



Delft University of Technology

Designing Search-as-Learning Systems

Câmara, Arthur

DOI

[10.4233/uuid:0fe3a6bb-1bc1-40e2-86b0-ec3d3aef9c77](https://doi.org/10.4233/uuid:0fe3a6bb-1bc1-40e2-86b0-ec3d3aef9c77)

Publication date

2024

Document Version

Final published version

Citation (APA)

Câmara, A. (2024). *Designing Search-as-Learning Systems*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:0fe3a6bb-1bc1-40e2-86b0-ec3d3aef9c77>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

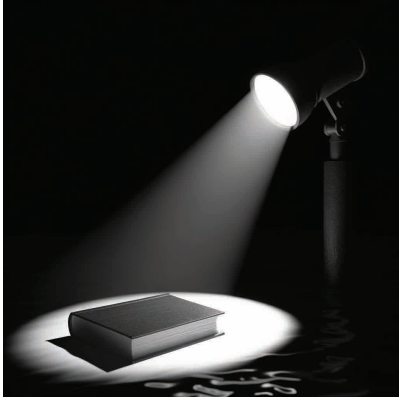
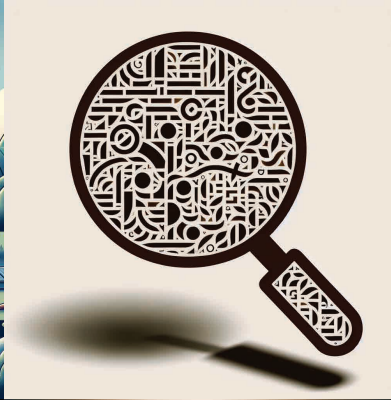
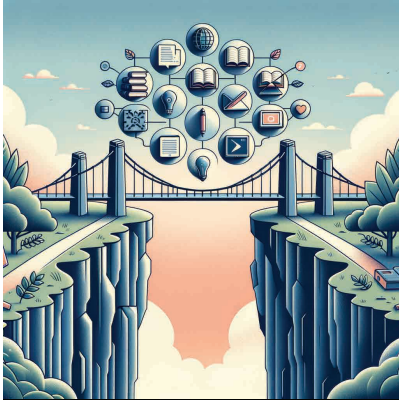
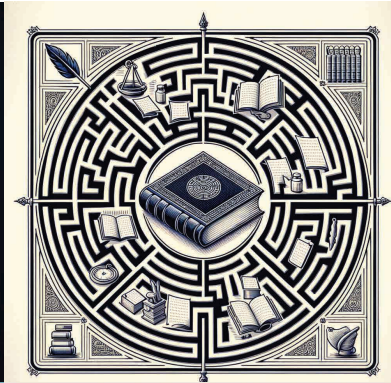
Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Designing Search-as-Learning Systems



Arthur Barbosa Câmara

Designing Search-as-Learning Systems

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on Wednesday 22 May 2024 at 15:00 o'clock

by

Arthur Barbosa CÂMARA

Mestre em Ciência da Computação
Universidade Federal de Minas Gerais, Brazil,
born in Belo Horizonte, Brazil.

This dissertation has been approved by the promotor

promotor: Prof. dr. ir. G.J.P.M Houben

promotor: Dr. C. Hauff

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. G.J.P.M Houben,	Delft University of Technology, promotor
Dr. C. Hauff,	Delft University of Technology, promotor

Independent members:

Prof. dr. ir. Alessandro Bozzon	Delft University of Technology
Prof. dr. ir. Arjen P. de Vries	Radboud University
Prof. dr. Jaime Arguello	University of North Carolina at Chapel Hill, USA
Dr. Julian Urbano	Delft University of Technology
Dr. Matthijs Spaan	Delft University of Technology, reserve member

SIKS Dissertation Series No. 2024–16

The research in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems. This research has been supported by NWO projects SearchX (639.022.722) and NWO Aspasia (015.013.027).



Keywords: search-as-learning, information retrieval, computer-human interaction

Printed by:

Cover: DALL-E 3 and GPT-4, prompted with the summary of this thesis

Style: TU Delft House Style, with modifications by Moritz Beller
<https://github.com/Inventitech/phd-thesis-template>

ISBN 978-94-6384-569-4

Instruct the wise and they will be wiser still; teach the righteous and they will add to their learning (...) If you are wise, your wisdom will reward you; if you are a mocker, you alone will suffer.

Proverbs 9:9 (NIV)

Contents

Acknowledgments	vii
1 Introduction	1
1.1 Search-as-Learning.	2
1.1.1 Front-End Interventions	4
1.1.2 Back-End Interventions	5
1.1.3 Measuring Learner’s Knowledge	6
1.1.4 Predicting Learning from Behavior.	7
1.2 Goals and Research Questions	9
1.3 Contributions	9
1.4 Thesis Origins	10
2 Searching to Learn with Instructional Scaffolding	11
2.1 Introduction	12
2.2 Related Work.	13
2.3 Instructional Scaffolding in SearchX	13
2.3.1 Topical Outlines	14
2.3.2 Variant AQE _{SC}	15
2.3.3 Variant CURATED _{SC}	15
2.3.4 Variant FEEDBACK _{SC}	16
2.4 User Study Setup.	17
2.4.1 Topics	17
2.4.2 Metrics.	17
2.4.3 Study Workflow	20
2.4.4 Study Participants	21
2.5 Results	21
2.5.1 RQ1: Impact of Scaffolding on Learning	22
2.5.2 RQ2: Search Behavior Analyses	24
2.6 Conclusions	26
3 RULK: A Framework for Representing User Knowledge in Search-as-Learning	29
3.1 Introduction and Related Work.	30
3.2 The RULK Framework	31
3.3 Implementing and Validating RULK	33
3.3.1 Implementing RULK.	35
3.4 Results	37
3.5 Caveats and Limitations	40
3.6 Conclusions	42

4	Searching, Learning, and Subtopic Ordering: A Simulation-based Analysis	43
4.1	Introduction	44
4.2	Related Work.	45
4.3	The Subtopic-Aware Complex Searcher Model (SACSM)	46
4.4	Experimental Method	48
4.4.1	Fixed SACSM Components.	48
4.4.2	Variable SACSM Components	49
4.4.3	Simulation Setup.	51
4.5	Results	51
4.6	Conclusions	54
5	Keep KALM and Search On: Unveiling Causal Connections in Search-as-Learning	57
5.1	Introduction	58
5.2	Related Work.	60
5.3	A Causal Model for Knowledge Acquisition	65
5.3.1	Structural Equation Modeling	65
5.3.2	Structural Model	66
5.3.3	Measurement Model	68
5.4	Datasets	68
5.5	Assessing KALM.	70
5.5.1	Assessing the Reflective Measurement Models	72
5.5.2	Assessing the Formative Measurement Models	74
5.5.3	Assessing the Structural Model.	78
5.6	Discussion	81
5.6.1	On Learning Metrics	82
5.6.2	Causal Relationships	84
5.7	Conclusion.	87
6	Conclusion	89
6.1	Summary of Findings.	89
6.1.1	Impacting Learner's Behavior and Knowledge Acquisition	89
6.1.2	Modeling Learner's Knowledge and Behavior	90
6.1.3	Explaining and Predicting Learner's Knowledge Gains	91
6.2	Ethical and Societal Implications	92
6.3	Moving Forward	92
	Bibliography	97
	Summary	125
	Samenvatting	127
	Glossary	128
	Curriculum Vitæ	131

Acknowledgments

It has been a wild journey. Through a global pandemic, lockdowns and two kids, I could not have done this without the support of many, many people that helped me, directly or indirectly, during these years. As a christian, I must recognize the many, many times, in both small and big situations, that my prayers were answered and I have found comfort in the roughest patches.

I am also extremely grateful for the woman who said yes to an out-of-the-blue invite of “Let’s move to the Netherlands”? My wife La’is has been the firm foundation where I know I can rely for anything for the last (almost) decade. As we use to say, “It is going to work out. Whatever that means”. I guess it did worked out, right? From this adventure, two amazing kids were born, Oliver and Logan. I cannot express how much joy and love their hugs and kisses bring to my life.

Even from a distance, I can still feel the cheering and support from my parents, Kleber and Neila, that provided me with the best foundation I could have asked for, and my brother Ivan and sister-in-law, Raquel, that are always there for me. Finally, my friends, Guilherme, Pedro, Guilherme, Filipe, Paula, Mari, Jacque and Deborah, that have been there for me, even when I was not always there for them.

Of course, the mentorship and guidance of my supervisors, Claudia Hauff and Geert-Jan was invaluable in these years. I am grateful for them for making me the independent researcher I am today. In special, I am specially grateful for Claudia’s guidance, support and patience during all these years. Thanks to her critical feedbacks and constant support, I have grown as a researcher, and I am a better person today.

We are lucky to have such a great group of colleagues at WIS: Alessandro, Marcus, Nava, Avishek, Sole, Asterios, Cristoph, Jie, Ujwal, Tim, Lixia, Agathe, Alisa, Andra, Andrea, Christos, Daphne, Dimitrios, Gaole, Garrett, Georgios, Guanliang, Petros, Kyriakos, Lorenzo, Manuel, Sara, Sepideh, Sihang, Shabnam, Shahin and Ziyu. A special thanks to (now, former) lambda lab, David, Nirmal, Felipe, Gustavo and Peide for the fun times and the great discussions.

Finally, this Ph.D. would not have been the same without other collaborators, like Dima, for her commitment to the RULK framework, and the great people I met during my internships at Bloomberg and Naver Labs: Edgar, Saher, and the whole Knowledge Graph team at Bloomberg, and Stéphane, Thibault, Carlos and Thibaut and the whole Search and Recommendation team at Naver Labs.

*Arthur Câmara
The Hague, March 2024*

1

Introduction

Learning is a fundamental facet of human life, being one of the essential attributes of human developmental psychology [1]. The famous Maslow’s hierarchy of needs, for instance, states that, after physiological and social needs are satisfied, “self-actualization” (i.e., education and learning) follows [2]¹. Whether driven by our innate curiosity or guided by a teacher, learning is how we absorb information, and learning shapes our behavior and understanding of the world.

In most contexts, learning involves discovering knowledge previously recorded and shared by someone else. This practice of externalizing and disseminating knowledge can be traced back to cave paintings [4], and it evolved into the extensive collections of documents and libraries from the ancient world [5]. With the advent of computers, massive collections of knowledge became digital, and the methods for searching this information in databases evolved as well, from simpler, statistics-based methods [6] to complex methods using Transformers-based neural-networks models [7].

With the rise of the Web and its unprecedented scale as a repository of human knowledge, access to knowledge has been democratized on an unforeseen scale in dynamic and interactive ways. Knowledge that was once confined to private collections or closed computer networks can now be (mostly) freely accessed by anyone with an Internet connection. However, this abundance of online information brings its challenges. Learners must now find relevant and suitable information that fits their initial information needs and adapt to their evolving knowledge levels [8, 9].

The field of Information Retrieval (IR) have at its core the problem of finding information that satisfies a user’s information need, such as a *novel knowledge*. Generally, user interactions with IR systems involve a user translating their information needs into a **query**, a short piece of text—usually in natural language—that is submitted to a search engine. The engine then *retrieves* and *ranks* a set of documents from within a large collection, such as the Web, that it estimates to be relevant to the user’s need [10].

Traditional IR systems are generally optimized for *ad-hoc* retrieval. In this scenario, given a user query, the search system retrieves and ranks documents according to their

¹Curiously, Maslow’s pyramid, one of the most commonly used visualizations of his theory, was never used by Maslow himself [3].

relevance to that single query [11]. However, in a learning setting, the user’s information need is unlikely to be satisfied with a single query. Rather, their interaction with the search system spans longer, encompassing multiple, distinct queries and documents [12, 13].

The knowledge acquisition process is intricate and usually requires multiple interactions with the search system [9, 14]. Users (or *learners* in this context) submit queries, consume information, reflect on their newly acquired knowledge, and repeat this process until they reach a satisfactory understanding of the topic [15].

This disparity between search systems optimized for single interactions and the multifaceted search scenarios required by learners has given rise to the field of Search-as-Learning (SAL), where research focuses on studying user behavior whilst using search systems for learning and in developing techniques and methods to support learners during their knowledge acquisition journey. As a subfield of Interactive Information Retrieval (IIR), that studies how learners interact with a search system interactively, beyond a single query, SAL has gained considerable traction in the early 2000s [8, 9, 16–19] and experienced a surge in interest in recent years [11, 13, 14, 20–43].

1.1 Search-as-Learning

Consider a “regular” Web user curious about a specific topic. For example, they may have heard about *Radiocarbon Dating* and want to know its reliability. The user then translates their information need (“I have heard about dating with radiocarbon, and I am curious about its reliability”) into a query that is typed into the search bar of a search engine: “radiocarbon dating reliability”. More likely than not, reading a single document would be enough for that user to satisfy their information need. Figure 1.1 shows an overview of this process.

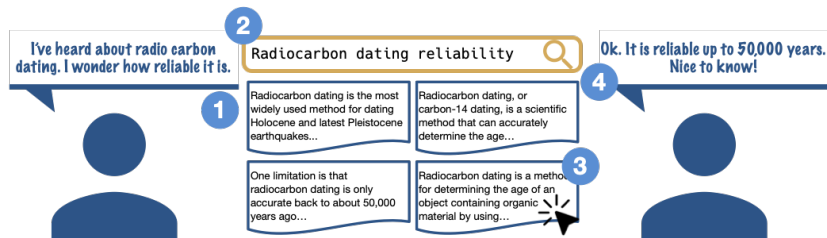


Figure 1.1: The process a traditional user of a search engine may take from translating an information need ① into a query and submitting it to a search engine ②, analyzing and clicking in a retrieved document ③, and finally being satisfied by the information found ④.

On the other hand, consider a *learner*, that is, a user seeking in-depth knowledge about the same topic. This user’s information need extends beyond a fleeting curiosity, encompassing a broader range of themes under the same subject and a deeper understanding of them. While their initial query may be the same (“radiocarbon dating reliability”), they are likely to visit more documents, spend time reflecting on the knowledge acquired,

and formulate subsequent queries based on this new (and evolving) understanding. This scenario is depicted in Figure 1.2.

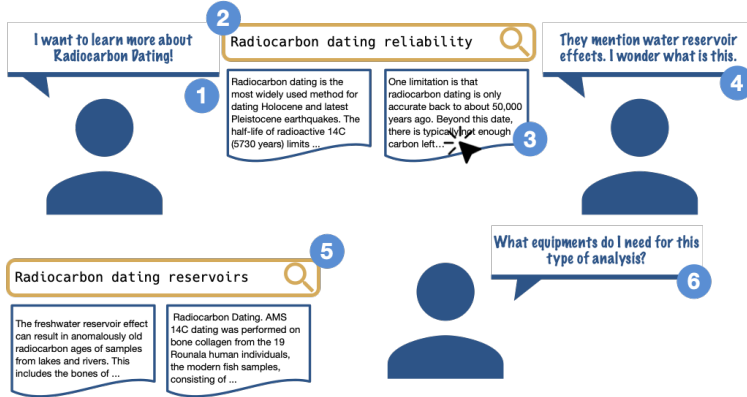


Figure 1.2: The process a *learner* may take from translating an information need **1** into a query, submitting that query to the search system **2**, clicking and reading a retrieved document **3**, processing the newly acquired information **4**, formulating new queries based on this new knowledge **5** and switching to another subtopic **6**.

In these examples, the regular user benefits from the fact that most search engines are optimized for single-query scenarios. Given a user query and minimal additional signals (such as the user's location [44]), these systems attempt to infer the user's knowledge need from that short text and rank documents to maximize the relevance of the top few results in the ranking to that single query.

The learner, on the other hand, typically engages with the search system at a more deliberate pace. Learners search for extended periods, submitting more queries and reading more documents than in other situations [9]. Therefore, Search sessions from learners differ from those of traditional search engine users in several key aspects:

- Learners are driven by tasks of higher cognitive levels [45, 46], such as analyzing and evaluating concepts and facts, rather than simply finding information such as a website or a phone number [47].
- Learners' sessions are longer, involving more queries submitted and more documents read over longer periods of time [20, 48].
- Consequently, the information needs of learners may change as they acquire new knowledge and evolve their own perspectives [49–51].

Therefore, the primary distinction between a traditional and a search-oriented search system lies in how they deal with the concept of a *search session*. While a regular searcher's session is usually short and with only one or two queries [12], a search system that is aware of the learner's session must keep track of the learner's exploration over extended periods,

encompassing sessions with longer reading times per document, multiple queries issued, and multiple read documents [13].

Works on SAL are usually interested in interventions made to the search system itself (e.g., improving the retrieval algorithm or the user interface) and what impact these interventions have in the learner’s behavior and knowledge acquisition process, usually displaying a complex interplay between the learner, the search system, and the knowledge the learner is trying to acquire.

Despite this complex interaction between multiple actors, some broad themes can be observed across multiple SAL studies. Generally speaking, all studies are interested in changes to the learner’s knowledge due to their searching behavior. A common theme is to use the learners’ behavior to infer their knowledge gains at either the *end* [52–56] or *during* their search session [34, 43, 57–60].

Another recurrent theme is to propose interventions to the search system itself, either in the *front-end* (i.e., the user interface) [61–66] or in the *back-end* (i.e., the retrieval and ranking algorithms) [29, 31, 67–69] and measure the impact that these interventions have in the learner’s behavior and knowledge acquisition process. In the following sections, we present these themes in more detail and discuss trends in the SAL literature.

1.1.1 Front-End Interventions

When using a search engine, users primarily interact with two distinct interfaces. The first is the search engine’s main page, where users submit their initial query. The second is the search engine results page (SERP), where the search engine displays its retrieved documents, reversely sorted by their estimated relevance and users interact with these documents. Since Google became the main Web search engine in Western countries in the early 2000s, the main page for most search engines has traditionally been simple, featuring little more than an empty white box and a button to submit a query². However, Web search engines’ SERP has evolved drastically recently. It has transitioned from the traditional “ten blue links” to a complex interweave of organic and sponsored results, entity cards, related and suggested queries, and multimodal features like videos and images [70].

Studying how different interventions in the front-end of a search engine can impact a learner’s behavior is a vital area of investigation in SAL. For instance, researchers have explored how specific novel features from “general-purpose” search engines, such as entity cards [41], non-linear search interfaces [71], query suggestions [72], and multimodal features [37], influence learning.

Particularly intriguing are front-end interventions tailored specifically towards the learning experience. While some changes might be replicated in general-purpose search engines, they are often designed with objectives like enhancing exploration, aiding recall of facts found by the learner, and assisting the learner in organizing their new knowledge. For example, progress bars tracking learners’ exploration of a subject have been experimented with by Umemoto et al. [61].

Methods such as highlighting [64, 73] and tagging documents [63] have been explored to improve the learner’s recall. For assisting learners in organizing their new knowledge,

²With the advance of conversational search and LLMs, however, this is slowly changing toward a more chat-like interface.

techniques like note-taking [64, 65] and mind-mapping [66] have proven helpful in structuring both search sessions and acquired knowledge.

It's notable that while studies that propose these types of changes usually display significant differences in user behavior, they exhibit mixed success in enhancing learners' knowledge acquisition. Some studies have shown improved learning-related metrics, such as knowledge retention [63, 65, 74]. In contrast, perhaps due to their small number of participants, other studies have not shown small or non-significant improvements in learners' knowledge acquisition [62, 64, 65]. Explaining why some interventions perform better than others is a multifaceted issue. A commonly discussed reason is the potential information overload caused by including excessive additional information, which may hinder navigation, especially when dealing with unfamiliar topics.

In this thesis, we discuss these topics in Chapter 2, proposing multiple front-end interventions to a search system and exploring their impact in a learner's search session.

1.1.2 Back-End Interventions

While the front-end of a search engine provides the user with an interface to submit queries and interact with retrieved documents, the back-end serves as the backbone of the search system, where documents are matched and ranked according to their relevance to the users' queries.

The information need of a learner is markedly distinct from the information need of a "regular user" [46, 75–77]. Therefore, it is crucial that the *back-end* of the search system (i.e., the retrieval and ranking algorithm themselves) is optimized for the former so that the search system can better support the learner's information needs.

In a SAL scenario, the relevance of a given document no longer depends only on the terms from the user query. While this is the hypothesis considered in traditional keyword-based retrieval methods like BM25 [78], a learning-oriented search system should also take into consideration factors specific to the learner (e.g., their previous and current knowledge on the topic), the topic being explored (e.g., its intrinsic complexity), to the documents being ranked (e.g., their textual content and complexity), and to the search session itself (e.g., the documents previously read by the learner and content of previous queries).

Picture a learner at the beginning of their search session with little to no knowledge about their learning topic. In this case, the learner's initial information need will likely be unclear, and their queries will likely contain broad and generic terms related to the topic.

Following our example of a learner interested in "radiocarbon dating", if, towards the beginning of their session, they submit a query "issues with radiocarbon dating", a document with the title "reservoir effects in radiocarbon dating", will, more likely than not, be too complex for them. However, later in their session, this same learner may consider that same document highly relevant, as their knowledge has evolved enough that the concepts covered and terms used in that document are now more familiar. In this setting, an ideal SAL system should be aware of the evolving knowledge state of the learner and adapt its ranking models accordingly.

Hence, relying uniquely on the document's content and the query without regard to the learner's current knowledge state is likely to result in suboptimal rankings. Most general-purpose search systems employ keyword-based retrieval methods, such as BM25, or semantic matching models, based on deep neural networks such as BERT [79, 80].

Therefore, in the context of SAL, the issue of estimating the relevance of a document to a learner’s information need is not only a matter of estimating the similarity between the document and the query but also a matter of estimating the learner’s current knowledge state and how it evolves.

1.1.3 Measuring Learner’s Knowledge

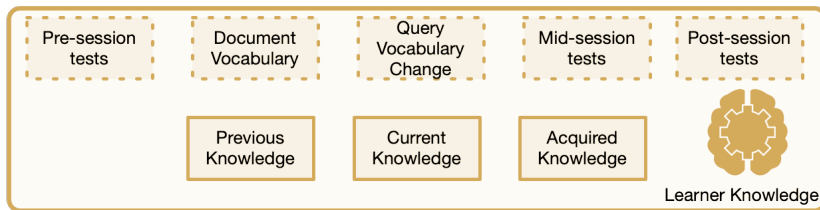


Figure 1.3: Tracking and predicting a learner’s knowledge state requires inferring hidden variables, such as their knowledge level before and after their search session (depicted as solid boxes), using proxy measurable variables, such as their scores in questionnaires and changes in their querying vocabulary (depicted as dashed boxes).

As SAL systems become more and more studied, much attention has been paid to how we measure learning, which is far from trivial [81]. With proper measurement of the learner’s knowledge, we can understand the true impact of proposed interventions on the search system, allowing for a better understanding of types of interventions that are more (or less) effective regarding learning outcomes [42]³. Additionally, if one can measure knowledge *during* a search session, it would enable search systems to dynamically adapt to the learner’s current knowledge state, providing them with a more personalized experience and search results more tailored to their current knowledge level.

In most SAL studies, the first hurdle to be overcome is how to *measure* learner’s knowledge. Usually, we rely on some pre- and post-session tests to measure the learner’s knowledge. Any difference between the learner’s scores in these two tests is attributed to their search behavior, such as the documents they found during their session. The difference between the two scores is considered the learner’s learning (or *knowledge gain*) during the session. One example of a test used to measure if a learner has learned about specific terminology of their topic of interest is the vocabulary knowledge scale (VKS) questionnaire [82]. In this questionnaire, learners are asked to rate their familiarity with a set of terms, from 1 (“*I don’t remember having seen this term/phase before*”) until 4 (“*I know this term/phrase*”) [62, 64, 83].

While useful and cheap to compute, multiple-choice questionnaires (such as the VKS) are mostly useful for lower levels of cognitive learning tasks [24, 46]. If we are interested in higher levels of learning, such as analysis and evaluation of information, other types of tests are needed. Some approaches include free-recall tests, where the learner is asked to write down as many facts as they can remember about the topic [34, 73], asking the learner to answer open-ended questions with short texts [28, 59, 62], the writing of extended summaries of the topic [28, 84] and even the evaluation of learners’ mindmaps [35, 58].

³For the sake of simplicity, we use the term learning, or knowledge gain, to describe a change in knowledge by the learner

One issue with evaluating written texts is that, while useful for laboratory or smaller-scale studies, they are unfeasible for larger-scale studies or real-world applications. Evaluating essays and other long-form texts is a time-consuming task, requiring experts in the topic and not trivially automated, making it an expensive and impractical approach⁴.

Therefore, one approach is to rely on *proxy* measures hypothesized to correlate to the learner's knowledge. Such measurements must be *easy* to capture so that estimating the learner's knowledge becomes computationally inexpensive, making it possible to be computed in real-time, and *observable* so that they can be measured implicitly without disrupting the learner's search session. Some of these indicators are illustrated in Figure 1.3.

Regardless of the usage of such proxy indicators, measuring the difference in learners' knowledge before and after their search session leaves a considerable blank space: what happens to the learner's knowledge *during* their search session? Simply prompting learners during their search sessions with the same questions used in the pre-and post-tests is rarely viable, as it would disrupt the learner's search session and likely impact their behavior [59]. Nevertheless, measuring learner's knowledge during their search session is crucial to understanding how it evolves, especially if we are interested in adapting aspects of the search system to dynamically adapt to the learner's knowledge state [57].

Here, one approach gaining traction is to automatically estimate the learner's knowledge state using variables such as the content of documents read during the search session and changes in the vocabulary of the queries submitted [21, 31, 43, 57, 58, 85]. In the setting, two challenges appear. First, it is unclear how to estimate how much novel knowledge the learner has acquired, given noisy signals such as the vocabulary of a document and a query. Second, assuming a reliable estimation of the learner's knowledge, how to use this knowledge to dynamically adapt the search system to the learner's knowledge state.

We discuss some of these challenges in depth in this thesis, especially in Chapter 3, where we propose a framework designed specifically for tracking the evolution of learners' knowledge, and in Chapter 4, where we simulate a learner's behavior while interacting with a search system where we keep track of their knowledge state with an evolving probabilistic language model.

Regarding the measurement of learners' knowledge gains, we use the VKS model as the primary evaluation metric in Chapters 3 and 2. We also discuss the reliability of the proxy indicators of learning and how they are causally linked to other more direct observable measurements of the learner's behavior in Chapter 5.

1.1.4 Predicting Learning from Behavior

Another common theme across SAL is to, instead of measuring (or estimating) the knowledge of the learner, try to *predict* how much *learning* (i.e., increase in knowledge) occurred during their search session. By developing an accurate method to predict the learner's knowledge gains, we can gain a deeper understanding of what behavioral (e.g., the learner's attention and effort) and system (e.g., retrieved document quality and complexity) variables have a larger influence in the learner's knowledge acquisition process.

