 → 4 occurring in the **V+SE**/**V+SE20** conditions. With respect to time-on-task (Table 4) our **V+SE20** participants spent on average just four more minutes searching than our **V+SE** participants, despite the quite different minimum task times (for **V+SE** the total task time is set to 20 minutes, for **V+SE20** it is 20 minutes for just the search part); indeed Figure 5 shows that only a minority of participants quit the task immediately after reaching the minimum task time.

Lastly, we also consider to what extent the appearance of vocabulary items in the clicked documents and on the SERP (snippets) is indicative of knowledge state changes. This investigation is inspired by the observations reported in [19] where users were found to draw terms from the SERP and viewed documents to formulate subsequent queries within a search session. We here focus on the absence or presence of the tested vocabulary items within the viewed documents and the SERP and bin the vocabulary items according to the knowledge state change they underwent. The results are shown in Table 6. Although one might expect a particular trend (the more often a vocabulary item appears in the viewed documents, the higher the knowledge gain), there is actually none across the single-user search conditions; presence or absence of vocabulary items is not sufficient to approximate knowledge gains. Even more surprisingly, in the **CSE** condition we observe the opposite: for vocabulary items that remain largely unknown a larger percentage of documents contain those terms than for vocabulary items our participants increased their knowledge on. These two results point to the fact that a valid proxy of learning needs to measure much more than term absence/occurrence.

5.3 Collaborative Search As Learning

With **RQ3** we aim to explore whether our instance of a search as learning task can benefit from users collaborating together. Table 3 shows that in contrast to our hypothesis, collaborative search

does not lead to increased learning gains compared to the other conditions. On the contrary, we observe our collaborative search participants to perform significantly worse than participants in the video-only condition (*ALG* of 0.32 vs. 0.23, i.e. **CSE** participants learn one word less than **V** participants on average) as well as both video+search conditions. One explanation can be found in the fact that despite spending more time within our search system than participants in all other search conditions, time is spent on the collaboration process—time that is not spent reading documents, as evident in Table 5, where **CSE** participants have the lowest average reading time compared to all other conditions. The median number of chat messages collaborative pairs wrote was 13 (minimum 1, maximum 54) and of all clicks, 6.15% came from entries in the query history and bookmarking widget that their partners made. As backed up by prior works on collaborative search systems, collaborators are efficient at sharing the bookmarking work: the average number of bookmarks per participant is indeed the lowest in the **CSE** condition, as participants here have a partner to contribute bookmarks as well.

Figure 6 provides us with another interesting insight with respect to our participants' academic background (high school certificate, undergraduate degree and graduate degree); here, the spread of realised learning potential within each background and condition is shown. Participants with an undergraduate degree show consistently higher gains than participants with a high school certificate (across all conditions the median realised learning potential is higher for undergraduates). Surprisingly, participants with self-reported graduate degrees do not follow this trend consistently, they perform especially poorly in the **SE** condition. The participants within the **CSE** condition show very similar learning potential (more so than participants in other conditions), likely due to the fact that the pairing of participants was random, instead of being based on a shared academic level. We also find that the spread in realised learning potential is small, there are few positive outliers.

Given the comparably low learning gains in the collaborative search condition, we explored whether our participants experienced particular difficulties finding information during the task, a question we included in the post-test.

Table 7 shows the result of an open card-sort approach (here, we merged the two search+video conditions); two of the authors independently sorted the 126 open answers submitted for this question in the into groups, discussed differences and then created a composite of the two results. We found seven categories of difficulties. Surprisingly, in the **CSE** condition nearly half of the participants (48%) indicated to not have encountered any difficulties, a higher level of satisfaction with our search system than participants in the single-user search conditions (where 38% reported no issues).

Interestingly, while in the **CSE** and **V+SE(20)** conditions, searching for the right information was the most often reported difficulty (e.g. "*No difficulty in finding information, however most websites gave more of an overview around physical treatment methods and did not use technical language or use the terms used in the lecture.*"), in the **SE** condition, where participants neither had a lecture video as a basis, nor a partner to exchange information with, this issue was only mentioned 12% of the time. Thus, the participants that achieved the highest learning gains overall, self-reported the largest difficulties with the search phase.

Table 7: Overview of participants' self-reported difficulties.