As discussed in Section 1.1.3, we mostly rely on proxy measurements correlated to learners' knowledge when interested in measuring their actual knowledge at a given point of their search session. The reasoning is that, due to the complexity of the learning process,

⁴With the advance of LLMs models, this could change in the near future

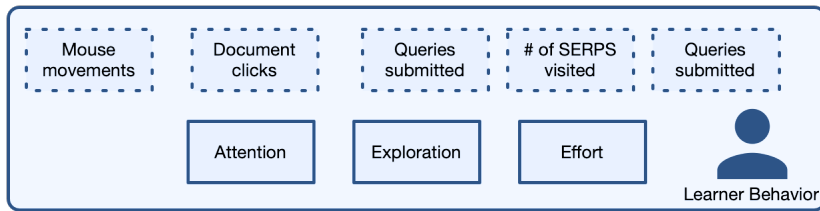


Figure 1.4: Measuring and predicting a learner’s behavior while searching requires inferring the value of latent variables related to their internal state (depicted as solid boxes) using proxy observable variables captured by the search system (depicted as dashed boxes).

directly measuring knowledge is too expensive, slow, and disruptive to the learner to be used in most scenarios. Similarly, learner *behavior* is a complex and multifaceted concept and can be influenced by many unobservable variables, such as their attention, motivation, and effort employed while exploring documents and formulating queries.

Therefore, we again use observable variables as proxies to the learner’s internal state. For instance, we can correlate a learner’s effort in exploring the topic of interest to the number of unique query terms used. When trying to model the learner’s attention while reading a document, the time spent reading the document and the number of mouse movements can be used as proxy variables. These underlying, latent variables and some commonly used proxy metrics are illustrated in Figure 1.4.

With a set of existing observable metrics that are cheap to observe and compute, researchers commonly look into machine learning models for their predictions of in-session learning [52–54, 86]. These models, varying from simpler linear regressions to random forests, boosted trees, and neural networks, are sometimes trained with hundreds of different metrics as features. These metrics can be as simple as the number of documents visited by the learner in their search session and as complex as the number and direction of mouse movements and the reading difficulty of the documents visited.

Such works usually aim to identify a set of variables that are good predictors of learning [53, 87]. Therefore, one under-explored research path is using *causal* methods to analyze these correlations. Such methods can be used to both predict and explain the learner’s knowledge gain [88]. While a few recent works have tried to cast the learning process with a causal lens [55, 56, 89], this is still an incipient field of research in SAL and even in the broader IIR field.

In this thesis, we cast the learning process as a causal model in Chapter 5. We discuss how to model the complex interplay between the learner, the search system, and the knowledge they are trying to acquire using causal methods. We also discuss how existing learning metrics, especially ones focused on multiple-choice questionnaires, cannot explain learning. We also present an in-depth discussion of proxy measures of learning and search behavior and their corresponding higher-level latent variables, such as the learner effort and the quality of documents retrieved.

1.2 Goals and Research Questions

Given the discussion above, it is clear that the field of Search-as-Learning is complex and multifaceted. It covers aspects ranging from the most technical and algorithmic, such as the design of search interfaces and retrieval methods, to the most human and psychological, such as the learner's behavior and knowledge acquisition process.

However, the full potential of SAL research blossoms at the intersection of these complexities. It is not enough to design a new retrieval method or search interface. Instead, the fundamental questions, and, therefore, discoveries, arise when we try to understand how these changes impact real-life learners' behavior and knowledge acquisition process. Therefore, exploring this intersection, this thesis has the following original research questions that permeate all the chapters:

- ORQ1** What changes in the search engine can significantly impact learners' behavior and knowledge acquisition process?
- ORQ2** How can we model the learner's behavior and knowledge changes throughout their search session?
- ORQ3** What behaviors and metrics best explain and predict a learner's knowledge gains at the end of their search session?

The first question, while broad, is harder to answer, given the broad scope of possible changes in a search engine. However, it is arguably one of the most critical questions, the final goal of all SAL research. Any proposed intervention in a search system to enhance the learning experience follows from this question.

Given the complexity of the first research question, **ORQ2** goes one step deeper and tries to understand how the learner's behavior and knowledge changes throughout their search session. Given the enormous space of possible interventions, being able to model a learner and their interactions with a search engine and how their knowledge evolves as they submit queries and interact with documents, we have a clearer view of how a given intervention may impact the learner's behavior and, consequently, their learning.

Finally, **ORQ3** asks how the learner's behavior and knowledge changes throughout their search session and how these factors interact to explain and predict a learner's success in their learning goals. This question allows us to understand the most critical factors influencing the learner's knowledge acquisition and what factors should (or should not) be further explored in future research.

1.3 Contributions

This thesis comprises four main chapters, each looking at the research questions above with a different lens and contributing to the SAL field in different ways. Here, we summarize the main contributions of this thesis, how they relate to the research questions above, and in which chapter they are discussed.





- ✓ In Chapter 2, we show that, while impacting the learner's knowledge acquisition is not always feasible, we can significantly impact their behavior by applying scaffolding techniques from the field of education to the search system. We help answer **ORQ1** by

demonstrating that, when explicitly showing to learners the list of subtopics of their learning topic, their searching behavior can drastically change. Learners in these sessions tend to explore more and search for longer.

- ✓ Chapter 2 also discusses that learners can be overwhelmed by too much feedback on their search performance. We answer **ORQ1** by showing that while progress bars can be a helpful tool to guide learners in their search sessions, they can also be detrimental to their experience if not used carefully.
- ✓ We propose RULK in Chapter 3, a framework for estimating and tracking a learner’s knowledge as they interact with a search system. It answers **ORQ2** by modeling the learner’s knowledge as an evolving probabilistic state constantly updated as the learner interacts with documents and submits queries.
- ✓ Chapter 4 introduces a simulating agent for IIR users specifically designed for SAL research. With this agent, the SACSM, we simulate how different types of learners interact with search systems, simulating prototypical learning strategies to help answer **ORQ2**.
- ✓ Chapter 5 introduces the KALM, a causal model that connects latent variables—such as the learner’s effort while formulating queries and the quality of the documents they read—to the learning outcomes of their search session. KALM helps us answering **ORQ3**.
- ✓ Chapter 5’s KALM also addresses **ORQ2** by providing a model for how these latent variables interact through a causal lens, and how changes in the effort the learner puts into formulating queries reflect in the quality of the documents they read and, consequently, in their knowledge gain.

1.4 Thesis Origins

The chapters of this thesis have origins in different research papers published (or in the process of publishing) during my Ph.D. No paper was written in isolation, and all chapters resulted from collaboration between me and my co-authors. Here, we list the papers that originated each chapter and where they were initially published.

- Chapter 2**  *Arthur Câmara, Nirmal Roy, David Maxwell, Claudia Hauff*. 2021. Searching to learn with instructional scaffolding. In CHIIR 2021 [62].
- Chapter 4**  *Arthur Câmara, David Maxwell, Claudia Hauff*. 2022. Searching, learning, and subtopic ordering: A simulation-based analysis. In ECIR 2022 [90].
- Chapter 3**  *Arthur Câmara, Dima El-Zein, da-Costa-Pereira, Célia*. 2022. RULK: A Framework for Representing User Knowledge in Search-as-Learning. In DESIRES 2022 [60].
- Chapter 5**  *Arthur Câmara, Claudia Hauff*. 2023. Keep KALM and Search On: Unveiling Causal Connections in Search-as-Learning. Under Review.

2

Searching to Learn with Instructional Scaffolding

In this chapter, we propose and experiment with strategies for improving learning outcomes based on the principle of instructional scaffolding, a concept borrowed from the learning sciences. When scaffolding is employed, instructors provide learners with support throughout their autonomous learning process. This contrasts with a traditional classroom, where the instructor leads the learning process. While scaffolding effectively improves learning in both digital and traditional learning contexts, it has not been studied in the context of SAL. Therefore, in this chapter, we study the hypothesis that using scaffolding techniques within a search system can effectively help learners achieve their learning objectives while searching. As such, this chapter investigates the incorporation of scaffolding into a search system employing three different strategies (as well as a control condition): (i) AQE_{SC}, the automatic expansion of user queries with relevant subtopics; (ii) CURATED_{SC}, the presenting of a manually curated static list of relevant subtopics on the search engine result page; and (iii) FEEDBACK_{SC}, which projects real-time feedback about a user's exploration of the topic space on top of the CURATED_{SC} setting. To investigate the effectiveness of these approaches for human learning, we conducted a user study ($N = 126$) where participants were tasked with searching and learning about topics such as 'genetically modified organisms'. The dataset derived from this chapter is also used in all other chapters of this thesis. We find that (i) the introduction of the proposed scaffolding methods in the proposed topics does not significantly improve learning gains. However, (ii) significantly impacts search behavior. Furthermore, (iii) immediate feedback of the participants' learning (FEEDBACK_{SC}) leads to undesirable user behavior, with participants seemingly focusing on the feedback gauges instead of learning.

This chapter is based on the following paper:

📖 Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. 2021. Searching to Learn with Instructional Scaffolding. In CHIIR. ACM, 209–218 [62]. This paper received the best student paper award at CHIIR 2021 🏆.

2.1 Introduction

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 [53] has shown that learners, on average, are aware of only 40% of the different aspects of 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 [91–94]. Using scaffolding, an instructor or teacher provides *guidance* to learners through various means 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. The scaffold is withdrawn by the end of the learning process, as no more guidance should be required.

When translating the idea of instructional scaffolding to digital learning, Hill and Hanafin [16] proposed several 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 is important”. In this chapter, we explore to what extent conceptual scaffolds—which have been shown to be beneficial for human learning in digital environments—are beneficial for learning while searching.

To this end, we propose three different strategies for incorporating scaffolding into learners’ search sessions: (i) AQE_{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) $CURATED_{SC}$, 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 [95] (in contrast to AQE_{SC} the learner here is explicitly aware of the subtopics related to the main topic); and (iii) $FEEDBACK_{SC}$, which projects real-time feedback about the user’s exploration of the topic space on top of the $CURATED_{SC}$ visualization. This is inspired by recent works like ScentBar [61] and von Hoyer et al. [33], 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 [96] 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 measured 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 To what extent the Introduction of scaffolding impacts the behavior of the learners?

Our main findings can be summarized as follows. (i) The proposed scaffolding methods are shown not to be significantly effective for increasing learners’ knowledge gain, with gains ranging from 30% to a detrimental effect of 7% compared to the control condition. (ii) The type of scaffold significantly impacts learners’ search behavior. We also show that the scaffolding components heavily influence the participants’ queries. (iii) Participants

in the CURATED_{SC} and FEEDBACK_{SC} conditions were more engaged with the platform, issued more queries, viewed more documents, and spent more time searching. At the same time, the FEEDBACK_{SC} cohort exhibited behavior, indicating that they focused on the feedback gauge more than the actual learning process.

2.2 Related Work

In Chapter 1, we discussed current research in SAL and how *proxy measures of learning* are employed to evaluate a learner’s knowledge gain at the end of their session. Therefore, this section briefly discusses some main findings related to how adapting a search engine to a learner may directly contribute to their learning outcomes.



Retrieval System Adaptation

Some works have explored the adaptation of the retrieval system to support learning. Syed and Collins-Thompson [67] designed a retrieval algorithm specifically for vocabulary learning by ranking documents according 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 lead to significantly higher learning gains (a follow-up study showed this to hold over a long period as well [31]). However, it should be noted that the user study fixed the documents to read for each topic instead of allowing participants to search and adapt the retrieved results on the fly. Recently, Syed et al. [97] investigated whether automatic question generation can improve learning outcomes *while reading* a document. Although the improvement in learning outcomes was limited to learners with low levels of prior knowledge, it is not far-fetched to imagine such an interface component to be incorporated in a search system.

Visualization of Search Progress

Lastly, we want to point to the work on ScentBar by Umemoto et al. [61] 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 several intrinsically diverse tasks showed that users were better able to determine when to stop searching for relevant information when the amount of missed information was made visible to them.

2.3 Instructional Scaffolding in SearchX

We implemented our scaffolding variants as part of SearchX [96], a modular, open-source search framework that provides quality control features for crowdsourcing experiments and fine-grained search logs¹. Figure 2.1 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

¹Behaviors logged include document dwell time, clicked documents, mouse hovers, document snippets shown on screen, bookmarked documents, etc.

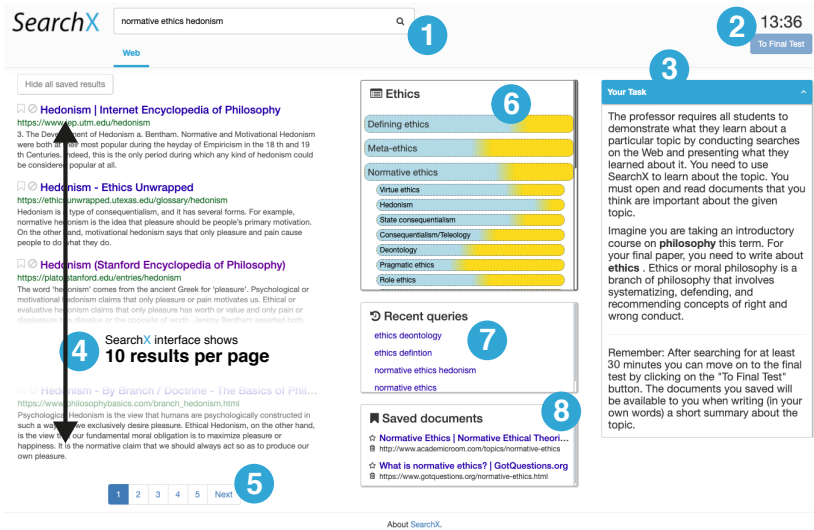


Figure 2.1: The SearchX interface: the eight annotated interface components are described in Section 2.3. Note that the scaffolding component (displaying the Ethics topic) shows the $\text{FEEDBACK}_{\text{SC}}$ scaffolding variant—complete with yellow progress gradients.

the scaffolding component, with the $\text{FEEDBACK}_{\text{SC}}$ 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 discuss our scaffolding variants after introducing the approach behind our topical outlines.

2.3.1 Topical Outlines

A key ingredient of all our scaffolding strategies is the topical outlines for each learning topic (cf. Figure 2.1, where the scaffolding component shows part of the outline for the topic Ethics). Effective outlines are typically hierarchical in nature [16, 98], 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.

Such outlines can either be created by instructors [99, 100] or automatically (known as *outline generation* [101]). The latter is desirable as it is scalable and not dependent on a domain expert’s availability—this is 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 [102]². This can be considered as employing a *crowd of experts* [103] for creating the outline. A concrete example outline from Wikipedia for the *subprime*

²We note that topical outlines can also be extracted from textbooks 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 2.4).

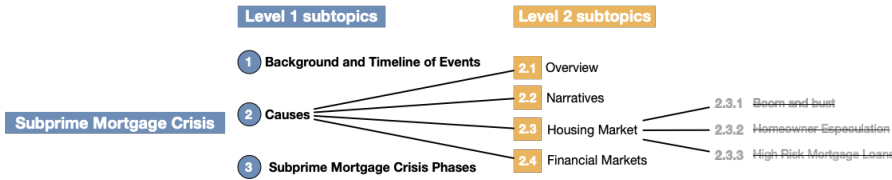


Figure 2.2: Hierarchical topic structure for the topic *Subprime Mortgage Crisis*. Topic and structure derived from TREC CAR 2017 [102]. Note that third-level subtopics (and deeper) and footnotes/references are excluded (illustrated in the figure by use of ~~strikethroughs~~).

mortgage crisis topic is shown in Figure 2.2. Each outline was manually cleaned and we only consider subtopics up to two levels deep (we refer to those levels as *L1* and *L2*, cf. Figure 2.2), and we remove generic subtopics that occur across most topics (such as *References*).

2.3.2 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 CURATED_{SC} and FEEDBACK_{SC}) or the backend. In the backend, we can either modify the retrieval function (as proposed by Syed and Collins-Thompson [31, 68]) or reformulate the *to-be-submitted* queries. We chose the latter setup, as this is agnostic to the employed search engine (*Bing* in our 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 minutes³, 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; 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 2.1, 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.

2.3.3 Variant CURATED_{SC}

As mentioned, the next two scaffolding techniques focus on changes to the front-end. Here, we explore to what extent making learners *explicitly aware* of the topical outline impacts their search behaviors and knowledge gains. The first variant, CURATED_{SC}, is as seen in Figure 2.1, 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 mentioned, the compo-

³As we set a minimum task time of 30 minutes in our study, this is a reasonable setup.

nent has fixed dimensions but can be scrolled anytime. While the task description does not point explicitly to the component (as seen on the right of Figure 2.1), 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 subtopic can itself be broad enough to be split into several subtopics. Explore the subtopics as much as you can.

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

2.3.4 Variant FEEDBACK_{SC}

While CURATED_{SC} presents a static component to the learner that does not change during the search session, in FEEDBACK_{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 2.1, inspired by Umemoto et al. [61]. 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 must happen in (near) real-time to avoid a noticeable lag.

When ten search results (documents) are retrieved from the Bing Search API for a given query, we compute their semantic similarity for each document/subtopic pair. To this end, we tokenize both document and subtopic⁴, and extract their sentence embedding using a pre-trained BERT-base model [79]⁵. 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*. 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 above. Each document can thus contribute to the progress score of multiple subtopics. We consider the subtopic's progress bar completely 'filled up' when the aggregate similarity score reaches 10. This constant is determined based on the search session length and the number of subtopics present.

⁴For tokenization, we employ <https://github.com/huggingface/tokenizers>.

⁵Here, 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>.

2.4 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.

2.4.1 Topics

We used a subset of the 117 training topics from the TREC CAR 2017 [102] 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 ten topics based on their difficulty and complexity, judged by 17 STEM graduate students⁶. Finally, we removed three 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 2.4.3 for how users were assigned to each topic). Ultimately, we worked with the seven topics in Table 2.2. 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.

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 [53, 59, 67, 83, 87] (i.e., users are queried about their knowledge of the concepts *before and after* the search session). We resorted to a vocabulary knowledge test to evaluate domain knowledge. To this end, two of the authors manually selected ten concepts per topic (listed in Table 2.1) 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 articles. 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*).

2.4.2 Metrics

We evaluate the knowledge gain of a concept by utilizing the *Vocabulary Knowledge Scale* (VKS) [82] across four levels (in line with [59, 83]):

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.*

⁶Each assessor received all 48 topics in a randomized order and was asked to select the ten that appeared most difficult to them for learning about. Finally, the ten topics selected most often were chosen as our topic set.

Table 2.1: Overview of the ten concepts per topic 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.

Topic	Concepts
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

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 score of 1. Choosing level (4) indicates the participant is confident in their assessment, and we assign it a score of 2⁷.

In line with [31, 68, 83, 104, 105], we utilize realized potential learning (RPL) as our main learning gain metric. RPL depends on the absolute learning gain (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 learning gain (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:

⁷We note that this scoring scheme is equivalent to the *fine-grained setup* employed by Moraes et al. [83].

Table 2.2: Overview of the topics used in our study, with associated statistics. Two-way ANOVA tests revealed no significant differences in the average number of queries between topics ($F(6, 99) = 2.01, p = 0.07$) or between the average number of bookmarks ($F(6, 99) = 0.41, p = 0.87$).

	Business cycle	Ethics	Genetically modified organisms	Irritable bowel syndrome	Noise induced hearing loss	Radiocarbon dating considerations	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)



Figure 2.3: RPL examples: \triangle represents vk_s^{pre} and ∇ represents vk_s^{post} . Here, $n = 10$. Note that MLG is dependent only on vk_s^{pre} , while ALG is the difference between vk_s^{post} and vk_s^{pre} . RPL is defined by the ratio between ALG and MLG.

$$\begin{aligned}
 ALG &= \frac{1}{n} \sum_{i=1}^n \max(0, vk_s^{post}(v_i) - vk_s^{pre}(v_i)) \\
 MLG &= \frac{1}{n} \sum_{i=1}^n 2 - vk_s^{pre}(v_i) \\
 RPL &= \frac{ALG}{MLG}.
 \end{aligned}$$

Here, $vk_s^X(v_i)$ 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 2.3.

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 done by Syed and Collins-Thompson [31], where vocabulary items such as earth and temperature were mixed with more technical vocabulary items). We do

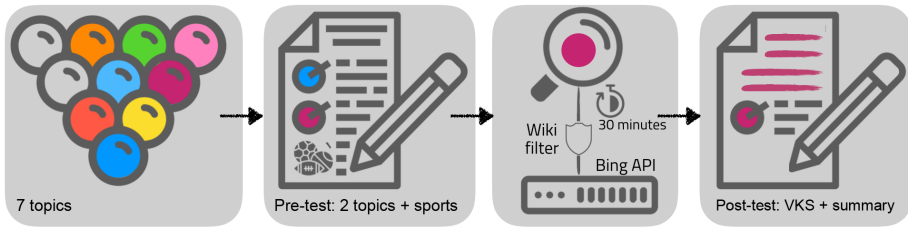


Figure 2.4: Overview of the flow of our user study. From seven topics, learners take a pre-test on two random topics and “sports” as a sanity check topic. The topic on which the learner scores the least is selected for their search session. A session lasts at least 30 minutes, using the SearchX platform with the Bing search API. Pages from Wikipedia and mirrors are filtered. At the end of the session, learners take a post-test and are asked to write a summary of what they learned.

not believe this to be necessary based on how we selected our concepts, as they are similarly difficult. Some prior works have also manually annotated participants’ summaries or mind maps to derive qualitative and quantitative metrics [24, 35, 36]. We leave the analyses of the user summaries we collected in this manner for future work.

2.4.3 Study Workflow

The flow of our user study is presented in Figure 2.4; it is implemented within our SearchX instance. Two of our seven topics are randomly selected when a participant enters the study. 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 to learn about during the search session. We introduced the simulated learning task as shown in Figure 2.1, 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 Wikipedia⁸. 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.

Participants could search, view, and bookmark documents during the search session. In the post-test, we asked them again about their knowledge of the ten concepts for their topic. In addition, we asked them to write a summary (100 words minimum)⁹ about the topic. We note here that the knowledge tests require understanding but no application or

⁸We blocklisted a total of 72 domains. All subtopics were submitted to the Bing Search API, with the top 10 results returned. Each result was inspected to determine whether it came from a Wikipedia clone in our blocklist.

⁹Specifically, 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 for your fellow students who read it to get a first idea of the topic without having to search for it themselves. Please write your summary here (at least 100 words).”

synthesis (i.e., higher-level cognitive processes of learning [45]) of the materials—here, we make the bookmarked documents available to our participants.

2.4.4 Study Participants

We conducted our study on the *Prolific platform*¹⁰ across three days. To ensure responses of high quality, we required our participants to have at least 15 previous submissions, an approval rate of 90+%, and be fluent 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 were actively using our search system instead of running down the timer). Of the valid participants, 65 were male, 59 were female (2 withheld gender information) with a median age of 27 (minimum 18, maximum 63). 44 participants reported the 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 levels of education.

We report in Table 2.2 the number of participants per topic. The topic '*radiocarbon dating considerations*' had the maximum number of participants assigned (21), while '*genetically modified organisms*' and '*irritable bowel syndrome*' the minimum (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 collected answers for 1260 VKS questions and 126 essays. 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*¹¹ and *incorrect*¹². We find that 25.2% of the vocabulary scores self-assessed as (3) were correct; 65.9% was partially correct; and the remaining 8.9% were incorrect. Among the self-assessed definitions 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.

2.5 Results

We now turn to addressing our research questions proposed in the start of this chapter. 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 ($p < 0.05$)¹³.

¹⁰<https://www.prolific.co/>

¹¹Partially correct definition example of *tinnitus* (i.e., noise induced hearing loss topic): "*hearing loud sounds in one's ears.*"

¹²Incorrect definition example of *genomes* (genetically modified organism topic): "*the amount of chromosomes.*"

¹³For further investigations, An anonymized version of the data is available at <https://github.com/ArthurCamara/searchx-scaffolding>

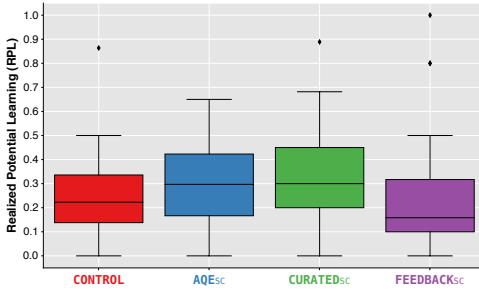


Figure 2.5: RPL over the four different conditions.

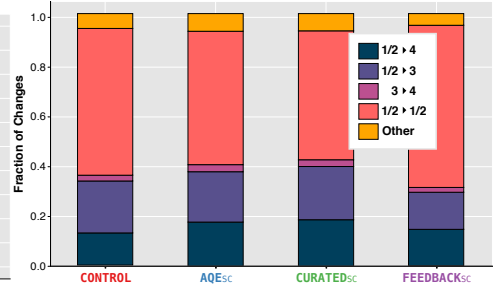


Figure 2.6: Fraction of change in answers between the pre- and post-test VKS questionnaires.

2.5.1 RQ1: Impact of Scaffolding on Learning

In Figures 2.5 and 2.6, we present the RPL across the four conditions (each one with between 28 and 36 participants, and an average search session duration¹⁴ of more than 36 minutes, cf. Table 2.3), and a more fine-grained presentation of the knowledge changes.

Table 2.3: Mean (\pm standard deviations) of RPL and search behavior metrics across all participants in each condition. [†] indicates two-way Anova significance, while ^C, ^d, ^U, ^F indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) increases vs. CONTROL, AQESc, CURATEDSc and FEEDBACKSc respectively.

	CONTROL	AQESc	CURATEDSc	FEEDBACKSc
I. Number of participants	30	28	33	36
II. Search session duration (minutes)	36m33s($\pm 12m15s$)	39m59s($\pm 10m59s$)	41m31s($\pm 13m6s$)	38m15s($\pm 12m46s$)
III. RPL	0.26(± 0.18)	0.30(± 0.16)	0.31(± 0.20)	0.24(± 0.24)
IV. Number of queries [†]	5.13(± 2.61) ^{U,F}	5.29(± 2.98) ^{U,F}	11.09(± 6.99) ^{C,d}	11.86(± 7.60) ^{C,d}
V. Fraction of query terms coming from topical outline [†]	0.26(± 0.28) ^{U,F}	0.33(± 0.31) ^{U,F}	0.58(± 0.34) ^{C,d}	0.58(± 0.29) ^{C,d}
VI. Fraction of topical outline terms used for querying [†]	0.04(± 0.04) ^{U,F}	0.05(± 0.04) ^{U,F}	0.32(± 0.23) ^{C,d}	0.34(± 0.24) ^{C,d}
VII. Average time between queries (minutes)	5m57s($\pm 5m26s$)	6m31s($\pm 8m31s$)	3m31s($\pm 2m45s$)	3m52s($\pm 4m40s$)
VIII. Average time between document close and next document load (secs.)	60.15(± 27.17)	68.06(± 33.44)	74.42(± 45.14)	57.32(± 39.13)
IX. Average document dwell time (secs.)	76.77(± 51.14)	100.61(± 61.59)	92.15(± 97.60)	55.33(± 51.04)
X. Number of unique documents viewed [†]	14.77(± 8.85)	10.96(± 4.08) ^F	14.09(± 7.95)	18.50(± 9.56) ^d
XI. Number of unique document snippets viewed [†]	97.47(± 47.37) ^F	81.07(± 44.58) ^{U,F}	136.42(± 76.97) ^d	152.44(± 84.23) ^{C,d}
XII. Fraction of retrieved documents that would affect the gradient bar	0.39(± 0.19)	0.48(± 0.21) ^{C,U}	0.41(± 0.18) ^F	0.35(± 0.19)
XIII. Fraction of clicked documents that affected the gradient bar	0.62(± 0.19)	0.72(± 0.16)	0.72(± 0.20)	0.71(± 0.15)

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

¹⁴We consider the search session duration as the time between the first query issued by the learner and the time they close the last viewed document.

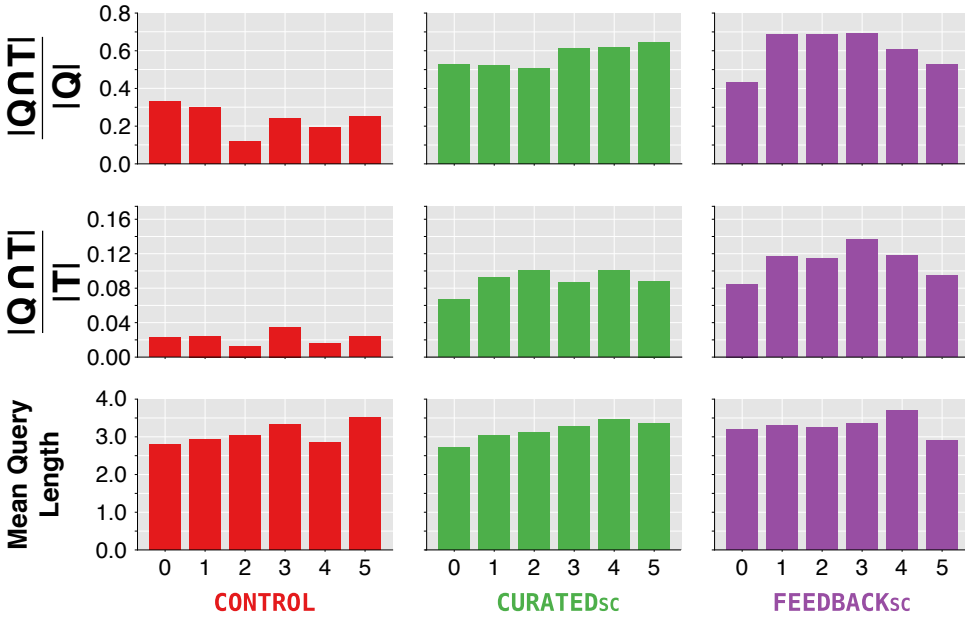


Figure 2.7: For each row: (*top*) the fraction of query terms taken from topic outlines; (*middle*) the fraction of topic outline terms used for querying; and (*bottom*) the mean query length, over 5-minute blocks (*x* axes) of the 30-minute search session, considering: CONTROL (*left*); CURATED_{sc} (*center*); and FEEDBACK_{sc} (*right*). Here, we consider the first query instance as the start of the first interval.

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} 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} greatly impacted 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 benefit the learning experience (as also envisioned, among others, by von Hoyer et al. [33]). We further investigate possible reasons for this finding in Section 2.5.2.

In Figure 2.5, we present the distribution of RPL scores across the four study variants. RPL does not cross 0.5 for CONTROL, AQE_{sc} and FEEDBACK_{sc}—except for outliers. This means that, on average, participants in those cohorts gained knowledge on less than half of the concepts they had little to no prior knowledge on. The long top whisker on the boxplot for CURATED_{sc} participants shows that, although the RPL varies more than other conditions, it is potentially more beneficial for learning.

To further analyze how the conditions differ, we provide a detailed breakdown of the knowledge state transitions in Figure 2.6. 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 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, so 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 topics.

2.5.2 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 2.3.

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 [61]). 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 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 2.3 (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 CURATED_{SC} and FEEDBACK_{SC}. At the same time, this is the case for 33% and 26% on average for AQE_{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 CURATED_{SC} and FEEDBACK_{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 2.7, we break down this comparison of query terms and topical outline terms further by splitting our search sessions into five-minute intervals and computing $\frac{|\mathcal{Q}_n \cap \mathcal{T}|}{|\mathcal{Q}|}$ and $\frac{|\mathcal{Q}_n \cap \mathcal{T}|}{|\mathcal{T}|}$ separately for each interval. We find that participants in the CONTROL

condition were not picking up more topical outline terms over time (even though they have read more documents on the topic by each passing interval). However, we see a slight increase over time for CURATED_{SC} and FEEDBACK_{SC}, which then drops again in the later stages of the search session.

Too Much Feedback Considered Harmful

Previous works [53, 59, 83, 87] have shown the number of queries issued to be a good proxy for learning. In our work, this finding holds for CURATED_{SC}, though not for FEEDBACK_{SC}: on average, a similarly high number of queries were submitted, but the learning gains for FEEDBACK_{SC} are low. For completeness, the bottom row of plots in Figure 2.7 shows mean query lengths across time: as search sessions progressed, queries tended to become longer.

We hypothesize that FEEDBACK_{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 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 2.3, row IX): it is on average 55 seconds in the FEEDBACK_{SC} variant, which is significantly lower than that of the CURATED_{SC} and AQE_{SC} variants (with an average document dwell time of 92 seconds and 100 seconds, respectively). At the same time, FEEDBACK_{SC} participants viewed, on average, the most documents and the most document snippets (Table 2.3, rows X and XI).

Finally, we look into the probability of a user opening a document that would change the progress bar. A higher number here implies that users are actively looking for “good” documents to click, which would increase the progress bar, instead of documents that would be relevant. On average, in all cohorts, 66% of the documents retrieved would produce some change on these bars. However, a slightly higher fraction of the clicked documents, 70%, actually changed the progress bar. This is another clue that users are actively looking for snippets from documents that would produce a change in the progress bar (Table 2.3, rows XII and XIII).

In addition, in Figure 2.7, we consider the query length across time: that although query length is similar in every 5-minute interval during the search session for CONTROL, CURATED_{SC} and FEEDBACK_{SC}, the percentage of query terms coming from outline terms and percentage of outline terms used for querying is higher for CURATED_{SC} and FEEDBACK_{SC} throughout the session. This observation follows from Table 2.3, rows V and VI. Figure 2.7 (middle row) also shows that, as the search session progresses, participants from FEEDBACK_{SC} tend to use more terms from the outlines than their CURATED_{SC} counterparts.

To explain the large gap between the results of CURATED_{SC} and FEEDBACK_{SC}, Swinnen et al. [106] 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. [107] corroborate these findings in 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.

2.6 Conclusions

In this chapter, we have explored three strategies to introduce instructional scaffolding into a web search system to improve a learner's knowledge gain during the search process. These strategies were: (i) automatic query rewriting (AQE_{SC}) which is agnostic to the search backend; (ii) a curated static topical outline (CURATED_{SC}); and (iii) a curated topical outline with instant feedback on the exploration of the topic space (FEEDBACK_{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 impact learning outcomes. However, we opened a new research venue, showing that scaffolding significantly changes user behavior on several metrics. This is shown by our analysis answering **RQ2**, where we show that explicit scaffolding (namely CURATED_{SC} and FEEDBACK_{SC}) significantly alters users' behavior in some important search metrics, like dwell time, number of queries issued and number of clicks. This is important and should lead to further investigation into using this behavior difference to support learners better.

Additionally, we have speculated that the discrepancy in behavior between CURATED_{SC} and FEEDBACK_{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. The difference in dwell time corroborates this, as the FEEDBACK_{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 do learners respond to best? 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 be personalized depending on users' domain expertise and other user characteristics. While exploration into (non-personalized) automatic outline generation [101] is relatively new, it is still unclear whether such slightly noisy outlines benefit users' learning outcomes. In addition, it remains to be seen to what extent the changes in user behavior hold across time (as explored by Syed and Collins-Thompson [31]) and whether users remain engaged over time when a scaffolding component is permanently introduced on the search interface. We also need to consider that we measured learning with a vocabulary knowledge task, which covers only the lowest cognitive levels of learning [108]. Is scaffolding beneficial for learners who face learning tasks that target higher cognitive levels of learning [84]?

Finally, this chapter also directly answers our first research question, **ORQ1**, from Chapter 1: *“What changes in the search engine can significantly impact learners' behavior*

and knowledge acquisition process”? We answer it by showing that not only our $\text{CURATED}_{\text{SC}}$ and $\text{FEEDBACK}_{\text{SC}}$ approaches can significantly alter learner behavior, but also that learners in the $\text{FEEDBACK}_{\text{SC}}$ cohort can be detrimental, as it seems to trigger exploration in detriment of quality. We also reach a similar conclusion that exploration does not necessarily lead to improved learning gains when analyzing the same dataset in Chapter 5.

3

3

RULK: A Framework for Representing User Knowledge in Search-as-Learning

In this chapter, we tackle the problem of tracking and predicting a learner’s knowledge state throughout their learning journey. This is a crucial piece for improving a learner’s experience using a SAL-oriented search system. Accurately estimating the learner’s knowledge while they search, with no human intervention, would allow researchers to dynamically adapt search results and the system’s interface, to better support the learning at their current knowledge level. Therefore, in this chapter, we propose RULK, a framework for representing and updating a learner’s knowledge state while they interact with the search system.

The intuition behind RULK is simple. Keeping an internal representation of the learner’s knowledge updated as the user progresses in their search session, the framework estimates how much the user knows (or still does not know) about a given topic by comparing the learner’s knowledge state to a target knowledge level. We implement two variations of RULK, each embedding the learner’s knowledge in a distinct latent space. The first, RULK_{KW} , uses a keyword-based representation of the knowledge. The second, RULK_{LM} , uses dense embeddings produced by a transformer-based language model. Our experiments show that the estimations of user knowledge produced by RULK correlate with actual user knowledge, clearing the path to future learning-focused search systems to provide an even better user experience.

3.1 Introduction and Related Work

One of the recurring themes for research in learning-oriented IR systems is to consider the learner’s learning goals while retrieving documents [31, 68, 109, 110]. One significant influence on these developments is the concept of the Anomalous State of Knowledge (ASK), introduced by Belkin et al. [15].

The ASK principle proposes that individuals possess an internal knowledge model of the world, which continuously evolves as they acquire new information. Suppose they detect an anomaly, such as missing or contradictory information. According to the ASK principle, they feel compelled to seek additional information, often by using a search system, to resolve the inconsistency.

Nevertheless, most current search systems focus on retrieving documents relevant to a single query, neglecting a more comprehensive view of the user’s session and accumulated knowledge [11]. A user with a learning goal (i.e., a learner) engages with a search by submitting queries and consuming retrieved documents (e.g., reading texts or watching videos). Consequently, the learner’s internal cognitive state evolves throughout the session as they acquire the knowledge from the consumed documents [49–51].

Therefore, *representing* the learners’ cognitive (or knowledge) state and *updating* it during a search session is an important task for any SAL-oriented system. By estimating the learner’s knowledge state, a search system can more efficiently assist them in achieving their learning goals by incorporating these signals both in the back-end (e.g., in the retrieval and ranking algorithms) and the front-end (e.g., suggesting more relevant queries, and adapting the user interface) of the system [62, 67, 111].

To address this need, we propose a novel framework for **Representing User Learning and Knowledge (RULK)**. This framework combines concepts already existing in some SAL systems for representing and updating a learner’s state [31, 57, 85, 90] and introduces a novel component: an **estimator**.

Typically, existing systems with knowledge representation comprise at least two components: A **feature extractor** that transforms clicked documents into features, and an **updater** that adjusts the internal representation of the learners’ knowledge state. These components are generally implemented by operating within a latent space, usually a keyword-based space, defined by extracting weighted keywords from documents [57, 67, 85] or by embedding the documents into a dense document-level embedding space [62, 90].

In RULK, we add a third component to that mix, an **estimator** that *estimates* how close a learner is to reaching a “target knowledge state”. The estimation is performed by comparing the learner’s current knowledge state, as maintained by the updater, to a “target” knowledge (e.g., a list of keywords or documents relevant to the topic) represented in the same latent space. Although predicting learner knowledge is not novel [53, 53, 54, 87, 112], our framework aims to not only predict but also track the learners’ knowledge state throughout their sessions rather than only at the end ¹.

To illustrate how one might instantiate a system following the RULK framework in a SAL system, we implement two variations: RULK_{KW} , which utilizes keyword representations inspired by El Zein and da Costa Pereira [57, 85], and RULK_{LM} , which uses embed-

¹Due to a lack of available data of how learner’s knowledge change during the search session, we do not directly measure the correlation between the estimated and the real knowledge during the session in this Chapter. However, Section 3.4 does explore how the knowledge gain changes throughout the learner’s session

dings generated by a large language model, inspired by our work in Chapter 2. Using logs of users' interactions with a learning-oriented search system, we investigate each implementation's behavior, particularly the impacts of the estimator in predicting a learner's knowledge. Accordingly, we aim to answer the following research questions:

RQ1 Are the estimations of a learner's knowledge state produced by RULK correlated with their reported knowledge at the end of their session?

RQ2 How do RULK_{KW} and RULK_{LM} behave in sessions of different lengths?

3

By addressing these questions, we demonstrate that RULK can effectively estimate a learner's knowledge gain, with RULK_{KW} showing an overall higher correlation with the learner's assessed knowledge state at the end of their session as compared to RULK_{LM} . Additionally, we illustrate how combining RULK_{KW} with RULK_{LM} leads to a more robust estimation of the distance between the learner's knowledge and a target knowledge level, with less variability across different session lengths.

3.2 The RULK Framework

Our framework consists of three main components: the **feature extractor** (γ), the **updater** (σ), and the **estimator** (θ). Each component plays a role in how a learner's knowledge state, denoted as \vec{c}_{ks} , changes throughout their search session. Intuitively, when a learner reads a document d , γ transforms d into a fixed-sized numerical representation (i.e., an embedding) \vec{v}_d . Then, σ uses this representation to update an internal representation of their current knowledge state (\vec{c}_{ks}). Finally, θ compares \vec{c}_{ks} to a target knowledge vector (\vec{t}_{ks}) to get an estimation of the distance of the learner's current knowledge level compared to \vec{t}_{ks} . This distance is represented by \tilde{G} . This process is illustrated in Figure 3.1.

This section presents an overview of each of these components and how they update a learner's knowledge throughout their search session. We go into more details about how we implement each component when describing RULK_{KW} (a keyword approach) and RULK_{LM} (a language-model approach), two instantiations of RULK, in Section 3.3. We present a discussion on the caveats and assumptions we made in this chapter, as well as how one could address these, in Section 3.5

Feature Extractor (γ)

One of the main assumptions in SAL is that the *content* of the documents a learner read is the primary contributor to their learning gains during their searching session². Such documents could comprise various types of content a learner interacts with during their search session, such as web pages, textbooks, and videos.

The function of the feature extractor γ is then to project a document d clicked by a learner into an embedding \vec{v}_d with length m in a given latent space:

$$\vec{v}_d = \gamma(d). \quad (3.1)$$

²We discuss this assumption in more depth in Chapter 5.

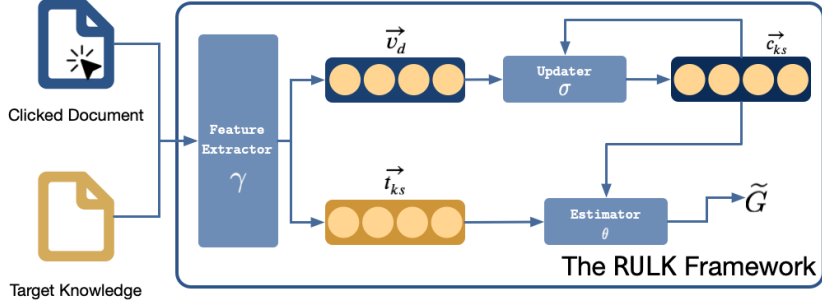


Figure 3.1: The RULK framework and its main components. First, a clicked document d is transformed into \vec{v}_d by γ . Next, σ updates the learner’s current knowledge state \vec{c}_{ks} with \vec{v}_d . Finally, θ compares \vec{c}_{ks} to a vector \vec{t}_{ks} , generated from a target knowledge document, to get an estimation of a learner’s knowledge level (\tilde{G}).

By standardizing the size m and encoding all documents in a common space (e.g., BERT or TF-IDF), γ enables RULK to compare and combine documents easily.

Another role of γ is to produce a “target knowledge state” (\vec{t}_{ks}) for the topic that the learner is exploring. This knowledge state is represented in the same embedding space as \vec{v}_d and, consequently, \vec{c}_{ks} . We assume that our SAL system has access to a set of “reference documents” that cover essential subtopics and themes for the topic (e.g., a textbook or a set of relevant documents) used as the target knowledge. We can then generate \vec{t}_{ks} by embedding this document according to Equation 3.1.

Updater (σ)

Within RULK, the learner’s knowledge is continuously monitored through an internal state vector, \vec{c}_{ks} , which has the same length m as \vec{v}_d embeddings produced by γ . Following the work of El Zein and da Costa Pereira [85], the updater (σ) refreshes \vec{c}_{ks} as new documents are read by the learner:

$$\vec{c}'_{ks} = \sigma(\vec{c}_{ks}, \vec{v}_d), \quad (3.2)$$

Here, \vec{c}'_{ks} represents the updated \vec{c}_{ks} vector, which is the refreshed state of the learner’s knowledge after engaging with a document d , represented by \vec{v}_d .

Estimator (θ)

RULK *estimates* the learner’s knowledge level (i.e., their distance from the target knowledge \vec{t}_{ks}) on a particular topic T during their search session by comparing their current knowledge state, \vec{c}_{ks} , to the target knowledge state, \vec{t}_{ks} :

$$\tilde{G} = \theta(\vec{c}_{ks}, \vec{t}_{ks}), \quad (3.3)$$

Here, \tilde{G} indicates an estimation of the learner’s knowledge level, compared to \vec{t}_{ks} , at a given point in time. The estimator (θ) implements a similarity function, such as the cosine

similarity. The rationale behind θ is that, as a learner progresses through their session, their knowledge state (\vec{c}_{ks}) gradually converges towards the target state (\vec{t}_{ks}). Because both vectors exist within the same latent space, the similarity between \vec{c}_{ks} and \vec{t}_{ks} can be seen as how near the learner is to attaining the knowledge contained in \vec{t}_{ks} .

3.3 Implementing and Validating RULK

In this section, we discuss how we validated RULK by describing our dataset, metrics, and implementation details for our two implementations of RULK, RULK_{KW}, and RULK_{LM}³.

Using our sample implementations, we apply the RULK framework to estimate the user's knowledge level, \tilde{G} , at the end of their session⁴. We analyze the logs from users captured during a user study to represent their current knowledge state \vec{c}_{ks} as it is updated along their search session. We initiate \vec{c}_{ks} as a vector of zeros for all users. (c.f. Section 3.5 for discussing our assumptions and caveats).

Dataset

To analyze our implementations of RULK, especially when estimating users' knowledge in a search session, we test our implementations of RULK on the dataset we initially used for the experiments in Chapter 2. The dataset was collected during a user study with crowdworkers recruited over Prolific⁵. Participants were asked to learn about a topic using a search system built on top of SearchX [96], an IIR research framework. We show some statistics of the dataset in Table 3.1. The full details of the dataset can be seen in Chapter 2.

When originally collected, the dataset did not contain the contents of the documents clicked by the participants, only their URLs. Therefore, we used the Wayback Machine API⁶ to fetch the documents with a date as close as possible to when the study was conducted (August 2020). Of all the documents, 33 did not have a snapshot, and neither are currently available. They were discarded from our experiments (i.e., we do not consider their impact on learners' knowledge).

The knowledge gains of a participant are measured by a VKS questionnaire [82] applied before and after their learning session. Such questionnaires ask learners to rate vocabulary terms relevant to the learning topic.

The results of the tests are used to compute two metrics: ALG and RPL. The ALG is simply the difference between a learner's score in the test applied before and after their search session. On the other hand, the RPL represents the fraction of knowledge the user acquired from the total knowledge they could obtain in their session. A learner that starts their session with no knowledge of the topic (i.e., scores a 0 in the pre-test) and reports full knowledge of the topic after their search session (i.e., scores 1.0 in the post-test) receives a RPL of 1.0. Alternatively, a user that reports knowing half of the vocabulary terms at the beginning of their session and then all of the vocabulary at the end of their session (i.e., scores 0.5 in the pre-test and 1.0 in the post-test) would also receive a RPL of 1.0, as they have both fully realized their potential learning gain during their sessions. For

³Our implementations can be found at https://github.com/ArthurCamara/RULK_SAL.

⁴While RULK can be used to estimate \tilde{G} at any point in time, the dataset used here only contains the knowledge level at the end of the session.

⁵www.prolific.com

⁶<https://web.archive.org/>

Table 3.1: Statistics, per user, extracted from the dataset used in Chapter 2.

	Total	Mean	Median
Number of users per topic	126	18.14 ± 2.79	19.0
Number of topics	7	—	—
Number of queries	1095	8.62 ± 6.47	7.0
Number of documents clicked	2116	16.66 ± 8.85	16.0
Number of snippets seen	15184	119.56 ± 72.43	105.0
Documents Clicked per query	-	2.78 ± 2.50	2.11
Session duration (minutes)	-	56.18 ± 14.58	54.05
Document dwell time (seconds)	-	79.94 ± 69.77	60.0
Pre-test scores (vks^{pre})	-	1.07 ± 1.60	0.00
Post-test scores (vks^{post})	-	6.21 ± 4.09	6.00
Absolute Learning Gain (ALG)	-	0.53 ± 0.38	0.50
Realized Potential Learning (RPL)	-	0.28 ± 0.20	0.25

more details on how the VKS questionnaire was designed and how RPL is computed, c.f. Section 2.4.2.

As discussed earlier in this Chapter, one of the strengths of RULK is that it can be used for measuring the learner’s knowledge throughout the session. However, with few exceptions (such as the work by Roy et al. [59]), data of knowledge gain throughout a learner’s session is not usually available. Therefore, while we explore how the learner’s knowledge evolves while they explore retrieved documents in Section 3.4, directly measuring the accuracy of RULK in this context should be tackled in future works.

Modeling a Target Knowledge State (t_{ks}^{\rightarrow})

As discussed in Section 3.2, The γ component of RULK has two functions. First, to produce the embedding \vec{v}_d of documents read by learners. Second, to model a “target knowledge state” (t_{ks}^{\rightarrow}) for the topic the learner is exploring.

An important consideration is how to define such a target. One possibility, commonly used in the Knowledge Tracing (KT) field [113], is to define a set of questions that the learner should answer correctly about the topic to be considered knowledgeable. Using this approach assumes that questions created by experts exist and that evaluating learners’ answers is cheap and computationally inexpensive, something unrealistic in most SAL settings. Also, disrupting learners’ search sessions may impact their behavior in negative ways [59]. Therefore, this is not ideal for the SAL scenario. Alternatively, defining a set of documents that should be visited is also possible. However, this would again require expert knowledge of the topic, and the coverage of the documents would be limited to the documents selected by the expert.

Instead, we follow an approach similar to the proposed by Syed and Collins-Thompson [67] and assume that our SAL system has access to a *reference document*. Such a document comprehensively covers essential subtopics and themes related to a topic (e.g., a textbook) and serves the target knowledge fed to γ to generate the vector t_{ks}^{\rightarrow} . In their work, the authors use the top results from Google using the topic’s title as a query as the reference document. Here, as our dataset contains topics derived from Wikipedia articles, we use

the respective Wikipedia article for each topic as the reference document as input for the γ component when generating \vec{t}_{ks} . It is also important to note that, during the user study, Wikipedia and pages deemed as mirrors or clones of Wikipedia were not shown to the participants, as these pages would likely be too similar to the reference document.

3.3.1 Implementing RULK

We present two example instantiations of RULK. The first, denoted as RULK_{KW} , relies on keyword-based features, whereas the second, RULK_{LM} , uses the BERT latent space. Therefore, the main distinction between these two implementations lies in what semantic space they employ for representing documents and the user’s knowledge. The former uses a bag-of-words representation, whereas the latter uses a dense vector space. In practice, they both share the same θ and σ components while differing in how they implement γ .

Both our implementations of RULK share the same implementation for the θ component, a cosine similarity between \vec{c}_{ks} and \vec{t}_{ks} :

$$\tilde{G} \approx \frac{\vec{c}_{ks} \cdot \vec{t}_{ks}}{|\vec{t}_{ks}| |\vec{c}_{ks}|}. \quad (3.4)$$

This sharing of θ is possible because both of our proposed implementations rely on latent spaces compatible with cosine similarity (i.e., a term frequency vector and a BERT embedding). However, other representations of \vec{c}_{ks} , \vec{t}_{ks} , and \vec{v}_d may require different similarity functions. For instance, if \vec{c}_{ks} and \vec{t}_{ks} were represented as probability distributions, the Kullback-Leibler divergence would be a more appropriate similarity function.

As for the Updater σ , following previous work [67, 85], we assume that the learner’s knowledge accrues monotonically. Specifically, as learners read documents d , represented by \vec{v}_d , σ implements an element-wise sum over all elements of \vec{v}_d and \vec{c}_{ks} .

RULK_{KW} Implementation

Our first instantiation of RULK is based on the frequency of relevant keywords a learner finds throughout their learning session. To do so, we adopt the learning model referred to as *vocabulary learning* [69], which operates at the lower levels of Bloom’s taxonomy [114]. This model posits that learners fulfill their learning objectives for a topic by acquiring a set of related vocabulary keywords.

Our proposal for RULK_{KW} also shares key ideas with the method proposed by Syed and Collins-Thompson [67] for improving the quality of learning-oriented search results. In both settings, the premise is that learners interested in learning about a topic should be exposed to a weighted set of relevant keywords $K = k_1, \dots, k_m$.

This set of keywords is sampled from “golden documents”. For Syed and Collins-Thompson [67], these are the top results from Google using the topic’s title as a query, and the keywords were selected according to their TF-IDF values in these documents.

For both of our implementations of RULK, we use the topic’s Wikipedia page as the “golden document”. Specifically for RULK_{KW} , we select the top-10 keywords (following Syed and Collins-Thompson [67]) from that article, as computed by the YAKE method [115], with a maximum n-gram size of 1, after stemming the keywords with the *Porter Stemmer* to avoid near duplicates. Finally, each keyword is assigned a weight based on its frequency

Table 3.2: Top-10 keywords extracted by Yet Another Keyword Extractor (YAKE) for each topic from their respective Wikipedia article. The article’s keywords are sorted by frequency, and the stemming is manually reversed for clarity.

Topic title	Keywords
Business cycle	treasury, bond, rate, business, cycle, notable, shorted, date, partisan, suggest
Ethics	ethics, relations, virtue, Epicurus, English, words, human, capacity, attention, norms
Genetically modified organisms	genetic, modify, crop, organism, engineer, food, GMO, European, wild, rabbit
Irritable bowel syndrome	found, IBS, patient, incur, postinfection, reduce, symptom, Carolina, technique, Novartis
Noise-induced hearing loss	hearing, protection, device, occupational, loss, permanent, due, compete, review, cause
Radiocarbon dating considerations	surface, water, give, bring, deep, thousand, year, upwell, mix, tree
Subprime mortgage crisis	financial, crisis, inquiry, commission, subprime, mortgage, loan, low, interest, rate

in the golden documents. In our case, the weight of a keyword is given by the number of its occurrences in the Wikipedia article for the topic. A list of the keywords used for the topics in our experiments can be found in Table 3.2.

We don’t use the same TF-IDF approach as Syed and Collins-Thompson [67] when selecting the keywords to avoid an unwanted bias in our results. A similar approach was used to select the keyphrases for the vocabulary learning questionnaire the user study participants were asked to perform (c.f. Section 2.4.1). Therefore, as our goal with the experiments is to verify if RULK agrees with the knowledge gain reported by the questionnaire results, using the same method would likely overestimate our correlations.