Category	SE	V+SE/V+SE20	CSE
No problems reported	38%	38%	48%
Task setup	19%	2%	—
Unclear focus	12%	8%	—
Searching	12%	40%	32%
Sensemaking	8%	10%	12%
Credibility of sources	7%	—	8%
Attitude	4%	—	—
Video >> search	—	2%	—

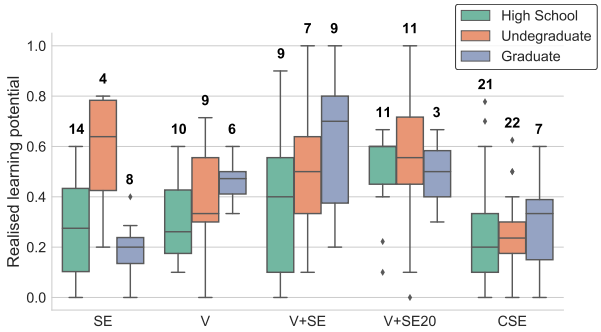

Figure 6: Relationship between highest academic degree and realised learning potential. The number of participants represented by each box is shown.

6 CONCLUSIONS

Is search a viable alternative to instructor-designed learning? This is the question we set out to answer. To this end, we carefully iterated over and designed a crowd-sourced user study with five different conditions that pitted instructor-designed learning (in the form of short high-quality video lecture as commonly found in online learning platforms) against three forms of search: (i) search-only, (ii) search as a support tool for instructor-designed learning and (iii) collaborative search. 151 participants spent a total of 138 hours in our experiment; they posed 897 queries and clicked on 1,512 documents. We measured their learning gains in a vocabulary learning task, which, though testing only lower cognitive skill levels [3], proved to be sufficiently challenging—not just for crowd-workers without a higher degree, but also for crowd-workers with a tertiary education (Figure 6).

We made a number of important findings: (1) participants in the instructor-designed learning condition reached up to 24% higher learning gains (measured in *RPL*) than participants in the search-only condition; (2) instructor-designed learning supported by search is superior to instructor-designed learning alone, leading to a 41% increase in realised potential learning; at the same time, these increases in learning gain do not translate into a higher confidence in the search process; (3) in our short-term learning task (approx. 20 minutes), collaborative search is not competitive as the collaborative overhead leaves less time for the retrieval and sensemaking steps of the search process: the learning gains decreased significantly compared to the video conditions. Considering that we pitted very high-quality lecture videos against search, we consider these

results as an indication that search can be a viable (though worse) alternative to instructor designed learning, especially in situations where no high-quality video material is available.

Our study has a number of limitations. Those are at the same time promising directions for future work: (i) so far, we have restricted ourselves to the vocabulary learning task—an open question is whether the findings are robust across a number of cognitive skill levels; (ii) with increased cognitive skill levels we also need to explore better sensemaking interface elements; (iii) the current study was designed to be completed by crowd-workers, which naturally restricts the possible task duration—a more longitudinal setup for instance with MOOC learners will enable a large-scale study with a larger number of conditions; and finally, (iv) we need to explore in detail the overhead of collaboration in the search as learning setting and designing interfaces to decrease the costs of collaboration.

ACKNOWLEDGEMENTS

This research has been supported by NWO projects LACrOSSE (612.001.605) and SearchX (639.022.722).