RULK_{LM} Implementation

Within the realm of IR, transformers-based language models [116], primarily influenced by BERT [79], have been found to excel in a variety of tasks [7, 80, 117–119], even without fine-tuning for the specific domain [120, 121].

Given the success of these transformers-based models, we assess whether such models can act as a feature extractor (γ) for RULK. Thus, we implement a transformers-based variant of RULK called RULK_{LM}, with a similar idea to our approach in Chapters 2 and 4 to track learner exploration of a topic.

The target knowledge \vec{t}_{ks} and the read document \vec{v}_d are represented by an embedding of fixed length m , as generated by the same language model. Given a document d (or, conversely, a reference document) with n sentences $\{s_1, s_2 \dots s_n\}$, γ generates, for each sentence s_i , an embedding of size m given by:

$$\vec{v}_{s_i} = \text{BERT}([CLS]; s_{i:l}; [SEP]), \quad (3.5)$$

where $;$ is the string concatenation operation, l the maximum input size of the model and $[CLS]$ and $[SEP]$ are special BERT tokens. \vec{v}_d (conversely, \vec{t}_{ks}) is then given by an element-wise sum over all \vec{v}_{s_i} .

For implementing RUL_{LM} , we use a 6-layer MiniLM [122] model with a hidden dimension size of 384. The model was also fine-tuned on the MsMarco dataset [123], as made available in the SBERT framework [121]⁷. Our choice of model is justified by its high-quality embeddings in multiple semantic representation tasks⁸.

Due to the model's restriction in input length (texts should have up to 382 tokens), we split the documents into sentences using the *NLTK*'s implementation of the *Punkt Sentence Tokenizer*, feed each sentence individually into the γ and sum their respective embeddings. We also experimented with other common approaches, such as truncating the document, averaging the sentences, and only using the sentence most similar to the user's query (MaxP [124]). Nevertheless, they all resulted in a worse or similar performance.

Mixing RUL_{KW} and RUL_{LM}

The latent spaces of RUL_{KW} and RUL_{LM} rely on two different assumptions. While RUL_{KW} captures token-level, syntactic similarities, RUL_{LM} captures a higher-level semantic representation of the documents. Therefore, we combine both approaches to provide a more comprehensive representation of the user's knowledge. Therefore, we propose a third implementation of RULK, RUL_{KW+LM} , that combines both RUL_{KW} and RUL_{LM} by interpolating their estimations of the user's knowledge gain. We define the updater θ for RUL_{KW+LM} as:

$$\theta_{LM+KW} = (\alpha)\tilde{G}_{LM} + (1 - \alpha)\tilde{G}_{KW}, \quad (3.6)$$

where \tilde{G}_{LM} and \tilde{G}_{KW} are the estimated knowledge level as predicted by RUL_{KW} and RUL_{LM} , respectively, and α is a hyperparameter for combining the weight of the two models. Experimentally, using a 5-fold cross-validation, we found that $\alpha = 0.48$ yields the better results. It is important to note that RUL_{KW+LM} does not directly modify a learner's \vec{c}_{KS} . Rather, it changes the implementation of θ by using both \vec{c}_{KS} , from RUL_{KW} , and RUL_{LM} , to estimate the learner's knowledge level.

3.4 Results

At the beginning of this chapter, we proposed two research questions, measuring different aspects of RULK:

RQ1: Are the Estimations of a Learner's Knowledge State Produced by RULK Correlated with Their Reported Knowledge at the End of Their Session?

To answer our first research question, we test the validity of RULK by computing the Pearson's correlation between the learner's self-reported knowledge level at the end of their sessions and the *estimated* knowledge level \tilde{G} , as measured by each implementation's θ . Our null hypothesis is that the reported and estimated knowledge levels are unrelated. We show the results of this analysis in Table 3.3.

As shown in Table 3.3, on the set of all learners, RUL_{KW} has a positive and statistically significant correlation with RPL, ALG and the learners' post-test scores. Additionally, as shown in the last three columns, the estimated knowledge level generated by all our implementations is highly correlated among themselves, with values above 0.7 in all cases and

⁷<https://huggingface.co/sentence-transformers/msmarco-MiniLM-L6-cos-v5>

⁸https://www.sbert.net/docs/pretrained_models.html

Table 3.3: Pearson’s correlation between estimated learning gains using a given implementation of RULK and reported learner’s learning. Values in **bold** indicate the best correlation against a learning metric. All correlations are statistically significant, $p < 0.001$

	ALG	RPL	POST	RULK _{KW}	RULK _{LM}	RULK _{KW+LM}
RULK _{KW}	0.3033	0.3088	0.3412	—	0.7808	0.9447
RULK _{LM}	0.2936	0.2930	0.3399	0.7808	—	0.9425
RULK _{KW+LM}	0.3164	0.3189	0.3609	0.9447	0.9425	—

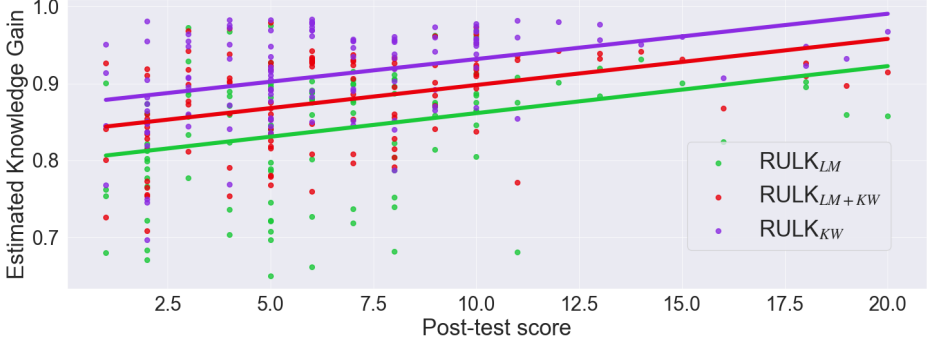


Figure 3.2: Scatter plot between the self-reported knowledge of a learner, measured by their post-test VKS score and the knowledge level as estimated by RULK_{KW}, RULK_{LM} and RULK_{KW+LM}.

values above 0.9 when RULK_{KW} and RULK_{LM} are compared with RULK_{KW+LM}, implying a high degree of consistency between our implementations. This result shows that, regardless of implementation, RULK is a viable option for representing learner knowledge.

Finally, Figure 3.2 shows a scatter plot between a learner’s knowledge state at the end of their search session and their post-test scores. From the plot, all three implementations overestimate the knowledge state of a learner. This discrepancy is mainly due to how \vec{t}_{ks} is implemented. As it sources from a single source (i.e., Wikipedia), learners can quickly “saturate” the knowledge from that document alone. Therefore, further calibration of \vec{t}_{ks} , or a more complex σ , where not all of the content of the document is added to \vec{c}_{ks} , would likely lead to better results.

We also show in Figure 3.3 a t-SNE projection of how the same learner’s \vec{c}_{ks} progresses throughout their search session in both vector spaces for RULK_{KW} and RULK_{LM}. In both cases, the \vec{c}_{ks} moves towards the \vec{t}_{ks} as the session progresses.

It is interesting to note how, in the RULK_{KW} setting, \vec{c}_{ks} overshoots \vec{t}_{ks} considerably. The same behavior can be observed for multiple users. This phenomenon mainly happens because of how γ is implemented for RULK_{KW}. The number of occurrences of a given keyword in the Wikipedia article defines the values in \vec{t}_{ks} . Therefore, if a learner encounters a keyword more times than it is present in \vec{t}_{ks} , \vec{c}_{ks} will “overshoot” the target. Multiplying \vec{t}_{ks} by a large constant solves this, with no changes in the correlations reported in Table 3.3, as θ is implemented as a cosine similarity, which is insensitive to the scale of the vector.

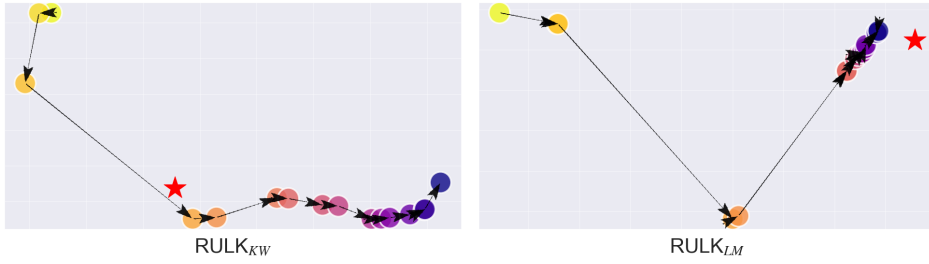


Figure 3.3: t-SNE visualization of the evolution of the \vec{c}_{k^s} vector for a learner learning about the topic “Noise-induced hearing loss” and the \vec{t}_{k^s} (depicted as a red star \star) in both $RULK_{KW}$ and $RULK_{LM}$ vector spaces

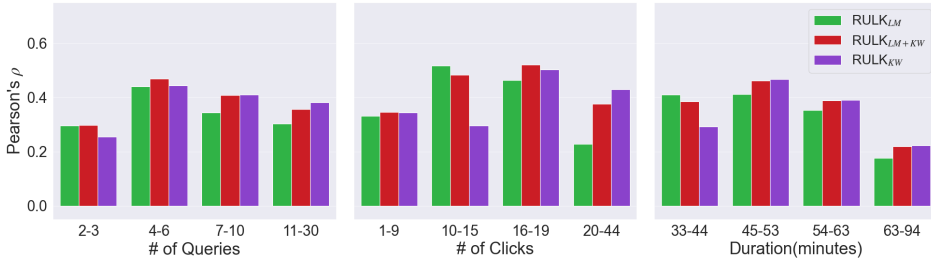


Figure 3.4: Pearson's correlations between estimated and measured (post-test scores) knowledge level, stratified by number of queries, clicks, and session duration.

RQ2: How do $RULK_{KW}$ and $RULK_{LM}$ Behave in Sessions of Different Lengths?

We can better understand how each implementation behaves as a learner session gets longer by measuring how $RULK_{KW}$ and $RULK_{LM}$ behave for learners with distinct session characteristics. In Figure 3.4, we show Pearson's correlation between the estimated knowledge level of a learner and their reported post-test scores for different quartiles of learners, stratified over the number of queries, clicks, and their session duration.

As a first takeaway, considering the length of the learner's sessions, all three implementations of RULK show a similar trend as sessions grow longer. The correlation between estimated learners' knowledge and post-test scores decreases as learners spend more time searching. In this last quartile, the Pearson's ρ of $RULK_{KW}$, $RULK_{LM}$ and $RULK_{KW+LM}$ is of 0.223, 0.176 and 0.2192, respectively. However, this same sharp drop in correlation is not observed when considering the number of queries or documents clicked. The implication is that, as learners spend more time searching, the linear behavior of σ is not enough to capture the complexity of the knowledge state of a learner.

Another interesting finding is that, while $RULK_{KW+LM}$ overperforms $RULK_{KW}$ and $RULK_{LM}$ in the set of all learners as seen in Table 3.3, this is less consistent when stratifying the learners. $RULK_{KW+LM}$ consistently overperforms both $RULK_{KW}$ and $RULK_{LM}$ when both methods have somewhat similar performance. For instance, when learners submitted less than 11 queries or when they clicked between 16 and 19 documents. However, if the discrepancy between $RULK_{KW}$ and $RULK_{LM}$ is large, such as in the second quartile of documents

clicked, $RULK_{KW+LM}$ shows a performance between the two methods. This result shows that, while one approach is better in some scenarios, an approach considering multiple features, such as $RULK_{KW+LM}$, leads to a more robust implementation of RULK.

Directly answering our research question, both $RULK_{LM}$ and $RULK_{KW}$ see a drop in their ability to estimate learners' knowledge as their session grows longer (i.e., more queries, clicks, or a session with longer duration).

However, this drop in quality is more pronounced for $RULK_{LM}$ than $RULK_{KW}$. This difference indicates that, as the sessions grow longer and the users click on more documents, issue more queries, and read for longer periods, $RULK_{KW}$ can better handle the increasing magnitude of \vec{c}_{ks} .

3.5 Caveats and Limitations

This chapter proposes a framework for tracking a learner's knowledge stage throughout their search session and estimating the learner's knowledge state at the end of their session based on that information. Therefore, we made several simplifying assumptions in our implementations of RULK. This section discusses these assumptions, motivating future works to improve upon them.

Modeling the Learner's Knowledge throughout the Session

In the opening of this Chapter, we stated that one of the main objectives of RULK is to be a framework that allow researchers to track how a learner's knowledge evolve as they explore documents during their search session. However, mainly due to limitations in data availability, we do not directly measure the correlation between the estimated and the real knowledge at different points in their session. Rather, we mainly focus on how the estimations of RULK correlates to the learner's knowledge state at the end of their search session.

While we do discuss (and show in Figure 3.3) how a learner's knowledge is tracked throughout their session, we hope that future works, ideally with user studies crafted with this goal in mind, can measure the learner's knowledge state at different points in their search session, and validate how a method inspired by our RULK framework can accurately track the learner's knowledge state throughout their session.

Documents Contents

Our implementations and experiments rely only on the textual content of the documents. However, the RULK framework is not necessarily limited to text. Any content represented in an embedding space that can be compared against a "target knowledge state" can be used, given a suitable implementation of γ . As an example, our subsequent work El Zein et al. [43] uses a knowledge graph to represent (i.e., "embed") documents and the \vec{t}_{ks} .

Another possibility is incorporating multimedia content, such as videos, in γ . It has been shown that videos positively correlate with knowledge gains [37, 54, 83, 125] in a SAL context. Combined with recent works using LLMs that successfully generate textual descriptions of audio-visual content from videos [126, 127], we believe that this is a promising avenue for future works extending RULK.

Knowledge Acquisition

When implementing the σ in both settings, one of our assumptions is that learners assimilate all the knowledge encoded in \vec{v}_d and that no forgetting occurs. While this is a common assumption [57, 67, 85], However, previous work has shown that many factors, such as the time the learner spends on the document (i.e., dwell time) [28, 87, 128], the document complexity [25, 86, 112], and the learner’s familiarity with the topic [18, 36, 59, 129, 130] may have an impact on how much of its content is acquired. Future works expanding RULK should incorporate these factors into RULK by changing Equation 3.2 to consider these factors.

Another assumption we made is to consider that the user has no previous knowledge about the topic. (i.e., we initialize \vec{c}_{ks} as a vector of 0s). However, it has been observed that users have at least some knowledge about the topic they are searching for [59, 62, 85, 87]. This prior knowledge could be incorporated into RULK by initializing \vec{c}_{ks} with a vector representing the user’s prior knowledge, potentially by using the learners’ VKS results on their pre-tests to infer this starting point.

Estimating the Learner’s Knowledge Level

RULK relies on the fact that learners’ knowledge levels can be represented as an embedding in a latent space. We do not, however, make any assumptions about such spaces. While we implemented σ as a cosine distance between two embeddings, we envision future research that applies more complex estimations of the learner’s knowledge level. For instance, works that use machine learning methods with learner behavior features, such as the ones by Yu et al. [53] and Liu et al. [131], or web-level features, such as Yu et al. [112]. Incorporating this information into the latent space for embedding the learner’s \vec{c}_{ks} and \vec{t}_{ks} could lead to more accurate estimations of the learner’s knowledge level.

The idea behind RULK is somewhat similar to some concepts from the Knowledge Tracing (KT) literature [113, 132, 133]. Under the KT setting, systems try to predict a learner’s performance in future questions by modeling how their knowledge evolves, updating that knowledge as learners answer questions or exercises during a period that usually spans a few months (e.g., a school semester). Modern KT approaches leverage methods such as Transformers-based neural networks [134–137] and Graph representation methods [138, 139].

However, KT approaches are not directly applicable to SAL. KT methods rely on many questions answered by learners over longer periods, unlike the comparatively shorter period encountered in a typical SAL session of no more than one hour. For instance, some of the most commonly used datasets for KT are the ASSISTments datasets [140, 141]. These datasets comprise grade-school students’ math exercises and answers collected during a school year. However, the *content* learners have read or been exposed to while learning is absent. Such content is crucial for the SAL setting, where the primary assumption is that learner knowledge is acquired by searching, clicking on documents, and reading them. Therefore, while KT techniques and ideas could be applied to SAL, this transfer is not straightforward.

3.6 Conclusions

In this chapter, we introduced RULK, a framework for **R**epresenting **U**ser **L**earning and **K**nowledge, containing three main components: The feature extractor γ generates embeddings from the documents read by users. The updater σ maintains a vector representing the user’s knowledge, and the estimator θ estimates the learner’s knowledge level during their search session. We hope RULK can pave the way for SAL-oriented search systems that benefit from incorporating this data, like ranking functions that can better find documents according to the user’s current knowledge and adaptations to the user interface to support the learner better.

To demonstrate how future works can instantiate RULK, we implement two example variants: RULK_{KW} and RULK_{LM} , each relying on a different method for knowledge representation. While the former uses extracted keywords, the latter uses a transformers-based language model to generate semantic representations of texts. Through our experiments, we show that both implementations can, to a certain degree, estimate the actual user knowledge level, with RULK_{KW} leading to a slightly higher correlation with self-reported knowledge and a combination of both ($\text{RULK}_{\text{KW+LM}}$) leading to an even higher correlation. We hope that our framework can be helpful for researchers aiming to incorporate users’ knowledge in search systems, primarily when focusing on learning settings (i.e., SAL), and that our work can spark discussion on how to estimate and track user knowledge.

This chapter also directly addressed our **ORQ2** from Chapter 1: “*How can we model the learner’s behavior and knowledge changes throughout their search session?*”? By studying RULK, we demonstrate a practical method of modeling a learner’s behavior.

In the remainder of this Thesis, however, we mostly do not re-use RULK. This is mainly for two reasons. First, the idea behind RULK arose later in the process of the Ph.D., as the paper that originated this chapter was originally published after the other chapters. Second, RULK is purposely high-level, as it was initially proposed as a discussion paper on DESIRES 2022 [60], and incorporating it in the other chapters would not be trivial as simple implementations, as shown by the results in Section 3.4, are not always enough to capture the full extend of a learners’s knowledge acquiring process. Nevertheless, we hope that RULK can be expanded and used in future works, as in some of our work incorporating a named entity recognition step and a knowledge graph in the framework [43].

4

Searching, Learning, and Subtopic Ordering: A Simulation-based Analysis

4

*In this chapter, we propose a novel model of searcher behavior, the Subtopic-Aware Complex Searcher Model (SACSM). This chapter continues answering our proposed **ORQ2**: “How can we model the learner’s behavior and knowledge changes throughout their search session”? Here, we propose that existing searcher models fail to capture one important aspect of SAL sessions: That a learner’s information need is not atomic, but rather, composed of multiple aspects, such as subtopics of a larger topic of interest, as explored in Chapter 2. Therefore, we propose to augment the Complex Searcher Model (CSM), an existing model, with the SACSM—modeling aspects as subtopics to the user’s need. We instantiate several agents (i.e., simulated users) with different subtopic selection strategies, which can be considered as different prototypical learning strategies (e.g., should I deeply examine one subtopic at a time, or shallowly cover several subtopics?). We also report on the first large-scale simulated analysis of user behaviors in the SAL domain. Our results in this chapter demonstrate that SACSM, under certain conditions, simulates user behaviors accurately when compared with the user data collected for our experiments in Chapter 2.*

This chapter is based on the following paper:

- 📖 Arthur Câmara, David Maxwell, and Claudia Hauff. 2022. Searching, Learning, and Subtopic Ordering: A Simulation-Based Analysis. In *ECIR*. ACM, 142–156 [90].

4.1 Introduction

Over the years, a series of models¹ that describe searcher behavior have been defined [49, 50, 142]. These often provide a post-hoc explanation of—and reasoning behind—the actions of a searcher during information seeking. One of the main drawbacks of such models is their lack of predictive capabilities: we can neither use these models to investigate what is likely to occur in different instantiations of a retrieval system; nor can we use them for simulating user behavior.

Indeed, models examining searcher behaviors with predictive power [143–145] have only recently been explored in the field of IIR. Such models enable us to relate changing costs (e.g., the cost of examining a document) to changing searcher behaviors. Prior works employing these models have investigated how searchers interact with ranked lists [146], the impact of different browsing costs on a searcher’s behavior [147, 148], and stopping behaviors on SERPs [149, 150]. Search topics are usually considered *atomic* in these works, with a simple need for information. That is, over a *search session*², a single topic is considered—with retrieved documents being either relevant or non-relevant to that one topic.

These works do not consider the different *aspects* that may constitute a wider topic. This chapter introduces the first user behavior model that incorporates such thinking. More specifically, we take as a starting point the Complex Searcher Model (CSM) [151], a model that considers a user’s interactions throughout a search session (over multiple queries), and extend it to yield the Subtopic-Aware Complex Searcher Model (SACSM)—which, by considering the aspects as subtopics of a larger information need, models: (i) subtopic selection; and (ii) subtopic switching steps in the search process.

To ground our work with the SACSM, we explore the effect of different strategies for switching subtopics for multiple types of users within the SAL domain. As a learner’s complex information needs can often be decomposed into several subtopics, a natural question is *how searchers should tackle the different subtopics to learn efficiently*.

To answer this question, we present an exploratory study of the SACSM where we simulate different types of learners as *agents*³, and compare these to each other, examining the effect their search behavior has on their ability to discover documents containing important keywords, as well as how they navigate throughout the subtopic space.

We instantiate a series of agents that subscribe to the SACSM—with four tunable parameters that control their simulated searching behavior: (i) **learning speed** (λ), or how fast agents incorporate novel terms into their vocabulary; (ii) **exploration** (ξ), or how willing agents are to explore each subtopic; (iii) **tolerance** (τ), or how willing an agent is to click on a search result snippet; and (iv) **subtopic switching** (φ), the strategy that agents employ to navigate through subtopics. As such, we present the first SAL study that employs simulation to examine the search behaviors of learners. By grounding a series of simulated agents with interaction data from the user study from Chapter 2, we run extensive simulations of interaction to address the following research question:

¹In this chapter, we refer to a *model* as a **model of user behavior**.

²We consider a *search session* as interactions with a search interface, which can include the issuing of multiple queries—and the examination of multiple documents.

³Agents are *simulated users* that can make judgments as to the relevancy/attractiveness of information *without* recourse to relevance information [152].

RQ How do **subtopic switching** (φ) strategies for learning-oriented search tasks affect the search behavior of simulated agents?

To answer **RQ**, we measure behaviors by tracking how specific measures—the *number of keywords found*, the *order of keywords found*, and *subtopic exploration*—evolve over an agent’s session. We argue that to be considered effective, a strategy should allow an agent to: (i) discover as many keywords as possible in the early stages of the session; and (ii) help the agent to complete the subtopic space exploration in as few steps as possible.

The main findings of this chapter are: (i) subtopic switching strategies that prioritize ordering in the subtopic picking process yield improved discovery of keywords and exploration of subtopics; and (ii) the SACSM is enough to instantiate agents that display behavior similar to real-world learners in a SAL context. Findings suggest that the SACSM is a high-quality model that provides a solid step in approximating searcher behaviors in the SAL domain. This is vital for works that rely on large-scale simulations, such as reinforcement learning for training new rankers optimized for human learning, as well as quickly evaluating new interfaces and algorithmic changes cheaply—all in a simulated environment.

4.2 Related Work

Models of Searcher Behavior

Models of searcher behavior typically fall into one of two categories: (i) descriptive models [19, 50, 142, 153, 154], allowing us to gain an intuition about the search process; and (ii) models that are expressed in more formal (mathematical) language [143, 145, 155–158]. The latter category of model provides *predictive power* about why users behave in a certain way. As such, they can be used as the basis of *simulations of interaction* [159]. Here, a model of searcher behavior that provides a credible approximation of reality can be used to ground simulations to examine what may happen under given circumstances.

Despite the advantages that simulations provide, formulating such descriptive models is non-trivial. Contemporary SERPs for example are complex user interfaces, with new components (e.g., *entity cards* [71]) added all the time. In contrast, searcher models typically assume a simple SERP in the format of the traditional *ten blue links* [70]. Numerous studies have been undertaken on this more simplistic design, such as the cost of scrolling [147, 160, 161], typing [162, 163] or response time lag [164, 165].

Subtopics

The IR community primarily considers the notion of subtopics from a system-centered point of view, with prior works focusing on ranking functions optimized for subtopic retrieval and result diversification [166–170]. Automatic subtopic (structure) extraction has also been investigated, generally based on a given starting query or document [171, 172]. The influence of subtopic characteristics on users has not been frequent in IIR. One exception is our previous work from Chapter 2 that provided study participants with a list of subtopics and (visual) indicators about the extent of their subtopic exploration. However, we did not study the impact of subtopic ordering on users on that chapter.

4.3 The Subtopic-Aware Complex Searcher Model (SACSM)

In this chapter, we augment the CSM [173, 174] to be subtopic-aware, turning it into the SACSM. The CSM is a conceptual model of the IIR process (or a *search session*), describing the flow of activities and decisions that a searcher undertakes when interacting with a search engine.

The CSM is built on the work of other conceptual models of the IIR process, such as the models of Baskaya et al. [157] and Thomas et al. [158]. Conceptual models provide us with the necessary scaffolding from which we can expand and develop the model further for a SAL context—and instantiate the model so that we can run our simulations of interaction [159].

We illustrate the SACSM in Figure 4.1; it includes a series of additional activities and decision points (compared to CSM) pertaining to the idea of **subtopic selection**, with novel components highlighted in **blue**. Key activities are represented as boxes □, with decision points undertaken by subscribing agents represented as diamonds ◇. Upon starting at ●, a user (or, in the case of a simulation, an agent) following the SACSM will first examine the given **topic** A. SACSM then directs the agent to examine a list of the provided **subtopics** B for the given topic, before then deciding **what subtopic** C to examine in detail. From here, the agent will **consider several potential queries** D to issue about the selected subtopic, before **selecting a query** E to issue F. The agent will then obtain an “overview” of the SERP G, and decide whether to **enter it** [149] H—and if they do, they begin to **examine a snippet** I. If the present snippet is **sufficiently attractive** J, the agent will **click the associated link** K, and **assess the document** L for usefulness and/or relevancy, before deciding to **continue on the SERP** M (and examining further snippets if so). If not, the decision to **continue with the current subtopic** N is made. If this is the case, further queries are issued E—meaning that the snippet and document examination activities are repeated for the results of the new query. This also means that subtopic exploration can entail multiple queries. If the agent decides to **abandon the subtopic** N, they must then decide whether to **stop the search session** O. This process is repeated until all subtopics have been exhausted by the agent P, or some other condition is met—such as running out of session time.

Note that compared to previous instantiations of the CSM [149, 151, 173, 174], we have removed activities and decision points about assessing documents for relevance. Unlike simple search sessions, with *atomic* information needs, a SAL task generally has a more complex and nuanced need [25]. Therefore, we are interested in examining the *content* of documents (and thus learning from them)—not simply whether the documents themselves are considered relevant, as has been the norm for prior simulations of interaction [175].

In order to keep track of terms/concepts that are examined by agents subscribing to the SACSM (as vocabulary learning is a typical manner to measure learning gains in SAL [31, 59, 62]), we must also incorporate some *state* within it. This state model was considered in the study by Maxwell and Azzopardi [151, Fig. 3] through the User State Model (USM), which “*represents the user’s cognitive state*”. Instead of representing the USM as a global, session-based model accumulating state and knowledge of the information examined, we consider a state model for the individual subtopics examined by agents. Each subtopic state consists of a representation of the terms observed by the agent to help them

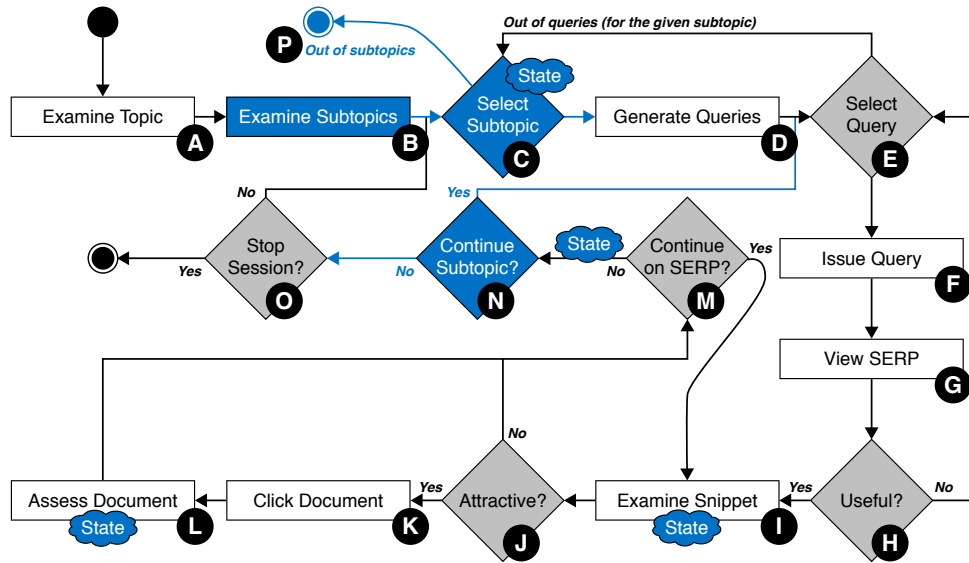


Figure 4.1: The *Subtopic-Aware Complex Searcher Model* (SACSM). Changes from the CSM are highlighted in blue. Refer to Section 4.3 for more about the sequence and shapes.

identify fundamental terms about the subtopic, which is used for query generation and determining what snippets (based on the snippet text provided by the underlying retrieval system) should be clicked on, with the corresponding document examined in more detail. Agents following this model only accrue knowledge when examining documents in full, deterministically deducing whether a document is worth examining without recourse to any relevance judgments. The state model is updated at points represented by ☁ in Fig. 4.1.

While this thesis’ chapter was initially published before the work we conducted in Chapter 3 with the RULK model, it is worthy to compare the differences between RULK and SACSM. First, the RULK embeds learners’ knowledge into latent spaces (Term frequency counter and Bidirectional Encoder Representations from Transformers (BERT)) that, at least in our implementations, are not trivially decoded into texts. As discussed in Section 4.4, this is an important feature of the SACSM, requiring the agent’s knowledge state to be translated into queries frequently. A second important difference is that the SACSM explicitly separates the agent’s knowledge state into subtopics, while RULK encodes both the user knowledge state and a target knowledge state (e.g., a Wikipedia article) into a single vector representation. While it would be possible to use RULK to track the agent’s knowledge state for each subtopic, combining these subtopic-wise states into a single vector to be compared by the RULK’s estimator (θ) would require considerable changes to the implementations proposed in Chapter 3.

Nevertheless, we believe that it is possible to combine the SACSM with RULK in future work, as both models are complementary. One possible approach would be, instead of encoding the user’s knowledge state into a single vector, to encode it in multiple textual forms, such as a collection of summaries of the pages a learner found, as related to each subtopic, and use a Large Language Model (LLM) to generate queries, while embedding

these summaries into similar latent spaces as RULK's \vec{t}_{ks} and \vec{c}_{ks} .

4.4 Experimental Method

This section describes the details of our instantiation of the SACSM and our simulations. We start by defining how we instantiate each of the components of the SACSM from Figure 4.1, dividing them between fixed (i.e., no difference between agents) and variable components (i.e., changes between each agent). We can instantiate agents that simulate users with different characteristics by tweaking the variable components. For example, an agent with a high λ (*how fast am I at learning new terms?*), low ξ (*how much content should I explore?*) and low τ (*how liberal am I at clicking links?*) simulates a learner that can quickly absorb new concepts, while only skimming through documents and clicking on almost all documents presented to them. We also outline our search setup, simulation setup, the datasets, and topics (and subtopics) used.

4.4.1 Fixed SACSM Components

We instantiate the SACSM in various ways to evaluate how different **subtopic switching** strategies performed for different types of users. Although the SACSM has many activities and decision points to instantiate, we fixed several of these to reduce the space we were required to examine.

Query Generation

We use the $QS3^+$ querying strategy proposed by Maxwell and Azzopardi [151], where three query terms are selected from a language model learned from the documents the agent has already explored (plus the topic description). Previous user studies in the SAL domain [62, 64] have shown that three query terms per query is reasonable and close to what real-world searchers use.

SERP Examination

Considered by Maxwell and Azzopardi [149], SERP examination strategies provide users with the ability to survey a SERP before committing to examining it in detail. Here, We choose to reduce the complexity of our agents (and explored space) and use the *Always Examine* approach—agents always enter the SERP and examine at least one result snippet.

User Interaction Costs

To realistically mimic how long agents should spend on each phase of their search process, we present in Table 4.1 the costs (in seconds) from the interaction data from the user study from Chapter 2. Note the high document examination cost—as participants of the user study were attempting to formulate ideas about concepts, they spent on average longer on documents when compared to other, non-SAL based studies (e.g., [149]). We also note that the total session times influence the stopping behaviors of agents since, when agents reach the time limit of their sessions, they automatically stop—regardless of the number of remaining queries to be issued, as generated by the $QS3^+$ strategy.

Table 4.1: Interaction costs grounding our agents, as derived from the data from Chapter 2.

Time required to...	Value (in seconds)
...issue a query	9.42
...examine a SERP	2.00
...examine a result snippet	3.00
...examine a document	80.00
Total session time	2400

Snippet-Level Stopping Strategies

Different *snippet-level stopping strategies* can be employed, generally classified between *fixed* (i.e., the agent will evaluate snippets until a certain depth) or *adaptive* strategies (i.e., the number of snippets evaluated may change depending on factors like agent state, presented snippet content, etc.). We use only a fixed snippet-level stopping strategy for our agents, where agents examine snippets to a depth of 10. This is a reasonable depth to examine, and avoids issues with SERP pagination.

4.4.2 Variable SACSM Components

For this study, agents can be instantiated using four variables according to the type of user to be simulated. An overview of these variables is presented in Fig. 4.2.

Subtopic Switching (φ)

We propose four different strategies for agents to select and switch between subtopics during their search sessions. These implement the **Select Subtopic** decision point, as shown in Fig. 4.1. To determine whether agents have explored a subtopic sufficiently, we use a method similar to our approach in Chapter 2 for tracking subtopic exploration. Each clicked document is embedded using SBERT [121] and compared—using the dot product—to pre-computed embeddings for each subtopic of the current topic, as extracted from their *Wikipedia* articles⁴. Therefore, each document an agent clicks will update an internal state tracker for each subtopic, summing how much the agent ‘explored’ each subtopic. We evaluate four strategies.

- **Greedy** For this strategy, an agent examines each subtopic in turn, according to the order provided by the respective Wikipedia article, only deciding to move to the next subtopic when they have achieved a certain level of progress. Intuitively, this would be the most rational type of user since they follow a subtopic ordering optimized for human understanding (i.e., the order comes from a Wikipedia page). In other words, they will attempt to master one subtopic before moving to the next (prescribed) topic.
- **Greedy-Skip** Instead of the above, an agent subscribing to **Greedy-Skip** moves to the next subtopic with the *next lowest completion value*. This instantiated agent attempts to minimize the number of documents to be read by querying in a domain with lesser knowledge.

⁴Refer to Section 4.4.3 for more information on the use of Wikipedia articles.

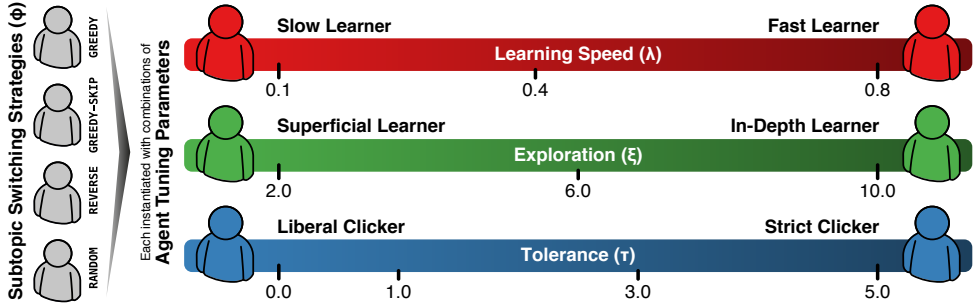


Figure 4.2: Overview of the four variable parameters for instantiating simulated agents.

- **Reverse** This strategy is similar to **Greedy**, but the agent examines the subtopics in reverse order, as presented in the corresponding Wikipedia article. The rationale here is that an agent attempts to game the system by first learning the most complex subtopics *before* moving to easier ones.
- **Random** This strategy randomly selects a new subtopic after each query, with no predefined order. This strategy models a non-rational learner and is a lower bound for our experiments.

Learning Speed

(λ) This parameter is the same λ from the language model proposed by Maxwell and Az-zopardi [151] (i.e., the Jelinek-Mercer smoothing of language models). It controls how much an agent relies on their acquired knowledge (i.e., novel terms) when considering whether or not a given snippet should be clicked. The language model is updated whenever the agent clicks on a relevant snippet. In addition, a *Maximum Likelihood Estimator* [176] is used to decide if a given snippet is attractive. An agent with a low λ gives lower weights to terms learned during the session, simulating a *slow learner*. An agent with a higher λ , in turn, mimics a user that quickly incorporates new terms, being a *fast learner*. In our simulations, we use $\lambda \in [0.1, 0.4, 0.8]$.

Exploration

(ξ) This parameter controls how much the agent should explore a subtopic before being satisfied by what it has '*learned*'. A lower number implies that such an agent is only '*skimming*' through the topic, inspecting only a few documents per subtopic. In contrast, a higher value implies that such an agent is willing to explore deeper within each subtopic. We trial $\xi \in [2.0, 6.0, 10.0]$ for the simulations reported in this paper.

Tolerance

(τ) Finally, this parameter is the threshold that controls how attractive a snippet should be to be clicked [176]. An agent with low τ is a *strict clicker*, clicking on fewer, '*safer*' snippets, while a higher τ implies a *liberal clicker* agent, more willing to explore. In our simulations, we trial $\tau \in [0.0, 1.0, 3.0, 5.0]$.

Table 4.2: # of subtopics and distinct keywords (**KW**) for each topic. We determine the ten KWs with higher TF-IDF for each subtopic on the respective subtopic section on Wikipedia. A KW may appear in the top ranks of several subtopics. KW difficulty is given by the age-of-acquisition, as proposed Kuperman et al. [177].

Topic	#Subtopics	#Unique KWs	KW Difficulty
Ethics	6	49	10.85
Genetically Modified Organism	5	33	9.97
Noise-Induced Hearing Loss	8	56	8.85
Subprime Mortgage Crisis	8	52	9.81
Radiocarbon Dating	4	35	9.77
Business Cycle	4	32	10.70
Irritable Bowel Syndrome	10	72	9.88
Theory of Mind	8	67	9.63

4.4.3 Simulation Setup

The setup of our experiments follows that of the user study we presented in Chapter 2.

In this chapter, we use the same eight topics extracted from the TREC CAR 2017 dataset [102], as shown in Table 4.2. Subtopics were also derived from the TREC CAR dataset: they were extracted from first-level headings of the respective *Wikipedia* articles as they were in December 2017—the dataset’s creation.

Our study uses the *Bing Search API* to provide a ranking for queries issued by real-world users and our simulated agents. We used a manually curated blocklist⁵ of URLs serving *Wiki*-style clones to filter results returned from the Bing API to prevent agents from encountering a single page that would give them all the information on all subtopics at once. This encourages agents to examine multiple documents and issue queries to find information pertinent to their learning task. Ten results per page were presented to agents to match our stopping strategy (see Section 4.4.1).

4.5 Results

By combining all values of ξ , λ , τ and φ , we instantiate 144 unique agents (using a modified version of the SimIIR framework [151]), and run each agent over all the topics shown in Table 4.2. Our version of SimIIR—and the raw outputs of our simulations—are available at https://github.com/ArthurCamara/simiiir_subtopics/. With some methods being non-deterministic, each agent was run ten times—with the average reported. In total, we ran a total of 11,520 simulations.

We show representative examples for each set of measures in Figures 4.3, 4.4 and 4.5. The x axes denote how many documents the agent examined during a search session in all plots. While values on y axes may seem low, they are averaged over many simulations with varying degrees of complexity.

Table 4.3 shows the average value for key measures over all agents of each φ over the eight topics. In the first row, we also show the measures from the FEEDBACK_{SC} cohort from

⁵<https://github.com/ArthurCamara/CHIIR21-SAL-Scaffolding/blob/master/data/blocklist.txt> (All URLs last accessed January 18th, 2022.)

Table 4.3: Overview of (average) measures across agents and subtopic switching strategies, and real learners extracted from the $\text{FEEDBACK}_{\text{SC}}$ cohort from Chapter 2.

Strategy φ	#Queries Issued	#Snippets Examined	#Documents Clicked
$\text{FEEDBACK}_{\text{SC}}(N=36)$	11.86(± 7.60)	152.44(± 84.23)	18.50(± 9.56)
Greedy	13.05(± 14.93)	133.37(± 176.29)	21.32(± 7.60)
Greedy-Skip	13.05(± 15.13)	133.23(± 177.26)	21.44(± 8.88)
Reverse	12.01(± 14.55)	123.28(± 173.34)	21.42(± 8.78)
Random	13.03(± 16.07)	117.61(± 155.80)	21.82(± 8.30)

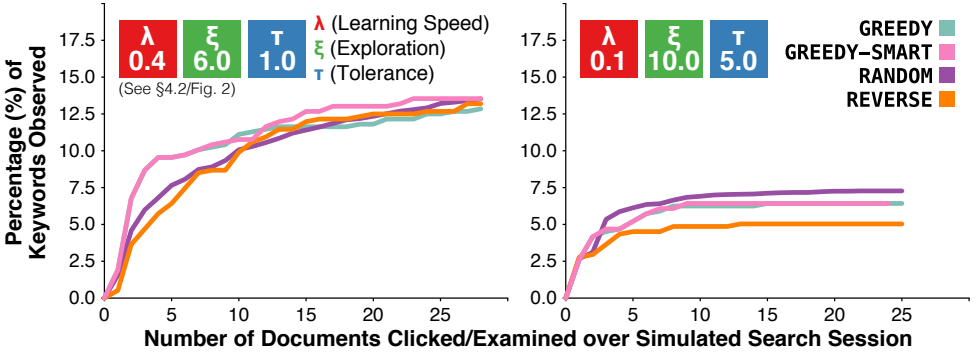


Figure 4.3: Accumulated percentage of the keywords seen for two different agents (averaged over all topics, weighted by the number of keywords) with varying ξ , λ and τ .

our user study. Our simulated agents are similar on these measures compared to how real-world learners would behave, with a similar number of queries, snippets examined, and documents clicked. While a high deviation is expected, recall that this is an average of 288 agents with a large variation in their parameters (compared to only 36 real-world learners). Results show that our agents are indeed similar to real-world learners. To address our **RQ**, we break down our analysis further into three sub-questions.

How Many Keywords can the Agents Find?

To measure how well the agents can find documents with a high concentration of potentially valuable keywords⁶, we extracted ten keywords for each subtopic from their respective paragraphs from the topic’s Wikipedia article⁷. To do this, we begin by ranking all terms from their portions of the articles (excluding stopwords) by their TF-IDF, with the IDF computed over the whole TREC CAR Wikipedia dump—and selecting the top 10 terms as keywords. We use this subtopic-wise approach (instead of extracting keywords from the whole article) to ensure a fair distribution of keywords over all subtopics, providing a less biased overview of how the agent is performing over the topic. Therefore, each topic has a different number of keywords, reflected by its number of subtopics. Table 4.2 shows

⁶As noted in Section 4.3, we do not have explicit relevance judgments.

⁷As an example, the following are extracted keywords for the topic *Ethics*: ethical, ontology, propositions, consequentialism, normative and principles.

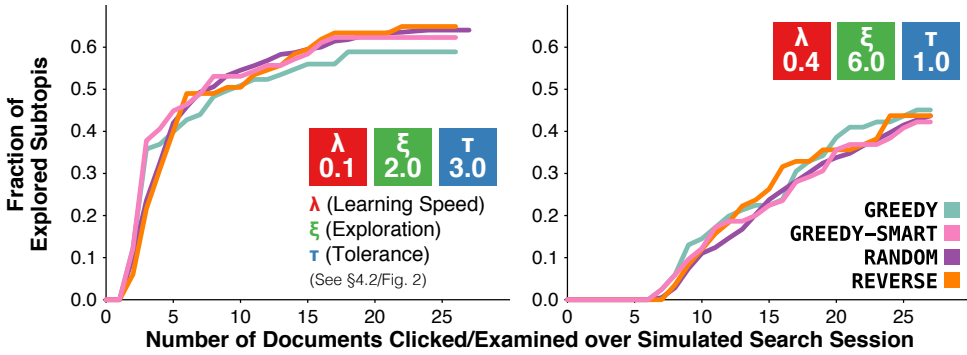


Figure 4.4: Fraction of fully explored (i.e., agent reached ξ value) for two different agents (averaged over all topics, weighted by the number of keywords) with varying ξ , λ and τ .

each topic’s many subtopics and unique keywords. This setup is similar to previous SAL user studies [59, 62, 64], where study participants were asked to define a list of concepts before and after their search session to evaluate their knowledge gain. We can mimic this setup throughout the entirety of an agent’s search session by requiring the keyword to appear at least a few times (in our case, five) in the documents ‘read’ by the agent.

For two agents, Fig. 4.3 shows how many keywords each approach for φ discovers during their search sessions after reading a certain number of documents. At the beginning of the search sessions, we observed that agents instantiated with **Greedy** and **Greedy-Skip** strategies found keywords faster than agents with **Random** or **Reverse**. However, this difference diminishes over time. This is expected since the subtopics ordering comes from Wikipedia articles, which are optimized for human understanding. Therefore, an agent that searches for subtopics in order has a higher probability of encountering documents with more keywords earlier in the session when compared with one that does not. We can also note that **Random** with higher τ found more unique keywords, given their high probability of clicking in any document.

Are the Agents Exploring Enough of the Subtopics?

Another way to measure how the agents behave is by investigating how their internal *Subtopic Trackers* evolve during the session (as explained in Section 4.3). If an agent can reach ξ for a given subtopic in a few documents, we can infer that they could quickly find documents related to that subtopic. Fig. 4.4 shows a similar trend to that observed previously, with agents using **Greedy** and **Greedy-Skip** strategies clicking on documents that advance their internal tracking faster. This implies that these strategies effectively lead the agents towards better documents faster.

Are the Agents Following the Order of the Subtopics?

While the previous measures show that the agents are effective in finding documents related to the topic, they fail to incorporate another essential learning feature, namely that keywords have dependencies. We assume that, for an agent to comprehend what a keyword means entirely, they have to comprehend at least some other, more basic concepts related to the topic at hand. Therefore, an agent that can find documents so that they will

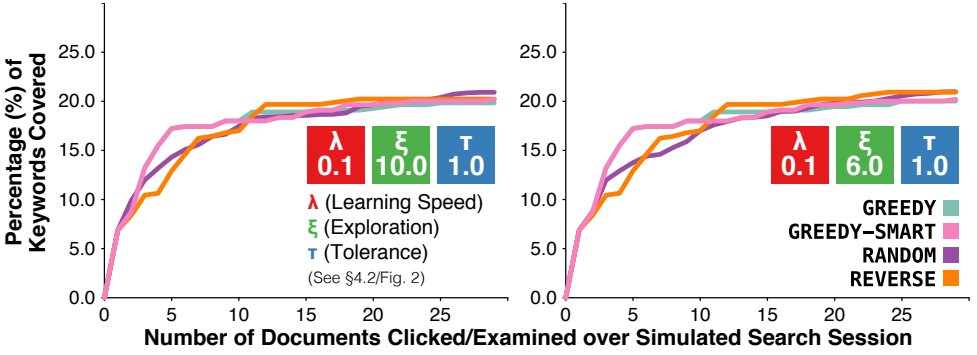


Figure 4.5: Fraction of keywords properly ‘learned’ by two different agents (averaged over all topics, weighted by the number of keywords) with varying ξ , λ and τ .

encounter more primary keywords for the topic (i.e., that appear earlier in the Wikipedia article) earlier in the session before facing more complex keywords (i.e., that appear later in the Wikipedia article) is more desirable for a SAL environment. For example, consider the keyword consequentialist for the topic *Ethics*. Before an agent can adequately understand what it means in this context, they probably need to understand other concepts, like *virtue* and *morality*. Therefore, for this analysis, we consider a keyword to be ‘learned’ after the agent has already encountered a certain number of keywords that appear before it in the original Wikipedia article. To account for possible noises in our keyword extraction method, we define this number as 50% of the keywords seen before the current one (e.g., the keyword consequentialist is the 19th out of 49 keywords to appear in the list of extracted terms for the topic *Ethics*).

Therefore, we only consider a given keyword as learned after the agent has learned at least 30% of the prior terms⁸. As seen in Figure 4.5, we see similar behavior to the one observed above, with **Greedy** and **Greedy-Skip** outperforming **Random** and **Reverse**, with the difference slowly disappearing throughout the search session. Again, almost all agents repeat this behavior.

These results show that our simulations are close to real users and that there is a clear difference between strategies, with **Greedy** and **Greedy-Skip** following the logical structure of the subtopics and generally being better strategies for agents exploring subtopics. Consequently, these should be considered when simulating agents for SAL scenarios.

4.6 Conclusions

In this chapter, we have proposed a novel user model for simulating agents focused on SAL tasks: the *Subtopic-Aware Complex Searcher Model*, SACSM. Recalling our original research question which considered how different subtopic switching strategies (ϕ) affected the behavior of simulated agents, we show that strategies that mimic a rational user (i.e., **Greedy** and **Greedy-Skip**) are more effective at *finding keywords*, *exploring subtopics* and *following subtopic structure* when compared to other strategies. With 11,520 simulations,

⁸This number was decided experimentally, as it was the best to distinguish between the different trial methods.

our study is the first (to the best of our knowledge) that focuses on simulated agents for Search as Learning, enabling future works in both SAL and IIR that may require large quantities of user data, such as Reinforcement Learning models and studies on how changes in the search system may impact the behavior of learners. To further help research efforts, we also make public our implementation of the SACS_M, built on top of the already established SimIIR framework.

Finally, this chapter directly addresses **ORQ2**, as defined in Chapter 1: “*How can we model the learner’s behavior and knowledge changes throughout their search session?*”? By simulating different types of learners, with different sets of parameters, we show that the searcher model of SACS_M can not only simulate learners with similar characteristics as real learners, compared to data from our user study from Chapter 2, but also simulate distinct styles of learners, allowing for mimicking of how learners interact with search engines during their learning sessions.

5

Keep KALM and Search On: Unveiling Causal Connections in Search-as-Learning

5

*Most recent SAL works, including other chapters in this thesis, focus on drawing correlations between single metrics and learning outcomes. While this is a simple way to pinpoint the impact of specific learner behaviors on their learning outcomes, its simplicity fails to capture the complex causal relationships in the SAL process. Therefore, this chapter answers **ORQ2**, How can we model the learner's behavior and knowledge changes throughout their search session, by modeling these relationships with KALM, a causal model for knowledge acquisition. Given its suitability for small, non-normal samples and predictive nature, we employ partial least squares structural equation modeling (PLS-SEM), a multivariate analysis technique, to analyze KALM on three user datasets.*

This chapter shows that, for different user cohorts, different variables have distinct impacts on their knowledge gains. While query quality is the main driver for learning outcomes in a broader population, undergraduate-level learners' exploration effort and essay' quality have the biggest impact. The implication is that SAL researchers should carefully consider their audience when designing learning-oriented search systems. Finally, we also discuss how relying only on multiple-choice questionnaires provides an incomplete view of the learning process, yielding a lower explained variance in learners' knowledge when compared to combining it and open-ended essay writing.

Our findings provide a nuanced understanding of knowledge acquisition during search, laying a foundation for future causality research in IIR. We believe that more causal-oriented studies like ours can drive more rigorous investigations in SAL, ultimately leading to more principled and effective interventions to enhance learners' experience in search systems.

5.1 Introduction

As discussed in Chapter 1, one of the main driving forces behind SAL research is the discrepancy between general-purpose search engine optimization and the actual learning process [67]. Traditional search engines are generally designed for ad-hoc search, assuming that users encapsulate their entire information need into a single natural language query¹. Therefore, such systems try to maximize the relevance of search results for this singular query. However, this assumption is not always consistent with the behavior of *learners*, who often have complex information needs that may require multiple rounds of interactions, where learners engage with the search system over longer periods, submitting multiple queries and interacting with multiple documents [25].

As a result, many studies have explored the connections between metrics derived from users' search interactions, such as the number of queries submitted and time spent reading documents, and some Knowledge Gain (KG) metrics. These studies mostly explore metrics derived from learner interactions with the search system and how they are *correlated* with such knowledge gain metrics [18, 36, 59, 77, 87, 128–130, 178–180].

However, some key issues persist. First, identifying suitable metrics to measure user learning is far from trivial. In a traditional classroom setting, instructors may grade tests or essays to assess individual student learning. These methods are not feasible in large-scale environments, like SAL. Here, learning metrics need to be *scalable*. We need metrics that require *no* human intervention to enable tracking of the learners' knowledge throughout their search session, as discussed in Chapter 3. And, at the end of their sessions, metrics with *minimal* or no human intervention, so evaluating learners' knowledge acquisition becomes feasible even on a larger scale.

To solve these concerns, researchers traditionally rely on metrics derived from the difference in scores between a multiple-choice questionnaire given to learners before and after their search session. These tests are generally based on a set of topic-related questions developed by experts or based on VKS scores, where learners self-report their degree of familiarity with some topic-related keywords².

The second issue with current SAL research arises because, while many prior works have successfully identified metrics *correlating* with users' knowledge gains, correlation does not necessarily imply causation. This means that these studies only provide researchers with a partial picture. Without causality, our ability to infer how manipulating an independent variable (e.g., the quality of documents retrieved) might affect a dependent variable (e.g., an increase in the user's post-test score) is limited.

To address this gap, we here depart from the traditional correlation-focused works in the field and look into the *causal* relationships between users' search interactions with the system and their knowledge acquisition during SAL sessions.

Our goals here are threefold: First, to explore to what degree knowledge gain metrics derived from multiple-choice questionnaires can be explained by the learners' actions while searching. Second, to understand if other metrics can provide a more comprehensive view of learning. Finally, to establish causal relationships between different facets of learners' interactions and their post-session knowledge gains.