REFERENCES

[1] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012. In *ACM SIGIR Forum*. 2–32.
[2] Saleema Amershi and Meredith Ringel Morris. 2008. CoSearch: A System for Co-located Collaborative Web Search. In *CHI '08*. 1647–1656.
[3] Lorin W Anderson, David R Krathwohl, P Airasian, K Cruikshank, R Mayer, P Pintrich, James Raths, and M Wittrock. 2001. A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy. *New York. Longman Publishing. Artz, AF, & Armour-Thomas* (2001), 137–175.
[4] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. 2016. Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *JASIST* (2016), 2635–2651.
[5] Kumaripaba Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. 2016. Beyond relevance: Adapting exploration/exploitation in information retrieval. In *IUI '16*. 359–369.
[6] Peter Bailey, Liwei Chen, Scott Grosenick, Li Jiang, Yan Li, Paul Reinholdtsen, Charles Salada, Haidong Wang, and Sandy Wong. 2012. User task understanding: a web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Kanagawa, Japan*.
[7] Nicholas Belkin, Toine Bogers, Jaap Kamps, Diane Kelly, Marijn Koolen, and Emine Yilmaz. 2017. Second Workshop on Supporting Complex Search Tasks. In *CHIIR '17*. 433–435.
[8] J Patrick Biddix, Chung Joo Chung, and Han Woo Park. 2011. Convenience or credibility? A study of college student online research behaviors. *The Internet and Higher Education* (2011), 175–182.
[9] Marc Bron, Jasmijn Van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. 2013. Aggregated search interface preferences in multi-session search tasks. In *SIGIR '13*. 123–132.
[10] Martynas Buivys and Leif Azzopardi. 2016. Pienapple search: an integrated search interface to support finding, refinding and sharing. *ASIS&T* (2016), 1–5.
[11] Robert Capra, Annie T Chen, Katie Hawthorne, Jaime Arguello, Lee Shaw, and Gary Marchionini. 2012. Design and evaluation of a system to support collaborative search. *ASIS&T* (2012), 1–10.
[12] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *IP&M* (2013), 1075–1091.
[13] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as Learning (Dagstuhl Seminar 17092). *Dagstuhl Reports* (2017), 135–162.
[14] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *CHIIR '16*. 163–172.
[15] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: an eye-tracking study of information usage in web search. In *CHI '07*. 407–416.
[16] Edgar Dale. 1965. Vocabulary measurement: Techniques and major findings. *Elementary English* (1965), 895–948.
[17] Abdigani Diriye and Gene Golovchinsky. 2012. Querium: a session-based collaborative search system. In *ECIR '12*. 583–584.
[18] Katherine A Dougherty Stahl and Marco A Bravo. 2010. Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher* (2010), 566–578.
[19] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the journey: a query log analysis of within-session learning. In *WSDM '14*. 223–232.
[20] Urs Gasser, Sandra Cortesi, Momin M Malik, and Ashley Lee. 2012. Youth and digital media: From credibility to information quality. (2012).
[21] Gene Golovchinsky, Abdigani Diriye, and Tony Dunnigan. The future is in the past: designing for exploratory search. In *IIiX'2012*. 52–61.
[22] Gene Golovchinsky, Jeremy Pickens, and Maribeth Back. 2009. A Taxonomy of Collaboration in Online Information Seeking. *CoRR* (2009).
[23] Roberto González-Ibáñez and Chirag Shah. 2011. Coagmento: A system for supporting collaborative information seeking. *ASIS&T* (2011), 1–4.
[24] Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of mooc videos. In *L@S '14*. 41–50.
[25] Ahmed Hassan Awadallah, Ryen W White, Patrick Pantel, Susan T Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *CIKM '14*. 829–838.
[26] Bill Kules and Robert Capra. 2012. Influence of training and stage of search on gaze behavior in a library catalog faceted search interface. *JASIST* (2012), 114–138.
[27] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Comm. ACM* (2006), 41–46.
[28] Meredith Ringel Morris. 2013. Collaborative search revisited. In *CSCW '13*. 1181–1192.
[29] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: An Interface for Collaborative Web Search. In *UIST '07*. 3–12.
[30] David Nicholas, Ian Rowlands, David Clark, and Peter Williams. 2011. Google Generation II: web behaviour experiments with the BBC. In *ASLIB*. 28–45.
[31] Sharoda A. Paul and Meredith Ringel Morris. 2009. CoSense: Enhancing Sensemaking for Collaborative Web Search. In *CHI '09*. 1771–1780.
[32] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* (2017).
[33] Sindunuraga Rikarno Putra, Felipe Moraes, and Claudia Hauff. 2018. SearchX: Empowering Collaborative Search Research. In *SIGIR'18*. 1265–1268.
[34] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R Jamali, Tom Dobrowolski, and Carol Tenopir. The Google generation: the information behaviour of the researcher of the future. In *ASLIB*. 290–310.
[35] Tuukka Ruotsalo, Jaakko Peltonen, Manuel JA Eugster, Dorota Głowacka, Aki Reijonen, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. 2015. Scinet: Interactive intent modeling for information discovery. In *SIGIR '15*. 1043–1044.
[36] Chirag Shah, Jeremy Pickens, and Gene Golovchinsky. 2010. Role-based results redistribution for collaborative information retrieval. *IP&M* (2010), 773–781.
[37] Laure Soulier and Lynda Tamine. 2017. On the collaboration support in Information Retrieval. *Comput. Surveys* (2017).
[38] Rohail Syed and Kevyn Collins-Thompson. 2017. Optimizing search results for human learning goals. *IRJ* (2017), 506–523.
[39] Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval algorithms optimized for human learning. In *SIGIR '17*. 555–564.
[40] Arthur Taylor. 2012. A study of the information search behaviour of the millennial generation. *Information Research: An International Electronic Journal* (2012).
[41] Jaime Teevan, Kevyn Collins-Thompson, Ryen W White, Susan T Dumais, and Yubin Kim. 2013. Slow search: Information retrieval without time constraints. In *HCIR'13*. ACM, 1.
[42] Marjorie Wesche and T Sima Paribakht. 1996. Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *Canadian Modern Language Review* (1996), 13–40.
[43] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *WSDM '09*. 132–141.
[44] Ryen W White, Gheorghe Muresan, and Gary Marchionini. 2006. Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems. In *ACM SIGIR Forum*. 52–60.
[45] Ryen W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* (2009), 1–98.
[46] Mathew J Wilson and Max L Wilson. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *JASIST* (2013), 291–306.
[47] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. 2011. Predicting users' domain knowledge from search behaviors. In *SIGIR '11*. 1225–1226.