¹With the rise of LLMs and conversational search, this is quickly evolving.

²The survey by Urgo and Arguello [42] covers 40 papers containing SAL studies, with over 25 of them using either of these approaches

Specifically addressing the first goal, we are interested in metrics that retain the scalability of multiple-choice questionnaires, with no (or minimal) human intervention during the assessment phase. For this, instead of relying exclusively on multiple-choice questions, we look into short essays written by learners at the end of their sessions. We use two simple metrics that require no manual human assessment: the number and the density of relevant vocabulary terms in these essays.

Therefore, the two main research questions we aim to address here are the following:

1. Can changes in KG, measured by metrics derived from the learners' answers to multiple-choice questionnaires, be causally explained by their search behavior?
2. What causal relationships exist between search interactions and knowledge acquisition in SAL, and how strong are they?

To address these questions, we introduce and analyze a novel causal model for knowledge acquisition in SAL, providing researchers with a comprehensive framework for investigating the interplay between various latent variables (LVs) within the search process (e.g., the effort a learner puts into exploring and the quality of submitted queries). This model, named Knowledge Acquisition Learner Model (KALM), not only enhances our understanding of how different behaviors influence users' knowledge gains but also shows how incorporating essays in the learners' evaluation significantly improves our ability to understand their learning gains.

To validate KALM and measure the strength of causal relationships between latent variables, we employ partial least squares structural equation modeling (PLS-SEM). PLS-SEM is a versatile multivariate analysis technique well-suited for causal modeling in complex scenarios like those encountered in SAL research [181–184]. Unlike traditional covariance-based structural equation modeling (CB-SEM), which focuses on minimizing the difference between measured and estimated covariances of the model's variables, PLS-SEM is an “explainer” model. That is, it aims to maximize how much of the variance of dependent variables (e.g., learner's knowledge gain) is explained by the independent variables in the model (e.g., the quality of the documents found by the learner) [181].

We argue that this “explainer” facet, combined with its lower sensitivity to sample size [183] and non-normal data distributions [185], makes it exceptionally suitable for exploring and validating complex causal relationships between learners' behaviors and their knowledge acquisition. By employing PLS-SEM, we assess the validity of KALM and leverage it to answer our research questions, looking into three datasets derived from previous SAL user studies with distinct characteristics. Furthermore, we illustrate the utility of PLS-SEM as a valuable tool for SAL researchers, hoping to further encourage its use.

Following the research questions above, this chapter has the following main findings: (i) Multiple-choice questionnaires do not fully capture the intricacies of knowledge acquisition. (ii) A mixed approach, combining multiple-choice questionnaires and simple essay-based metrics, better captures the complex knowledge-acquisition process while being more causally connected to other metrics from the user's interactions with the search system. (iii) When studying an heterogeneous population of learners, query quality is the latent variable that better explains their learning gains. However, for higher education

learners, their effort in exploring documents is the main predictor for their learning gains, mediated by their essays' quality.

5.2 Related Work

Measuring Learning

One issue, recently discussed in-depth by Urgo and Arguello [42], is *how* to measure learning. As discussed in that work and others [65, 186, 187], many approaches for measuring learning exist. Frequently, researchers rely on the difference in assessments between a knowledge test before and after a learner search session and assume that any difference between these two values is due to knowledge acquired while searching.

In their work, Urgo and Arguello [42] outline nine learning assessments that may co-exist in the same study. Specifically, in studies with (i) **self-reported** assessment, researchers ask learners to directly report how much they have learned [23, 28, 32, 35]. With (ii) **implicit measures**, researchers try to predict the learner's knowledge based on their search behavior [188]. In a study with a (iii) **multiple-choice questionnaire**, learners answer multiple-choice questions potentially created by experts [53, 68, 84, 87]. For (iv) **short-answers assessments**, learners answer open-ended questions with short and objective answers [40, 59, 62, 83]. In (v) **free-recall** studies, participants should list terms and phrases related to concepts and ideas from the learning topic [34, 63]. For (vi) **sentence generation**, learners should create sentences using a set of vocabulary terms. Learning is assessed by the correctness of these sentences [189]. (vii) **Mind mapping** is another technique for assessing learners' knowledge gains. By asking participants to build mind maps, researchers can estimate how this map grows and how many concepts are added [35, 58]. An (viii) **argumentative essay** requires participants to write a long-form essay defending a position regarding a potentially controversial topic. By assessing the thoroughness of the arguments, researchers can estimate the learners' knowledge of the topic [190]. Finally, writing a (ix) **summary or an open-ended essay** is a common way to measure learners' knowledge gains. Learners should summarize what they know about the topic or answer a question using a long-form essay [13, 28, 36, 84, 191].

Despite the variety of methods for assessing learning outcomes, Urgo and Arguello [42] report that the two most common forms, **summary or open-ended essay** and **multiple-choice questionnaire** are used by 40% and 38%, respectively, of the studies covered in that work³.

Therefore, in this chapter, we employ three different datasets for assessing the causal relations between multiple facets of learning. All of them use *multiple-choice* questionnaires for assessing user learning, with Scaffolding, from our previous user study, conducted for Chapter 2, also measuring learning with *short answers* and a *summary*. Lighting, from Otto et al. [13], also uses an open-ended *essay*, where learners were asked to discuss concepts they learned about during their search session. Finally, SearchWell, from Gadiraju et al. [87], uses only a multiple-choice questionnaire for assessing learners' knowledge gains.

³We disagree with their classification, however. For instance, our work from Chapter 2 is mentioned as containing short answers only, despite also having VKS questionnaires to discuss concepts they learned about during their search session.

Metrics Correlated with Learning

Previous studies have investigated the impact of user-related features, such as topic [87] and task [128] familiarity, the learner's prior knowledge of the topic [18, 36, 59, 129, 130], and level of learning-related skills of the learner [37].

As discussed earlier in Chapter 3, Bloom's taxonomy [114] and its updated version proposed by Anderson et al. [45] are commonly used in SAL studies. Related to this chapter, works that investigate different types of learning have found that a higher cognitive level is usually correlated with increased interactions by the user with the search system [77, 178–180], sometimes leading to less learning in higher-level processes [84] and with changes in searching behavior during search usually related to a change in the learning task [35, 56].

How interventions on the search system correlate with learning is also a common theme in SAL research. For instance, Demaree et al. [190] show that using a smartphone or a laptop while searching impacts user behavior but does not impact knowledge acquisition. Freund et al. [65] examined if a pure HTML interface, as opposed to one with scripts and ads, interferes with learning, showing that a cleaner interface is correlated with higher knowledge gains. Another approach is to measure how active learning strategies may lead to a better learning experience. Here, Roy et al. [40] showed that essays at the end of the search session could be improved by highlighting portions of the text while reading, a similar conclusion to Kammerer et al. [63], who showed that annotating material correlates with better learning outcomes. On the impact on user behavior, but not necessarily knowledge gains, Liu et al. [35] demonstrated that providing learners with a mind-map-creating tool allows researchers to discern better when learners change between learning tasks during their learning process. Finally, in Chapter 2, we demonstrated that providing users with visual feedback while learning changes their searching behavior, with constant feedback leading to more exploration, with no significant changes in knowledge gains.

One of the most frequent themes across multiple studies is how the searching behavior of the learner, measured by metrics ranging from number of queries issued to eye-gazing movement [13, 192–194], changes during the session or how it is correlated with an increase in knowledge. One of the most common results is that users who spend more time reading relevant documents (i.e., increased dwell time) display higher knowledge of the topic at the end of the session [28, 53, 87, 128, 129, 191]. Other metrics correlated with higher levels of learning are the textual complexity of submitted queries [28, 34, 87], the increase of novel terms in subsequent queries [25, 87] and the branchiness (i.e., how often learners revisit certain “anchor” documents) of the learners' session [25, 195]. Finally, the quality and diversity of the documents encountered by learners is also a metric commonly correlated with learning, albeit not as frequently, given the hard task of defining relevance [28, 68, 128].

Despite the myriad of works investigating how metrics correlate with knowledge gains in SAL, it is not uncommon to find conflicting outcomes. For instance, Vakkari et al. [196] differ from most studies above by finding an inverse correlation between dwell time and document usefulness. They argue that this discrepancy comes from a more laborious post-test, with a writing task instead of a multiple-choice questionnaire. Similarly, Vakkari et al. [55] showed that a higher time spent on each query affects different learning aspects. These results hint at the complex and involved process from querying to document reading to knowledge acquisition, something we study in this chapter.

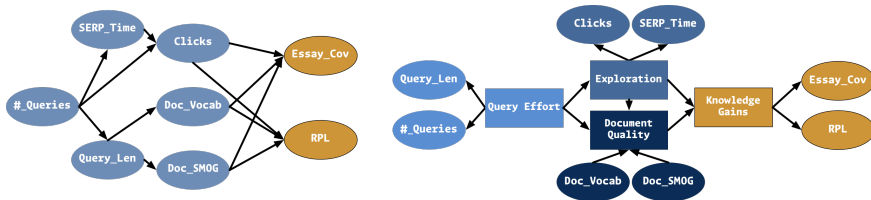




Figure 5.1: A toy example of a traditional path model (left) and a structural equation modeling (SEM) model (right). A SEM model consists of both a *measurement* and a *structural* model, while a path model only contains the structural part. In a SEM model, the measurement part describes how indicators (i.e., variables that are directly measured, here drawn as ellipses ) are related to constructs (e.g., quality of a document, here drawn as squares ). The structural model describes how these latent variables are causally related. The direction of an arrow $A \rightarrow B$ can be interpreted as “A causes B”.

Predicting Learning outcomes

While most studies cited above mainly correlated different features and metrics to learning gain, explicitly *predicting* the learning outcomes is also a recurrent theme. The intuition behind this line of research is that if we can accurately predict how much the learner has already learned, one can adapt the search engine accordingly, better serving the learners’ interests, something that we also discuss in Chapter 3 [43, 90, 197].

Works like the ones conducted by Yu et al. [53] and Gritz et al. [86] use machine learning models with dozens of features to try and predict learning outcomes. Yu et al. [53] show that a random forest model can predict knowledge gains with high accuracy, with features like dwell time and time per query being the most useful. Also using a random forest model, Gritz et al. [86] perform extensive feature selection on over one hundred features and show that the complexity of the documents read by learners measured by POS tags and verb clusters are good predictors of the knowledge gains of learners. Similarly, Otto et al. [54] explore how features extracted from documents, especially related to videos, can predict a user’s learning gain. Using 110 features, they demonstrate that the time a learner spends on multimedia pages (i.e., pages with videos) is a good predictor of their knowledge gain when combined with textual features of other documents in the learner’s session.

Using simpler, regression-based models, Guo et al. [52] trained models to predict the success of a user’s search session using fine-grained features such as cursor movements and scrolling. In a follow-up to their previous study, Yu et al. [112] trained different models for different learning tasks, using over 100 web and behavior features. Their results contradict their prior research [53], pointing to logistic regression as the best-performing model for predicting knowledge gains.

Instead of directly predicting the learners’ knowledge gain, researchers have also tried to predict related measurements, like the learner’s current state of knowledge [21] and task difficulty [198]. This type of research mainly supports the idea that a search engine should adapt its results as the learner’s knowledge of the topic evolves.

Causality in SAL

While the prospect of modeling the process by which a learner transitions from formulating queries to acquiring knowledge using a search engine is compelling, few works have

employed causal modeling to characterize this process.

One line of research, mainly spearheaded by Pertti Vakkari, is to use path analysis [199] to model the relationship between metrics and knowledge gains. Interestingly, most of these works rely not on multiple-choice questionnaires for assessing learning gains but rather on metrics derived from essays written by the learners, echoing our findings in this chapter that essay-based learning metrics should be preferred.

Path analysis tools draw correlation paths between directly observed metrics, such as the number of vocabulary terms used on queries, and dependent variables, such as learning gains. PLS-SEM, on the other hand, uses similar path analysis techniques over *latent variables*, not directly observable, but built by combining multiple observable variables. An illustration of the differences between a SEM-based model and a path analysis model can be seen in Figure 5.1.

For instance, Vakkari and Huuskonen [89] use a path model to analyze the behavior of medical students using a specialized search engine over six weeks. In that study, learning was measured by the score of an essay written by the students and evaluated by their teachers at the end of the period. The authors' main finding is that learners' effort, as measured by the number of search sessions performed by learners and the number of Medical Subject Heading (MeSH) terms "exploded" (i.e., terms from a defined, hierarchical vocabulary that learners can select to be included in the query), degrades precision (i.e., percentage of documents assessed useful). However, this effort also leads to higher essay scores. The authors hypothesize that this counter-intuitive finding is due to a compensatory mechanism by the learners. The increased effort by the learners led to a more careful examination of the documents retrieved, classifying fewer of them as useful but culminating in more structured learning (due to the MeSH terms) and a clearer understanding of the problem (evidenced by a higher correlation with problem formulating scores).

In a later work [196], the same authors propose another path model for relating the user behavior to the quality of search results. Interestingly, the authors estimate the quality of the documents a user finds by how many words from the documents are reused in the essays produced at the end of the search process.

That work shows that a longer time spent on each query (i.e., time formulating and examining the SERP produced by that query) negatively impacts the quality of the documents the user finds. However, a higher number of clicks leads to more reused words in the user's essay. Another finding in that study is that the number of useful clicks per query (i.e., the number of documents with portions of text copied into the final essay divided by the number of queries issued), their path model shows that more clicks lead to more useful clicks, but a longer dwell time per click is an indicator of less useful documents. This finding, contradicting previous works [28, 53, 87, 128, 129, 191] is attributed to differences in the task compared to previous research. In this 2019 study, the learning task mainly comprised copying parts of the documents users consider relevant and writing an essay based on these copied portions. The authors hypothesize that users could quickly identify relevant portions of the document and quickly copy them instead of spending time reading and analyzing the document's quality.

In a work that inspired this chapter, Vakkari et al. [56] used factor analysis and path models to analyze how certain search behaviors may improve essay quality at the end of a search session. They demonstrate how higher-level constructs (latent variables (LVs)),

such as query effort and click utility, may predict essay quality. These LVs are indirectly measured by directly observable variables (indicators), such as the number of queries for the query effort and the percentage of useful clicks for the click utility.

Their model, referred to here as Essay Quality Model (EQM), provides a useful starting point for developing a more comprehensive model for knowledge acquisition. It lays a solid statistical and theoretical foundation for the causal relationships between LV and learners' essay quality.

The authors divided their study participants into two groups according to their writing strategy. A "build-up" group participant gathers material aspect-by-aspect, constructing their essay accordingly. A "boil-down" group participant focuses first on finding large amounts of material, which is later edited for the essay.

The authors find that, for learners in the first group, the effort employed in the querying process, and of clicking in documents, the utility of the clicks, the volume of pastes, and the effort put into writing an essay have a positive and significant impact on the learner's essay quality. However, the second group's querying effort hurts their essay quality. The authors explain this finding by claiming that learners in the second group struggle more with writing their queries, leading to low efforts but higher diversity.

While useful as a starting point, the EQM has limitations. First, similar to their previous work [196], it relies heavily on metrics based on how often learners copy and paste content from documents. Second, the EQM does not incorporate direct measurements of the quality of queries submitted and documents read by the learner. This omission makes it challenging to design specific interventions that could improve learning gains in a search system. EQM uses simpler path analysis, which, although useful, can be less versatile in handling complex causal relationships compared to methods based in SEM.

In this chapter, inspired by these works, we aim to model the complex relationships between a learner's behavior and knowledge acquisition by defining higher-level constructs from lower-level observable metrics, similar to [56]. However, we propose expanding on the simpler path analysis concept by employing PLS-SEM.

PLS-SEM is a multivariate analysis technique for causal modeling that overcomes some of the main issues with traditional path analysis and CB-SEM. It is also more robust to two situations commonly encountered in SAL studies: small sample sizes [183] and non-normal data [185]. Additionally, it allows us to build constructs in a composite manner (also called formative constructs), where, unlike in other path analysis methods, lower-level variables can have a causal relationship to their related construct. This difference is also shown in our toy example in Figure 5.1, where the construct "Document Quality" is *causally defined* by its indicators "Doc_vocab" (i.e., the number of vocabulary terms in the document) and "Doc_SMOG" (i.e., the reading difficulty of the document). For a more thorough explanation of these topics, we provide an overview of our proposed model on Section 5.3.

As a practical example, the work by Vakkari et al. [56] defines a construct called "query effort". In their work, this construct is measured by metrics such as the number of queries and seconds spent querying. By using PLS-SEM, we can also define a construct of query *quality* that, instead of being indirectly measured by a set of metrics, is defined as being caused by indicators such as the average age-of-acquisition of the query terms and the ratio of relevant vocabulary terms present in the query.

Therefore, by leveraging PLS-SEM, our analysis in this chapter results in insights into how the different facets of this complex process of acquiring knowledge while searching are influenced by each other.

5.3 A Causal Model for Knowledge Acquisition

In traditional SAL research, the typical approach involves extracting a set of user behavior features from the usage logs of a search system, such as the number of queries issued or the time spent reading documents. This is the same approach we use in Chapters 3, 2 and 4. Then, learning is measured using straightforward metrics, such as the difference between a pre- and a post-search multiple-choice questionnaire or vocabulary test. By comparing these two data sets, we can draw correlations and speculate about the effects of certain behaviors on learning outcomes.

Although this approach has its merits and has been the prevailing methodology driving research in the field, it often oversimplifies the complexities of the knowledge acquisition process. It limits our understanding of what factors contribute to users' learning gains. Correlations do not properly capture many aspects of learning and search behaviors, especially when dealing with more complex latent variables rather than single metrics. For instance, how do we account for the quality of the queries issued or the effort put into exploring multiple documents? These factors are not easily quantifiable, but their interaction may play a critical role in learning.

Some works, such as EQM, proposed by Vakkari et al. [56] and discussed in Section 5.2, use more powerful statistical methods, such as path analysis, to try and understand the causality underlying this process. However, these are not without their problems, such as the need for large sample sizes and lack of LVs that are causally defined [181–184].

To ameliorate these issues, we propose the Knowledge Acquisition Learner Model (KALM). Similarly to EQM, KALM also models measures of users' effort in formulating queries and finding relevant documents, but it also models the *quality* of these artifacts. By incorporating a larger set of metrics and covering more facets of the search process, we aim to understand better the causal relationships between search behaviors and knowledge acquisition. An overview of KALM can be seen in Figure 5.2.

We begin this section with a brief introduction to SEM and an explanation of our rationale for using the partial least squares (PLS) variant instead of the more traditional CB-SEM method. We then describe KALM as a structural equation modeling, dividing it into two parts: *structural model* and the *measurement model* (c.f. Hair et al. [199]). While the former outlines the relationships between LVs, the latter establishes how each LV is connected to a set of measurable variables that provide a measure of the LV.

5.3.1 Structural Equation Modeling

Although SEM has been widely used in research in social sciences (c.f. Hair et al. [199]), its applications in SAL and general IIR are still incipient [88]. When engaging with search systems, learners acquire knowledge by submitting queries, assessing SERPs, and reading documents. SEM, by representing constructs (i.e., an abstract concept, such as the quality of a document) as latent variables, allows for the modeling of unobservable cognitive processes and individual differences that can impact learning outcomes.

A primary research objective in SAL, as discussed in Chapter 1, is in defining metrics that lead to higher knowledge gains during a search session [53–55]. Consequently, PLS-SEM [200] is suitable for analyzing the causal relationships between learners’ search behaviors and knowledge gains. PLS-SEM focuses on maximizing the explained variance of dependent variables, making it an ideal choice for predictive analysis and identifying potential causal relationships. In contrast, the more commonly used CB-SEM emphasizes model fit and theory testing, rendering it more suitable for hypothesis testing and theory confirmation [201].

PLS-SEM’s ability to handle complex models with multiple latent variables and relationships facilitates the exploration of intricate relationships between variables and better modeling the dynamic nature of cognitive processes during search sessions. Furthermore, PLS-SEM is less sensitive to sample size [183] and non-normal data distributions [185], making it a more practical choice for studies where such challenges are encountered, as is often the case in SAL research.

SEMs comprises two main components: the structural and measurement models. We discuss how we define each of these for KALM in Sections 5.3.2 and 5.3.3, respectively.

5.3.2 Structural Model

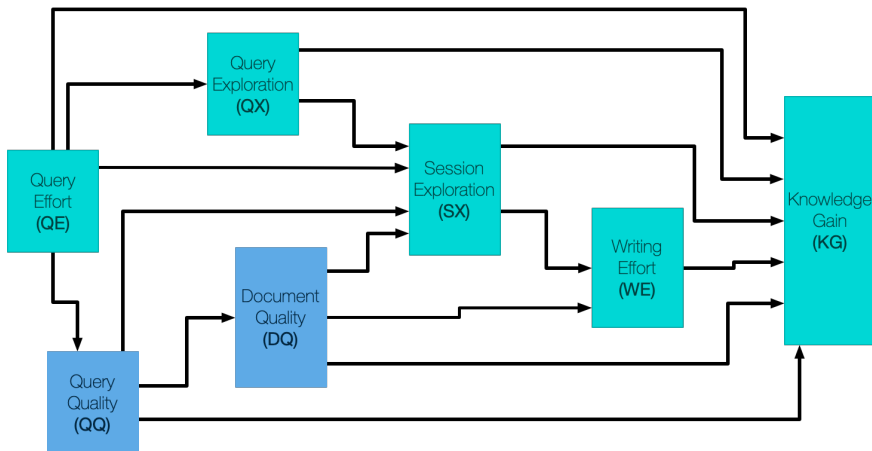


Figure 5.2: Overview of the structural model of KALM. Arrows between latent variables denote a causal relationship between them. Boxes in cyan and azure indicate reflective and formative latent variables, respectively.

The structural model outlines how each construct, represented by a LV, interacts with others. Each arrow in the model signifies a causal relationship between two LVs. Figure 5.2 illustrates the structural model for KALM. For defining what LVs to incorporate into KALM, we start from the EQM defined by Vakkari et al. [55] and expand on it by incorporating LVs related to metrics available in our datasets (c.f. Section 5.4). When defining the sequence of the LVs, we follow a natural order of query → documents → knowledge gain. Given the flexibility of PLS-SEM and its exploratory nature [200], there is room for experimen-

tation on how to refine further the LVs defined here. However, as shown in Section 5.5, KALM yields good validity for the datasets studied in this chapter. Therefore, KALM has the following latent variables:

- **Query Effort (QE):** Measures the learner's effort in their querying strategy. It considers metrics such as the total number of queries submitted by the learner and the time spent formulating queries.
- **Query Exploration (QX):** Gauges the amount of exploration the learner performs for each query. Here, metrics such as the number of documents clicked per query, the maximum SERP depth clicked, and the average time spent exploring each query are used.
- **Session Exploration (SX):** Measures how much exploration the learner performed during their total search session. This is determined by metrics such as the number of documents read, queries that resulted in a click, session length, etc.
- **Writing Effort (WE):** Considers the amount of effort the learner puts into writing their essays at the end of the session, measured by the length of the essay, number of essay terms originating from documents encountered by the learner, etc. Recall that we use these essays as one of the main components when assessing learner's knowledge gains and, therefore, hypothesize that this LV will have a sizeable impact on learners' knowledge gain.
- **Knowledge Gain (KG):** An endogenous (or dependent) variable that quantifies the amount of knowledge a learner has acquired by the end of their search session.
- **Query Quality (QQ):** Assesses the quality and diversity of each query submitted by the learner. This includes aspects like query term diversity and usage of vocabulary terms on a per-query basis⁴.
- **Document Quality (DQ):** Evaluates the quality of the documents found by the learner, covering aspects such as textual complexity, readability, and the presence of specific vocabulary terms.

It is worth noticing that the influence of each construct on a learner's knowledge acquisition varies across studies. As highlighted by Vakkari [202], there are discrepancies in the magnitude and even the signal (i.e., positive or negative) of the impact of variables. For instance, Mao et al. [203] reported a negative correlation between the number of documents visited by a learner and their learning gain, while Guo et al. [52] found the opposite. Such discrepancies may arise due to differences in experimental design, populations, or contexts across studies.

Despite these inconsistencies, the logical causal sequence from queries to documents and finally to knowledge gains is well-established. In this sequence, the queries submitted by the learner guide the search system in retrieving relevant documents, and the interaction with these documents subsequently contributes to knowledge gains.

⁴c.f. Section 5.4 for a discussion on how vocabulary terms are selected.

5.3.3 Measurement Model

The measurement model in SEM links LVs to measured variables (MVs) [204]. An LV is usually an abstract concept difficult to directly measure (e.g., Query Effort (QE)). In contrast, an MV, when connected to an LV, serves as an observable *indicator* providing a quantifiable value to the LV (e.g., the number of SERPs visited for each query)⁵.

The relationship between an LV and its indicators can be either *formative* or *reflective*. In reflective relationships, the MVs are assumed to measure the underlying LV equally well and be highly correlated; changes in the LV induce changes in all associated indicators. This model aligns with the one used in CB-SEM. Conversely, a formative model, also known as *causal-formative*, assumes that the MVs, when combined, determines the variability in the LV, effectively inverting the relationship.

Keeping this distinction in mind, we classify two of the LVs in KALM—DQ and QQ—as formative. This decision stems from the fact that both DQ and QQ are multifaceted constructs that incorporate various non-interchangeable MVs. These constructs necessitate a mix of diverse indicators to capture the quality of a document and a query [205].

Furthermore, these constructs—DQ and QQ—are primary intervention targets in SAL studies. Strategies to enhance the quality of retrieved results (e.g., retrieving documents with more terms related to the learner’s goal, as proposed by Syed et al. [97]) and to increase query diversity (e.g., encouraging usage of novel vocabulary terms, as suggested by us in Chapter 2) are common in SAL research. As such, establishing a causal relationship between specific indicators and these latent variables could inform future studies and help interpret past results.

5.4 Datasets

Table 5.1: Comparison between the three datasets analyzed in this study. Statistics are computed *after* filtering of participants with no queries, clicks, visits, or that interacted with documents not in English (for Scaffolding and SearchWell) or German (for Lightning).

	Scaffolding	Lightning	SearchWell
Participants	120	110	235
Topics	7	1	11
Source of participants	Crowd-sourcing (Prolific)	Undergraduate students	Crowd-sourcing (Mech. Turk)
Session length	42m55s ± 12m4s	24m36s ± 6m45s	6m15s ± 6m32s
Queries	9.37 ± 6.6	4.1 ± 3.0	2.02 ± 2.2
Documents Visited	21.55 ± 11.5	25.6 ± 21.9	2.55 ± 1.9
Multiple-choice?	✓	✓	✓
Essays?	✓	✓	✗
Avg. ALG	0.55 ± 0.4	0.2 ± 0.2	0.3 ± 0.2
Avg. RPL	0.29 ± 0.2	0.0 ± 0.0	0.5 ± 0.2
Vocab terms per topic	10	10	16.2
Avg. Essay length	132.26 ± 58.0	200.9 ± 73.4	—

⁵As in most PLS literature, we use the terms latent variable (LV) and “construct” interchangeably, as well as measured variable (MV) and “indicator”.

To ensure that our results are consistent across multiple scenarios, KALM should be capable of explaining causal relationships and identifying general trends in more than one search context. Therefore, we experiment with KALM in three distinct datasets of past SAL user studies, including our work of Chapter 2. This approach of using secondary (or archival) data when exploring PLS-SEM gives us a better and more effective interplay between theory and the available data, at the cost of a potentially worst model fit when compared to data collected for validating an existing model [206].

Therefore, we utilize three different datasets in our experiments, each with distinctive characteristics: *Scaffolding*, from Chapter 2, *SearchWell*, from Gadiraju et al. [87] and *Lightning*, published by Otto et al. [13].

While the datasets may vary considerably across multiple dimensions, this selection allows us to explore how the interactions between variables may differ in the different conditions of each user study. Table 5.1 highlights some differences across the three datasets.

One main difference is how each dataset recruited learners for their user studies. While *Scaffolding* and *SearchWell* recruited participants from crowdsourcing platforms (Prolific and Amazon Mechanical Turk, respectively), *Lightning* relied on undergraduate students. This distinction may play a considerable role in the result, as crowdsourcing platforms are known to have a considerable proportion of low-quality participants [207]. Additionally, the discrepancy in educational background may contribute to variations in their previous knowledge and learning potential. Finally, this difference also plays a role in the setting of each study. While crowd workers performed their searches on personal desktops, factors such as attention level or external distractions are not controlled in differing scenarios. Meanwhile, participants in the *Lightning* dataset performed their search sessions in a controlled lab setup.

Another distinction is in the selection of topics. Participants from the *Scaffolding* dataset were asked to learn about one of seven topics. Participants conducted a pre-test questionnaire on two random topics, and the one with the lower pre-test score was selected. As for *SearchWell*, participants were randomly assigned to one of ten topics. In the case of *Lightning*, all participants were required to search about the same topic. These differences may also influence results, as topics may have considerable difficulty differences.

Regarding session duration, *Scaffolding* required participants to search for at least 30 minutes before taking a post-test, while the other studies set no minimum duration. Conversely, *Lightning* set an upper time limit of 30 minutes, and *SearchWell* imposed no time constraints. On average, sessions from *Scaffolding* participants are 7 times longer than *SearchWell* participants, with 4.5 times more queries submitted and 8 times more documents. Participants in *Lightning* took 4 times longer in their session, submitted 2 more queries, and read 10 times more documents than in *SearchWell*.

Language-wise, participants in the *Scaffolding* dataset were fluent English speakers, while *Lightning* recruited German-speaking students. *SearchWell*, on the other hand, did not impose any prior language restrictions. To avoid significant discrepancies within the same dataset (e.g., the number of tokens in a query written in English or Russian can vary drastically), we excluded participants from *SearchWell* who issued queries in languages other than English or interacted with non-English documents. This further reduced the number of participants from 420 to 235.

All three datasets measure learning gains using a set of multiple-choice questions that

participants answer before and after the study. While Scaffolding focuses on a set of relevant keywords extracted from Wikipedia, Lightning reuses questions from a previous study on multimedia learning [208] and SearchWell employs the list of questions from the *TREC 2014 Web Track* dataset. Participants in Scaffolding and Lightning were also asked to write a short essay on what they have learned during their search session. Originally, these essays were not evaluated in the Scaffolding dataset, but rather used as a sanity check to verify whether their answers were coherent and filter potentially bad actors from the pool of participants. However, the authors from Lightning manually evaluated all the essays, scoring them on a scale from 0 to 10 on how many relevant concepts participants correctly defined in the essay.

As mentioned in our second research question, we want to understand if metrics based on vocabulary usage in essays can assess learner knowledge gains better. We also use the same vocabulary terms in some metrics related to the quality of the documents and queries issued by the learners (c.f. Tables 5.2 and 5.3). Therefore, we necessitate a vocabulary of topic-relevant terms for each topic.

For the Scaffolding dataset, we employ the list of keywords used in the original study for the vocabulary test. As for Lightning and SearchWell, we utilize YAKE [115] to extract salient unigram and bigram keywords for the questions used in pre- and post-tests questionnaires in these datasets. We then manually filter these to eliminate duplicates and lower-quality keywords and to keep at least one term per question.

The documents visited by participants were retrieved using the WayBack Machine API⁶, timed as closely as possible to the date reported in each respective study. In cases where URLs were not listed in the Web Archive index or were no longer available, they were omitted from the dataset. To replicate the original formatting of the pages as closely as possible, we employed Selenium⁷ for page rendering. We extracted the text using the `jusText` Python library⁸.

From all datasets, participants were excluded if they did not submit any queries, did not click on any SERP links, did not visit any documents, or if we could not retrieve any of their visited documents (i.e., the visited URLs were not accessible on the Web Archive). Table 5.1 showcases key statistics for these refined datasets.

Finally, a note on data availability. The same set of indicators was extracted for all three datasets whenever possible. However, as the datasets used here were not primarily collected to test an existing SEM model, some metrics may not be available for individual datasets. For instance, information on bookmarked documents is only provided by Scaffolding, while only Lightning provides a pre-test essay and scores for all participants' essays. The comprehensive list of keywords for each dataset, as well as the full set of processed logs, HTMLs, and extracted texts, is available on GitHub⁹.

5.5 Assessing KALM

The evaluation criteria for CB-SEM and PLS-SEM are fundamentally distinct. CB-SEM aims to minimize the discrepancy between observed and estimated *covariances* in both

⁶https://archive.org/help/wayback_api.php

⁷<https://www.selenium.dev/>

⁸<https://github.com/miso-belica/jusText>

⁹<https://github.com/ArthurCamara/cauSAL>

the dataset and the proposed model. This enables quality to be directly assessed through various χ^2 -based metrics. On the other hand, PLS-SEM's primary goal is to maximize the explained *variance* of dependent variables—in our context, the KG experienced by a learner after their search session. This necessitates a separate evaluation of each model component to gauge its predictive performance for each dependent variable and construct.

In assessing a PLS-SEM model like KALM, our first task is to evaluate the measurement model for each construct to ensure its validity. This entails analyzing the relationship between the indicators and the constructs, which requires a separate evaluation of both reflectively and formatively measured constructs, with several scores calculated during this stage to validate each component.

Two key values of each indicator are reported here: outer loading (l_i) and outer weight (w_i). l_i measures the correlation between the indicator and its LV. It can be interpreted as the *absolute* measure of the correlation between an indicator and its construct, regardless of the other indicators. On the other hand, the w_i measures the indicator's contribution to its LV. It represents the *relative* importance of the causal relationship between the indicator and its construct, considering the influence of the other indicators.

With a validated measurement model, we then turn our attention to the structural model, or path model, of KALM. This phase of the assessment focuses on measuring the significance and magnitude of the relationships between the constructs, as well as their *explanatory* power—That is, how well the exogenous variables can explain the variance of the endogenous variable KG.

In this step, the main metric of interest is the **path coefficient**. These values describe the strength of the relationship between two LVs. They are computed as regression coefficients when regressing from the dependent to the independent LVs and are interpreted as the size of one LV's causal effect on another.

As discussed in Section 5.4, our study is exploratory and based on secondary data—archival information (i.e., from previous studies) not initially gathered to validate an existing model. Therefore, we do not anticipate achieving an exceptional model fit [206]. Rather, our primary objective, answering the research questions proposed in the introduction, is to identify variables that better reflect the Knowledge Gain of a learner during their search session in the form of better indicators for the KG construct, and to identify the variables linked to higher knowledge gains, thereby providing a foundation for future SAL research.

All of our estimates of our PLS-SEM were computed using the SmartPLS software version 4.0.9.3 [209]. Statistical significance analysis was conducted with bias-corrected bootstrapping with $K = 10000$ re-samples. As the analysis is performed on secondary data, we assume a significance level of $\alpha = 0.1^{10}$. When reporting our results, we mostly adhere to the procedures and guidelines described by Hair et al. [210] for reporting results and assessing our model's quality. We suggest consulting Hair et al. [200], specifically chapters 4, 5, and 6, for a detailed explanation of the tests and steps involved. Finally, following the inverse square root method for minimum sample size [183], assuming a power level of 80% and significance levels of 10%, the minimum required sample size is 73 participants, below the size of the smallest dataset in our collection (110).

¹⁰As the data used in this chapter was not collected specifically for this study, we allow for a slightly higher significance level than otherwise [206].

5.5.1 Assessing the Reflective Measurement Models

When assessing the validity of the reflective measurement models, we evaluate *indicator reliability*, *internal consistency*, *convergent validity*, and *discriminant validity* for each construct. Recall that a reflective construct is defined by a set of indicators that are highly inter-correlated, consistently measuring the same latent construct. Table 5.2 presents the results of the reflective measurement assessment for all datasets.

First, *indicator reliability* is assessed using the outer loading for each indicator. Ideally, all indicators should exhibit high and statistically significant l_i . When $l_i > 0.7$, the construct can explain more than 50% of the indicator’s variance.

Next, we examine *internal consistency*, which measures the intercorrelation among the indicators. This is quantified using the reliability coefficient (ρ_a) [211]. A $0.6 < \rho_a < 0.95$ is considered acceptable [212, 213], indicating that the indicators of the LV are highly correlated and consistently measure the same underlying construct.

Convergent validity, which measures how well the LV’s variance explains the indicators’ variance, is also examined. The AVE metric is utilized for this purpose. A $AVE > 0.5$ signifies that the latent variable explains more than half of the variance in its indicators.

Lastly, we assess *discriminant validity*, which measures the distinctness of a construct relative to other constructs. The heterotrait-monotrait ratio (HTMT) measures the difference between intra— and cross—construct correlations. An HTMT below 0.9 for all pairs of constructs is considered acceptable.

Note that these guidelines, as discussed by Hair et al. [200] and [210], are general, and the specific thresholds may vary depending on the particularities of the study.

When assessing the results, removing certain indicators from a dataset may be necessary if the LV does not meet the criteria above. Removing problematic indicators strengthens the validity of the LV. As an example, removing QX_Dwell from Lightning increases QX’s AVE from 0.446 to 0.519 (above the 0.5 threshold) and ρ_a from 0.885 to 0.944. The removal of an indicator implies that, for that dataset, that indicator is not correlated enough with its respective LV. Although we still report their l_i , a removed indicator is excluded from the model and not considered in subsequent analyses.

An indicator is always removed if $l_i < 0.4$, implying a very weak correlation with the LV. Indicators within the range $0.4 \leq l_i < 0.7$ are considered for removal only if the LV does not meet the quality criteria for AVE and ρ_a and the removal of these indicators helps to bring the LV within the acceptable range for these measures. Removing an indicator signifies that, for that specific dataset, the indicator is not sufficiently correlated with the other measures of the same construct. Any indicators that have been removed are *Grayed-out* in Table 5.2 for clarity. Rows with indicators that differ considerably between datasets (i.e., were removed in some and kept in other datasets) are marked with ♦.

There are many reasons why one indicator may be retained for one dataset but not another. Mostly, this is due to small divergences in how each metric is calculated and in the population sampled for each dataset. As an example, the low l_i for QE_FormTime in the Lightning dataset can be explained by the higher level of familiarity of the participants with the search engine used. While Scaffolding and SearchWell required users to use a custom search engine, Lightning participants mostly used Google.

Another possible reason is how a metric is calculated in the provided user logs. While Scaffolding and SearchWell participants were expected to search *linearly* (i.e., each new

Table 5.2: Indicators, outer loadings (OL), reliability coefficient (ρ_a) and average variance extracted (AVE) for the reflective LVs for all datasets. Indicators for QX and QQ are averaged over all queries. Superscript * mean statistically significant results. Cells with dashes (—) are values that are not possible to compute on the given dataset. Indicators that are grayed-out are not considered in further analysis. Rows removed in some datasets but not others are marked with an indicator ♦.

LV	Indicator	Definition	Scaffolding			Lightning			SearchWell		
			OL	ρ_a	AVE	OL	ρ_a	AVE	OL	ρ_a	AVE
QE	♦ QE_FormTime	Time (s) formulating queries	0.567*			0.364*			0.810*		
	QE_Queries	Queries issued	0.937*	0.920	0.739	0.922*	0.935	0.870	0.879*	0.953	0.810
	QE_Toks	Query tokens	0.939*			0.961*			0.973*		
	QE_UniqToks	Unique query tokens	0.936*			0.902*			0.942*		
	QX_Clicks	Documents clicked	0.919*			0.824*			0.898*		
QX	QX_Visits	Documents visited	0.961*			0.689*			0.922*		
	QX_Docs	Unique documents	0.937*			0.882*			0.927*		
	QX_MaxDepth	Max. ranking of clicked docs	0.741*	0.949	0.690	0.457*	0.944	0.519	0.645*	0.916	0.673
	QX_SERPTime	Time (s) spent in SERPs	0.732*			0.709*			0.671*		
	♦ QX_Time	Time (s) between Queries	0.785*			0.897*			—0.114		
	♦ QX_SERPs	SERPs viewed	0.761*			0.427*			0.336*		
	QX_Dwell	Total dwell time	0.166			0.200			—0.235		
	QX_Bookmarks	Bookmarked docs	0.769*			—			—		
	SX_Visits	Documents visited	0.911*			0.650*			0.952*		
	SX_Docs	Unique documents	0.948*			0.894*			0.970*		
SX	SX_Domains	Unique domains	0.871*			0.840*			0.897*		
	SX_Clicks	Queries with clicks	0.712*	0.915	0.690	0.735*	0.885	0.567	0.605*	0.967	0.707
	SX_Duration	Session duration	0.431*			0.689*			0.560*		
	♦ SX_Dwell	Sum of dwell times	—0.164			0.522*			0.392		
	SX_Branch	Session's branchiness	0.948*			0.857*			0.953*		
	SX_Bookmarks	Bookmarked documents	0.661*			—			—		
	WE_EssLen	Number of terms in essay	0.513*			0.985*			—		
WE	WE_EssUniq	Unique terms in essay	0.924*			0.983*			—		
	WE_EssDoc	Essay terms from documents	0.937*			0.507*			—		
	♦ WE_EssQuery	Essay terms from queries	0.693*	0.940	0.652	0.109	0.901	0.728	—	—	—
	WE_AnsLen	Average length of VKS answers	0.513*			—			—		
	WE_Inc	Len. diff. between pre and post-essay	—			0.863*			—		
KG	♦ KG_ALG	Absolute Learning Gain	0.596*			0.332*			0.980*		
	♦ KG_Post	Post-test score	0.587*			0.619*			0.116		
	♦ KG_RPL	Realized Learning Gain	0.600*			0.347*			0.910*		
	KG_EssCov	% of all vocabulary terms in essay	0.848*	0.853	0.523	0.441*	0.813	0.555	—	1.239	0.903
	♦ KG_EssDen	Vocabulary density	0.851*			0.172			—		
	KG_EssALG	Essay's ALG	—			0.776*			—		
	KG_EssRPL	Essay's RPL	—			0.850*			—		
	KG_EssPost	post-test essay score	—			0.993*			—		

query should be issued on the same SERP and multiple browser tabs are prohibited), Lightning participants had no such constraint, with participants frequently opening multiple tabs and not returning to the original SERP. Therefore, computing the difference in time between the moment the last document for a query is closed and the subsequent query is issued is not trivial and potentially noisy.

An interesting finding from Table 5.2 is that most indicators based on multiple choice questionnaires show small l_i s for the KG construct in both Scaffolding and Lightning datasets. This suggests that essay-based metrics are more inter-correlated, creating a more cohesive representation of learners' Knowledge Gains (KGs). This finding starts answering our first research question (can changes in KG, measured by metrics derived from the learners' answers to multiple-choice questionnaires, be causally explained by their search interactions?) by showing that essay-based metrics are more cohesive, measuring the same underlying construct (i.e., KG).

We hypothesize that this happens due to the more nuanced aspect of essay evaluations, with higher variance and capturing a more comprehensive view of a learner's knowledge, as discussed by Urgo and Arguello [42]. This result also underscores the insufficiency of questionnaire-based assessments in capturing the full extent of learning, highlighting the need for diverse, robust metrics such as those derived from automatic essay evaluations.

After removing problematic indicators, all reported values of ρ_a and AVE fall within acceptable quality thresholds. While not shown in the table, the HTMT ratios (measuring the differences between intra- and cross-construct ratios) also reveal no issues, indicating that all LVs measure distinct concepts. With the reflective model assessment complete, we now turn our attention to the analysis of the formative model.

5.5.2 Assessing the Formative Measurement Models

Formative measurement model evaluation involves checking *convergent validity*, *collinearity*, and *significance and relevance* of indicator weights, with results provided in Table 5.3.

The *convergent validity* of a construct is determined by the correlation between the latent variable (LV) and an alternative reflective measure of the same construct. A LV is considered valid if the path coefficient is robust, ideally exceeding 0.7.

Collinearity issues may occur when an indicator is highly correlated with others within the same construct. Excessive collinearity is undesirable since a formative model is a linear combination of multiple distinct indicators. It can inflate the model's weights' standard error, reducing the probability of statistically significant weights. Additionally, it could lead to incorrectly estimated weights and potentially cause sign changes (i.e., an indicator with a positive impact having a negative weight). We use the variance inflation factor (VIF) metric to evaluate collinearity, quantifying the standard error inflation due to collinearity in an indicator. Acceptable VIF values are generally less than 5.0.

Lastly, the *significance and relevance* of each indicator are assessed. This involves evaluating the value and significance of the w_i (i.e., the relative causal contribution to the construct) and the l_i (i.e., the absolute importance of the indicator when other indicators are not considered).

Formative indicators with non-significant weights may still be retained due to the multifaceted nature of formative constructs. In the absence of VIF issues, an indicator covering an important aspect of the construct should only be considered for removal if it

Table 5.3: Indicators, outer weights (**OW**) and outer loadings (**OL**) for the indicators of the formatively measured LVs. Superscripts * indicates a value significantly different from 0. Indicators that are grayed-out are removed before further analysis. Rows removed in some datasets but not others are marked with an indicator ♦.

LV	Indicator	Definition	Scaffolding		Lightning		SearchWell	
			OW	OL	OW	OL	OW	OL
DQ	DQ_AoA	Avg. AoA	-0.335*	-0.025	0.506*	0.357*	0.266	0.154
	♦ DQ_ParNum	# of paragraphs	0.037	0.437*	-0.044	-0.248	0.329	0.12
	♦ DQ_ParLen	Avg. paragraph length	0.353*	0.415*	-0.080	-0.319	-0.456*	-0.490
	♦ DQ_SMOG	SMOG readability	-0.058	-0.377*	-0.283	-0.453	0.095	-0.165
	DQ_UniqRatio	% of unique terms	0.108	-0.369*	0.589*	0.393*	-0.326	-0.375*
	DQ_VocabRatio	% of vocabulary terms	0.772*	0.889*	1.013*	0.499*	0.648*	0.819*
	DQ_Vocab	# of vocabulary terms	0.228	0.703*	0.266	0.032	-0.073	0.183
	♦ DQ_ImgNum	# of tags	-0.150	-0.092	0.223	0.333	0.404*	0.338*
QQ	♦ QQ_AoA	Avg. AoA	0.191	0.247	0.137	0.244	0.242*	-0.018
	♦ QQ_IDF	Harmonic avg of terms IDF's	0.372*	0.033	0.245	0.412	0.440*	0.048
	QQ_TokNum	# of tokens	0.303*	0.179	0.855*	0.878*	1.175*	0.738*
	QQ_VocabRatio	% of vocabulary terms	0.967*	0.917*	0.441*	0.455*	0.635*	0.223
	♦ QQ_FormTime	Time (s) formulating	-0.081	-0.208	0.398*	0.382	0.066	0.045

meets the following conditions: (i) its w_i is statistically non-significant, (ii) its correlation with the construct is low ($I_i < 0.5$), and (iii) its I_i is non-significant.

Table 5.3 displays the w_i and I_i for all indicators for both QQ and DQ LVs.

We use a reflective version of DQ and QQ, Document Quality (reflective) (DQ_R) and Query Quality (reflective) (QQ_R), respectively, to assess the convergent validity. Their I_i , ρ_a and AVE can be seen in table 5.4. Path coefficients for all three datasets are strong (i.e., above 0.7) and statistically significant. For DQ, path coefficients of 0.807, 0.701 and 0.874 were found for Scaffolding, Lightning and SearchWell, respectively. For QQ, the path coefficients are 0.854, 0.869 and 0.916. These results suggest that the formative indicators are valid and can be used to measure the DQ and QQ constructs.

In the process of evaluating the significance and relevance of the indicators, we removed the indicator for the number of images on each page (DQ_ImgNum) from both the Scaffolding and Lightning datasets due to its non-significant outer weights and low outer loadings¹¹. Similarly, we also removed the indicators for the number and length of paragraphs (DQ_ParNum and DQ_ParLen) from the Lightning dataset and the SMOG readability index (DQ_SMOG) from the SearchWell dataset, all due to their low and non-significant outer weights and loadings. However, the Average Age-of-Acquisition (DQ_AoA) was retained in the Lightning dataset despite its low and non-significant outer weights and loadings. We made this decision because its removal resulted in changes in the significance of the outer weights and loadings of other indicators, suggesting its influence on the overall structure of the latent variable. For the same reason, we also kept QQ_AoA for Scaffolding. We also removed QQ_IDF from the Lightning dataset, despite its high loading, due to its high

¹¹One possible explanation for their removal for the Scaffolding dataset is because the HTMLs were rendered within the search system in a simplified version instead of directly in their original web pages (c.f. 2).

Table 5.4: Outer loadings **OL**, reliability coefficient (ρ_a) and average variance extracted (AVE) for reflectively measured version of the formative LVs for all datasets. Indicators are considered for all queries submitted by each learner. Superscript * mean statistically significant outer loadings.

LV	Indicator	Definition	Scaffolding			Lightning			SearchWell		
			OL	ρ_a	AVE	OL	ρ_a	AVE	OL	ρ_a	AVE
DQ_R	DQ _R _VocabTotal	# of vocab terms found	0.803*			0.974*			0.845*		
	DQ _R _VocabPct	% of vocab terms found	0.918*	0.864	0.789	0.875*	0.937	0.874	0.835*	0.814	0.724
	DQ _R _VocabUniq	# of unique vocab terms found	0.938*			0.974*			0.874*		
QQ_R	QQ _R _VocabTotal	# of vocab terms used	0.902*			0.722*			0.751*		
	QQ _R _VocabPct	% of vocab terms used	0.880*	0.960	0.818	0.858*	0.947	0.593	0.779*	0.707	0.566
	QQ _R _VocabUniq	# of unique vocab terms used	0.930*			0.722*			0.725*		

Table 5.5: Path coefficients, Total Effect and f^2 score for each latent variable (LV) into Knowledge Gain (KG). Values with a superscript* indicate statistical significance.

LV	Scaffolding			Lightning			SearchWell		
	Path Coeff.	Total Effect	f^2	Path Coeff.	Total Effect	f^2	Path Coeff.	Total Effect	f^2
QE	-0.179	-0.022	0.019	0.050	0.190*	0.001	-0.040	-0.069	0.001
QQ	0.094	0.558*	0.015	0.004	-0.069	0.000	0.073	-0.125	0.003
QX	-0.246*	-0.160*	0.058	-0.005	0.082	0.000	-0.110	-0.136*	0.007
DQ	0.625*	0.622*	0.303	0.096	-0.060	0.007	-0.322*	-0.321*	0.059
SX	0.127	0.157*	0.012	0.127	0.245*	0.011	-0.039	-0.039	0.001
WE	0.112*	0.112*	0.024	0.553*	0.553*	0.423	—	—	—

VIF value, caused by its high correlation with QQ_AoA. After these removals, we found no more VIF issues.

Some interesting findings can be seen in Tables 5.3. For the DQ construct, all datasets show strong and positive outer weights and loadings of the density of the vocabulary terms in the document (DQ_VocabRatio). In contrast, the count of vocabulary terms (DQ_Vocab) have mixed results across the datasets, ranging from positive and significant positive weights in the Lightning dataset to negative and significant weights for the SearchWell dataset. One possible explanation is that, due to the difference in the population of each dataset, vocabulary terms had a considerably higher impact on university students versus the general population. This outlines that, while the mere presence of vocabulary terms is not necessarily a strong predictor of the quality of a document, the *density* of relevant terms have a stronger causal relation to the quality of documents and should be prioritized when ranking documents in a SAL context.

Examining the textual complexity of the documents, we find varied results for the average age-of-acquisition of terms (DQ_AoA) and the readability of documents (DQ_SMOG). In the Scaffolding and Lightning datasets, more complex documents are associated with negative weights, suggesting a negative impact on document quality. Conversely, in the SearchWell dataset, more complex documents are associated with positive weights, indicating a beneficial impact. This discrepancy may be attributed to the differing levels of user interaction across the datasets. Participants in the Scaffolding and Lightning datasets interacted with over ten times more documents than those in the SearchWell dataset. Users preferred more complex and dense documents in shorter sessions, as indicated by the higher DQ_VocabRatio. This suggests a more focused or specialized search process.

However, simpler documents were a stronger prediction of its quality in longer sessions, as seen in the `Scaffolding` and `Lightning` datasets. This could imply that, while users explored a larger number of documents and spent more time making sense of the content, the complexity of the documents may have hindered their ability to comprehend the material fully. Interestingly, as discussed in Section 5.5.3, this strategy of preferring simpler documents in longer sessions appears more effective. The DQ LV is a stronger causal predictor for Knowledge Gain in these datasets.

Table 5.6: Summary of the indicators that were kept for each dataset and each LV. Indicators with a ✓ were kept for the analysis. Indicators with a ✗ were removed. Indicators with a — were not applicable to the given dataset.

LV	Indicator	Scaffolding	Lightning	SearchWell
QE	QE_FormTime	✓	✗	✓
	QE_Queries	✓	✓	✓
	QE_Toks	✓	✓	✓
	QE_UniqToks	✓	✓	✓
QX	QX_Clicks	✓	✓	✓
	QX_Visits	✓	✓	✓
	QX_Docs	✓	✓	✓
	QX_MaxDepth	✓	✓	✓
	QX_SERPTime	✓	✓	✓
	QX_Time	✓	✓	✗
	QX_SERPs	✓	✓	✗
	QX_Dwell	✗	✗	✗
	QX_Bookmarks	✓	—	—
	SX_Visits	✓	✓	✓
SX	SX_Docs	✓	✓	✓
	SX_Domains	✓	✓	✓
	SX_Clicks	✓	✓	✓
	SX_Duration	✓	✓	✓
	SX_Dwell	✗	✓	✓
	SX_Branch	✓	✓	✓
	SX_Bookmarks	✓	—	—
WE	WE_EssLen	✓	✓	—
	WE_EssUniq	✓	✓	—
	WE_EssDoc	✓	✓	—
	WE_EssQuery	✓	✗	—
	WE_AnsLen	✓	—	—
	WE_Inc	—	✓	—

Continued on next page

Table 5.6: Summary of the indicators (Continued)

LV	Indicator	Scaffolding	Lightning	SearchWell
KG	KG_ALG	✓	✗	✓
	KG_Post	✗	✓	✓
	KG_RPL	✓	✗	✓
	KG_EssCov	✓	✓	—
	KG_EssDen	✓	✗	—
	KG_EssALG	—	✓	—
	KG_EssRPL	—	✓	—
	KG_EssPost	—	✓	—
DQ	DQ_AoA	✓	✓	✗
	DQ_ParNum	✓	✗	✗
	DQ_ParLen	✓	✗	✓
	DQ_SMOG	✓	✗	✗
	DQ_UniqRatio	✓	✓	✓
	DQ_VocabRatio	✓	✓	✓
	DQ_Vocab	✓	✓	✓
	DQ_ImgNum	✗	✓	✓
QQ	QQ_AoA	✓	✗	✓
	QQ_IDF	✓	✗	✓
	QQ_TokNum	✓	✓	✓
	QQ_VocabRatio	✓	✓	✓
	QQ_FormTime	✗	✓	✗

In the case of the QQ construct, the ratio of vocabulary terms to the total number of query terms (QQ_VocabRatio) consistently shows strong and positive outer weights and loadings across all three datasets. This suggests that queries with more vocabulary terms significantly contribute to their quality. Similarly, the length of the query (QQ_TokNum) is a robust predictor of query quality, particularly in the Lightning dataset.

Furthermore, the use of more complex terms in the query, as measured by QQ_AoA and QQ_IDF, also substantially contributes to the QQ latent variable. Interestingly, the time spent formulating each query (QQ_FormTime) is only significant in the Lightning dataset. This suggests that for the more educated participants in the Lightning dataset, investing time in crafting a query is a strong predictor of its quality. However, the time spent formulating a query does not play a significant role in the other two datasets, representing a more general population.

We show a summary of the indicators retained or not for each dataset in Table 5.6

5.5.3 Assessing the Structural Model

After validating the measurement models, we analyze the structural models. This helps us understand the strength, direction, and significance of relationships between LVs and how well our causal model can explain (and predict) variance in the dependent variables.

The first step involves checking for collinearity issues among the latent variables (LVs). This check ensures that our predictors are not highly correlated with each other, which

could bias the regression results. As in the formative model assessment, we use the VIF metric, expecting values below 5.0.

If there are no collinearity issues, we evaluate the overall explanatory power of KALM using the R^2 metric, which measures the proportion of the variance in the endogenous construct explained by the model. Interpreting R^2 values isn't straightforward, especially in IIR, where there's a lack of precedent for SEM studies. While R^2 values of 0.75, 0.50, and 0.25 might be considered substantial, moderate, and weak, respectively [214], acceptable ranges can vary across research fields. For instance, in predicting stock returns, values as low as 0.1 are considered satisfactory [215]. Therefore, we interpret R^2 values in relative terms, comparing them across the datasets analyzed here.

We also report the f^2 effect size for each construct, which measures the change in R^2 if a specific LV is removed from the model. Effects of 0.02, 0.15, and 0.35 represent small, medium, and large effects, respectively.

Finally, we evaluate the significance and relevance of the paths within the model. A path with a larger coefficient suggests a stronger causal relationship between two constructs. We report not only the Path Coefficient for each LV towards the KG construct but also calculate its total effect, which combines its direct and indirect effects on KG. The path coefficient represents the direct effect, while the indirect effect captures the construct's influence through other intermediary constructs. The overview of all path coefficients for the three datasets is shown in Figure 5.3. Results for the R^2 and f^2 values can be seen in Table 5.5. Finally, detailed total effects values between all LVs is shown in Table 5.9

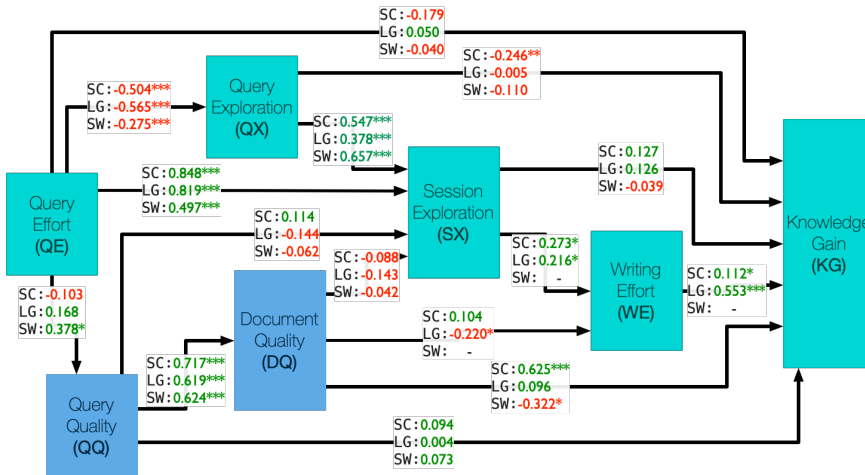


Figure 5.3: Path model with coefficients for KALM. Numbers after SC, LG and SW are the path values between two LVs in the Scaffolding, Lightning and SearchWell datasets, respectively. Values in green and red are positive and negative path coefficients. Superscripts *, ** and *** indicates p-values below 0.1, 0.01 and 0.001.

The VIF values for all LVs across the datasets sit comfortably below the 5.0 threshold, thereby indicating no collinearity issues. The highest values observed were 3.066 and 3.131 for the relationships between KG and QE in the Scaffolding and Lightning datasets,

respectively, and 2.298 between KG and QQ in the SearchWell dataset.

As for the R^2 value of the KG construct, we found values of 0.542 for Scaffolding, 0.349 for Lightning, and 0.092 for SearchWell. Given the lack of previous studies, interpreting these values is not straightforward. Nonetheless, comparing the three datasets, KALM shows significantly higher in-sample predictive power for the Scaffolding and Lightning datasets, while the R^2 value for SearchWell is noticeably lower.

The noticeable discrepancy between SearchWell and the other datasets regarding the low KG R^2 score comes primarily from the weights of the KG indicators shown in Table 5.2. This observation underscores our argument that a multiple-choice questionnaire might not fully capture the intricate and multifaceted nature of learning.

The negative path coefficient between DQ and KG in the SearchWell dataset, coupled with the higher outer weights of textual complexity metrics in the DQ construct, provides insight into some unexpected outcomes in the SearchWell dataset’s measurement model assessment. Unlike the strong R^2 score for Scaffolding, it seems that more complex documents (i.e., those with higher DQ_AoA values) might negatively impact the final knowledge gain in the SearchWell dataset, despite improving the document’s quality.

A possible explanation for this result comes from the duration of each participant’s session in the SearchWell dataset. For participants in the Scaffolding dataset the average session duration was almost 43 minutes and for participants in the Lightning dataset 24 minutes; participants in the SearchWell dataset searched for only 6 minutes on average. This discrepancy has considerable effects on user behavior. For instance, given that learners spend considerably less time reading, the complexity of the documents becomes more critical to the learner’s knowledge acquisition.

For the Scaffolding dataset, which had the highest R^2 value, DQ displayed a strong and significant positive effect on the KG construct, with a path coefficient of 0.565 and a medium effect size of $f^2 = 0.303$. Interestingly, QQ showed a large and significant total effect despite a small direct path coefficient. This outcome mainly stems from its path coefficient to DQ (0.722). It suggests that while a good query might not lead to knowledge gains, it retrieves quality documents that adds to the learner’s knowledge.

Interestingly, QX demonstrates a significant *negative* path coefficient and total effect on KG in the Scaffolding and SearchWell datasets, respectively. This result suggests that more extensive exploration, characterized by an increased number of documents and SERPs or more time spent on each query, doesn’t necessarily equate to knowledge gains.

The WE construct significantly affects the Lightning dataset. This is unsurprising, given that the learning evaluation on this dataset is heavily biased towards essays (c.f. Table 5.2). This indicates that the learner’s effort in creating a more comprehensive essay- reflected in longer essays with more terms derived from the documents they found- strongly impacts their knowledge gains.

One interesting finding also emerges when comparing the three datasets. SX and QE do not have any significant impact on the KG construct on the two datasets that relied on crowd-workers, Scaffolding and SearchWell. However, in the Lightning dataset, with more educated participants (i.e., university students), these constructs have a significant total effect on KG. Inspecting the path model carefully, we see that this is primarily mediated by the paths $SX \rightarrow WE \rightarrow KG$ with an indirect effect of 0.120 and $QE \rightarrow SX \rightarrow WE \rightarrow KG$, with a specific indirect effect of 0.098. The implication is that an increase in exploration, as

measured by more (and more diverse) queries and more documents and time spent in each query, may lead to higher knowledge gains in more educated learners but not necessarily in a broader set of learners.

5.6 Discussion

In the introduction of this thesis chapter, we outlined three main goals for our analysis. First, to assess if popular KG metrics based on traditional multiple-choice questionnaires are enough to capture the complex nature of users’ behaviors and learning. Second, to evaluate if alternative metrics, based on vocabulary usage on post-test essays, are better at this task. Finally, we explore the causal relationships between search behaviors and metrics and knowledge acquisition in SAL environments. In this section, we explore, in more detail, how our experiments answer these questions and how our findings agree (or differ) from other results in the literature. We show in Table 5.8 an overview of our findings and whether SAL literature agrees with them or not.

However, one important aspect underlying our discussion is the difference between the studied datasets. While this was already discussed in Section 5.4, we want to highlight some key differences here, as they play an important role when interpreting the results.

Regarding the type of participants, both Scaffolding and SearchWell recruited learners using crowdsourcing platforms, while Lightning relied on undergraduate students. On the evaluation of learning, our study from Chapter 2 relied primarily on the ALG and RPL metrics, derived from the difference between pre- and post-test multiple choice questionnaires. While an essay was also asked, it was not evaluated. In the original study for the Lightning dataset, learning was measured by a multiple-choice questionnaire and an open-ended essay. Essays were evaluated by the number of correct concepts present in them. While no further analysis of the data was provided, the dataset includes a learning metric similar to ALG. Finally, SearchWell reports learning by the absolute number of correct answers and the difference between the number of correct answers in a multiple-choice questionnaire before and after their search sessions. Another important distinction is in the *duration* of the search sessions. Participants in the Scaffolding dataset had to search for *at least* 30 minutes. Learners in the Lightning study had to search for *at most* for 30 minutes, and participants in the SearchWell user study had no upper or lower limits.

Table 5.7: Differences in R^2 when changing indicators for the KG LV. Values of AVE and ρ_a are still within the acceptable ranges defined in Section 5.5.1

	Scaffolding			Lightning		
	R^2	Δ	Indicators	R^2	Δ	Indicators
Full model	0.542	—	KG_ALG KG_RPL KG_EssCov KG_EssDen	0.349	—	KG_Post KG_EssCov KG_EssALG KG_EssPost KG_EssALGKG_EssRPL
Questionnaire	0.113	−79 %	KG_ALG KG_RPL KG_Post	0.225	−35 %	KG_ALG KG_RPL KG_Post
Vocab usage	0.558	3 %	KG_EssCov KG_EssDen	0.302	−13 %	KG_EssCov KG_EssDen
Essay assessment	—	—	—	0.302	−13 %	KG_EssALG KG_EssPost KG_EssRPL

5.6.1 On Learning Metrics

In response to our first research question, “Can changes in KG, as measured by metrics derived from the learners’ answers to multiple-choice questionnaires, be causally explained by their search interactions”?, and its follow-up “Are metrics based on vocabulary usage in learners’ essay better explained by the learners’ interactions”?, we found the answer in our analysis of reflectively measured variables, as discussed in Section 5.5.1 and in Table 5.7, showing the differences in R^2 if the indicators of KG are changed.

Our analysis highlighted the importance of essay-based metrics instead of multiple-choice questionnaires when evaluating learners’ Knowledge Gains (KGs). This was particularly evident when we looked at the KG construct in the *Lightning* dataset, where questionnaire-based indicators showed very low outer loadings, leading to their removal. The *Scaffolding* dataset also echoed this trend, presenting lower outer loadings for metrics based in questionnaires when compared to essay-based metrics such as *KG_EssCov* and *KG_EssDen*, representing the coverage and density of vocabulary terms, respectively.

Furthermore, our analysis revealed a limitation of the *SearchWell* dataset (and in other studies that rely on multiple-choice questionnaires only), which lacks a post-test essay. This deficiency resulted in lower explanatory power for the KG construct, as reflected by a substantially lower R^2 metric. This underscores the conclusion that relying solely on a multiple-choice questionnaire to assess learners’ knowledge gains falls short of capturing the nuanced and complex nature of learning. On the other hand, incorporating essay evaluations, as in the *Scaffolding* and *Lightning* datasets, considerably bolstered the R^2 metric, suggesting a more comprehensive representation of the learning process. This finding underscores the importance of leveraging diverse and robust learning metrics in evaluating learners’ knowledge gains in SAL environments.

This difference in R^2 values is also observed within the same dataset. As evidenced in Table 5.7, if we only use questionnaire-based metrics as indicators for KG in *Scaffolding* or *Lightning*, we observe a drop in the reported R^2 values of over 79% and 35%, respectively. Measuring only with *KG_EssCov* and *KG_EssDen*, we see a small 3% increase in the R^2 value for *Scaffolding* and a 13% decrease for *Lightning*. This finding implies that incorporating simple essay-based evaluation metrics, such as *KG_EssCov* and *KG_EssDen*, can drastically improve the amount of KG’s variability explained by the model.

Our findings echo the argument presented by Urgo and Arguello [42], emphasizing the importance of open-ended assessments, such as essay writing, in SAL evaluation. While expensive to measure, these methods allow for assessing a learner’s capacity to devise new solutions or provide unique representations of a domain, particularly during complex tasks. We also highlight the same concerns about the limited coverage of a topic, potential for guessing, and susceptibility to priming effects, where learners might focus only on concepts outlined in pre-search questions.

This also aligns with research outside the SAL field. For instance, in medical education, Newble et al. [216] demonstrated that multiple-choice tests tend to overestimate learners’ abilities. They advocated for a shift towards free-response items, which accurately reflect learners’ understanding and proficiency.

Our results underscore the issues of multiple-choice questionnaires discussed earlier and propose a viable alternative. Essay evaluations can be complex, expensive, and may require expert judgment. However, they show a clear advantage in assessing knowledge

Table 5.8: Comparison between our main findings and existing literature. References marked with ✓ agree with our findings, while those marked with ✗ disagree.

RQ	Finding	References
RQ1	Multiple-choice questionnaires fail to capture the full extent of the knowledge of the learner	Urgo and Arguello [42]✓ Newble et al. [216] ✓
	A mixed strategy should be preferred when evaluating learning gains	Pellegrino [217] ✓
RQ2	Query quality has a positive impact on knowledge acquisition	Gadiraju et al. [87]✓ Bhattacharya and Gwizdka [34]✓ Umemoto et al. [61]✓ Chapter 2✓
	Document quality has a positive impact on knowledge gains	Vakkari et al. [56]✓ Yu et al. [112]✓ Vakkari et al. [55]✓ Collins-Thompson et al. [28]✓ Syed and Collins-Thompson [67]✓
	Higher session-level exploration does not lead to better learning outcomes	Pardi et al. [37]✓ Bron et al. [218]✓ Yu et al. [112]✓ Collins-Thompson et al. [28]✗ Gadiraju et al. [87]✗ Lu and Hsiao [129]✗ Yu et al. [53]✗
	More query-level exploration does not lead to better learning outcomes	Yu et al. [112]✓ Vakkari et al. [56]✓ Yu et al. [112]✗ Vakkari et al. [56]✗ Pardi et al. [37]✗
	Effort when writing a query has a positive impact on knowledge acquisition (Lightning)	Vakkari et al. [196]✓ Vakkari et al. [56]✓ Vakkari et al. [55]✗
	The effort a learner puts into writing their essay is a strong predictor of knowledge acquisition	Liu and Belkin [219]✓ Vakkari et al. [56]✓ Vakkari et al. [55]✗

Table 5.9: Total effects between all LVs. Values with a superscript * indicate statistical significance.

LV	Dataset	QQ	QX	DQ	SX	WE	KG
QE	Scaffolding	-0.103	-0.504*	-0.074	0.567*	0.147*	-0.022
	Lightning	0.168	-0.565*	0.104	0.566*	0.099	0.190*
	SearchWell	-0.378*	-0.275*	0.236*	0.283*	—	-0.069
QQ	Scaffolding			0.717*	0.051	0.089	0.558*
	Lightning			0.619*	-0.232*	-0.186*	-0.069
	SearchWell			0.624*	-0.088	—	-0.125
QX	Scaffolding				0.547*	0.149*	-0.160*
	Lightning				0.378*	0.082	0.088
	SearchWell				0.657*	—	-0.136*
DQ	Scaffolding				-0.088	0.080	0.622*
	Lightning				-0.143	-0.251*	-0.060
	SearchWell				-0.042	—	-0.321*
SX	Scaffolding					0.273	0.157*
	Lightning					0.216*	0.245*
	SearchWell					—	-0.039
WE	Scaffolding						0.112*
	Lightning						0.553*
	SearchWell						—

gains. As showcased by the indicators KG_EssCov and KG_EssDen, which measure the coverage and density of vocabulary terms, respectively, essay-based assessments provide a richer perspective of learner’s knowledge acquisition.

Given our results, we advocate for a *mixed strategy* when assessing learning outcomes in SAL systems. This was also discussed by Pellegrino [217], asserting that the validity and fairness of inferences can be enhanced by providing learners with multiple ways to demonstrate their competence. Our findings corroborate this idea in two ways. First, the dataset that better integrated multiple-choice questions and essay assessments (i.e., Scaffolding) showed a higher R^2 score, indicating a more comprehensive capture of learning processes. Second, by measuring KG with multiple-choice questions and essay evaluations, we can greatly increase the explained variance in KG. Thus, integrating diverse assessment types seems to yield a more accurate representation of knowledge gains in SAL environments.

5.6.2 Causal Relationships

Addressing our second research question, “What causal relationships exist between search behaviors and knowledge acquisition in SAL, and how can these relationships be modeled and validated”? requires an examination of the structural model assessment of KALM, which was done in Section 5.5.3.

This analysis revealed that the latent variables DQ and WE exert a strong positive influ-

ence on learners' KGs, especially in the diverse participant set of the Scaffolding dataset. There, QQ also emerged as a significant factor influencing learning outcomes, mediated through the DQ construct.

These findings underscore the need for SAL systems to enhance not just their ranking models (as highlighted by Syed and Collins-Thompson [68]) but also to assist users in crafting more effective queries. Visual aids and suggestions, as suggested by us in Chapter 2 and Umemoto et al. [61], can potentially boost learning outcomes.

Interestingly, our findings differ from conventional wisdom in SAL. More extensive exploration (QX), or the effort a learner exerts exploring individual queries, was found to negatively influence KG in two of the datasets. This suggests that the quality of information is more important than its quantity. Additionally, in these same datasets, neither the effort dedicated to formulating a query (QE) nor the overall effort of a session (SX) had a significant impact on KG. These results underscore two important implications for the design of IR systems for learning: the need to enhance the ranking function to prioritize high-quality, relevant documents and to support learners' writing efforts. Ultimately, this shows that more exploration doesn't necessarily result in improved learning outcomes.

When comparing our results with previous research, we observe a dilemma similar to Vakkari [202], where contrasting results are reported for different behavior metrics. Concerning metrics similar to our QQ construct, both Gadiraju et al. [87] and Bhattacharya and Gwizdka [34] found that more complex queries correlate with higher knowledge gains, aligning with our findings in the Scaffolding dataset. Interestingly, Yu et al. [112], who also used the SearchWell dataset, found significant and positive correlations between document quality (analogous to our DQ_SMOG and DQ_ImgNum indicators) and learning outcomes. However, our study revealed a nuanced relationship: in this dataset, while DQ negatively impacted KG, DQ_SMOG had a small, non-significant positive outer weight and a negative outer loading with the DQ construct, while DQ_ImgNum had significant positive outer weights and loadings with KG.

On the other hand, the Scaffolding dataset exhibited a contrasting pattern. Here, DQ had a strong positive causal effect on KG, while DQ_SMOG was negatively correlated with higher document quality, and DQ_ImgNum showed no significant result. Despite these discrepancies, our results do not contradict those of Yu et al. [112]. While their study reported a significant linear correlation between individual variables and learning outcomes, our study provides a more comprehensive view of causality in knowledge gains, particularly given our multi-faceted KG construct. Our work aligns with Vakkari et al. [56], who presented EQM, a major inspiration for our study and found a strong positive impact between a similar LV and the quality of end-of-session essays. This is also supported by their previous work [55], where more useful documents were associated with higher learning gains. Furthermore, our findings agree with Collins-Thompson et al. [28], who found that learners engaging with higher quality documents displayed higher knowledge gains.

Exploration and effort metrics are common in SAL studies. In our research, we represent these aspects through two LVs, namely SX for session-wide exploration and QX for effort per query. We found that SX has a significant indirect impact in the Lightning dataset, whereas QX exerts a strong *negative* effect in the two crowd-sourced datasets.

These findings are in line with Pardi et al. [37] and Bron et al. [218], who found that some SX indicators (e.g., SX_Bookmarks, SX_Docs, SX_Dwell) are not significantly correlated

with KG, supporting our findings in the Scaffolding and SearchWell datasets.

Nevertheless, this differs from the prevailing notion in SAL literature that more time spent reading documents (i.e., higher `SX_Dwell`) is associated with higher KG [28, 53, 87, 129]. We observed this relationship only in the controlled `Lightning` dataset. Furthermore, some research indicates that `SX` may negatively correlate with learning, as Yu et al. [112] found that `SX_Duration` is negatively associated with higher Knowledge Gain levels.

Regarding the per-query exploration construct (`QX`), our study corroborates some findings by Yu et al. [112] and Vakkari et al. [56], who reported negative correlations between some `QX` indicators, such as `QX_Time` and `QX_Dwell`, and learning. However, their studies also posit a positive correlation between other `QX` indicators and learning gains, hinting at the nuanced relationship between exploration effort and learning outcomes.

Contrarily, several studies have found positive correlations between indicators of `QX` and KG. Yu et al. [112] also reported a positive relationship between `QX_MaxDepth` and learning gains. Similarly, Vakkari et al. [56] found in their EQM model that `QX` indicators like `QX_Clicks` and `QX_Dwell` positively influenced essay quality. Likewise, Pardi et al. [37] found that higher `QX_Dwell` was associated with better learning outcomes.

The latent variable `QE` has attracted considerable attention in SAL studies. In our research, `QE` was found to significantly impact KG only in the `Lightning` dataset. This finding aligns with Vakkari et al. [196], who reported a positive correlation between `QE_UniqToks` and learning gain. In their EQM model, Vakkari et al. [56] employed similar indicators to our `QE_Toks`, `QE_UniqToks`, and `QE_FormTime` for their `QE` LV. They discovered a positive path coefficient between `QE` and better essay quality during the build-up phase of essay writing but a negative effect during the boil-down phase, which differs from our findings. Their 2018 study [55] found that `QE_Queries` is negatively associated with essay quality, contrasting with our observations in this chapter.

Regarding `WE`, the effort a learner puts into their essay writing, our study found it to be the strongest predictor of learning gain for the `Lightning` dataset. This aligns with Liu and Belkin [219], who reported that learners spending more time on essay crafting achieved higher task performance. Similarly, Vakkari et al. [56] found that more time, more revisions, and a higher overlap between the essay and document (akin to our `WE_EssDoc` indicator) are also predictors of higher essay quality. However, their prior study [55] reported a negative correlation between these indicators and success in finding relevant documents, doffering from our findings here.

In SAL studies, a range of correlations—positive, negative, or non-existent—has been observed between various metrics and definitions of Knowledge Gain. Our study corroborates this variation, showing significant effects across different datasets between LVs and KG. This variability underscores that no two studies are identical, highlighting the conclusion made by Urgo and Arguello [42]: “Having precisely defined objectives can help researchers develop assessment items that measure specific types of learning. Learning is inherently multidimensional”. Therefore, it’s essential for researchers to clearly define the learning objectives specific to their SAL use case.

Our findings also illustrate that different learner groups may benefit differently from various facets of the learning process. For instance, crowd-sourced learners (as represented in the Scaffolding dataset) may benefit more from improved queries and ranking functions. In contrast, other learners, such as university students from the `Lightning`

dataset, may benefit more from extensive exploration and dedicated effort in essay writing. These insights further highlight the importance of tailoring the approach to each study's specific learning objectives and context.

5.7 Conclusion

Understanding the intricate relationships between learners' behaviors and outcomes in a SAL setting is not simple, and causal analysis provides methods to analyze it more principledly. This study leveraged our KALM and PLS-SEM to examine three distinct SAL user study datasets. Our findings indicate that factors impacting learning are distinct based on the learning process and learner demographics. For instance, we found that the quality of retrieved documents and learner queries significantly affect learning in a more general population, particularly in crowd-sourced studies. However, high levels of exploration per query tend to lower learning gains. Conversely, in controlled settings with more educated learners, diverse querying and extensive session exploration yield better open-ended essays and, thus, enhance learning outcomes.

The relationships between learners' behavior and outcomes in a SAL setting are intricate. Using causal analysis methods to model these relationships is crucial for understanding how interventions in search engines may influence learners' experience and learning. This chapter examined three distinct SAL user study datasets using our KALM. Leveraging PLS-SEM, we highlighted how various aspects of search affect learning differently depending on both the learning process and learner demographics.

Our experiments have shown that the quality of retrieved documents and learner queries significantly affect learning in a more general population, particularly in crowd-sourced studies. In contrast, higher levels of exploration per query lead to lower learning gains. Conversely, in controlled settings with more educated learners, the impact of these aspects is negligible. Instead, diverse querying and extensive session exploration yield better open-ended essays, greatly impacting learning outcomes.

Moreover, our research shows that traditional learning metrics, such as RPL and ALG, based on multiple-choice questionnaires, fail to capture the nuanced nature of learning. Instead, metrics reflecting the quality of learners' essays present a more comprehensive picture of the learning process. Therefore, we suggest that future SAL research incorporates open-ended learning assessments. Our results indicate that combining multiple-choice questionnaires with open-ended essay scores significantly enhances our ability to estimate and predict learners' knowledge gains.

While KALM was shown to be effective in estimating the causal relationships between learners' behavior and knowledge gains in certain scenarios, it is not without flaws. For instance, it does not explicitly consider a learner's prior knowledge or the differences between learning topics. We hope future SAL research will further evolve our model, unveiling more interesting and useful connections between search and learning behaviors.

Additionally, we encourage more research in IIR to consider causal methodologies seriously. Although simpler correlation-based methods like linear models have their utility in identifying specific metrics that may lead to learning, causal methods may reveal more complex interactions between users and search engines. This becomes particularly crucial as we move towards more personal relationships between search engines and users facilitated by Large Language Models and conversational interfaces.

6

Conclusion

In this chapter, we summarize the main findings discussed throughout this thesis, using the research questions proposed in Chapter 1. After that, we discuss the limitations of the work presented here and propose how future research in the SAL and, more broadly, IIR fields should be conducted moving forward.

6.1 Summary of Findings

To guide our discussion, we start by recalling the research questions proposed in Chapter 1:

- ORQ1** What changes in the search engine can significantly impact learners' behavior and knowledge acquisition process?
- ORQ2** How can we model the learner's behavior and knowledge changes throughout their search session?
- ORQ3** What behaviors and metrics best explain and predict a learner's knowledge gains at the end of their search session?

6.1.1 Impacting Learner's Behavior and Knowledge Acquisition

The first research question discusses the main interest of most SAL practitioners: “*What should I change to help learners in their knowledge acquisition journey?*”? A search system is a complex entity with many moving parts, providing many possible intervention points that can be changed to better support learners in both the front-end (i.e., the user interface) and the back-end (i.e., the retrieval system itself). Therefore, being able to provide guidance on which of these intervention points are more (or less) likely to provide the desired impact is of utmost importance.

In Chapter 2, we discussed that impacting a learner's knowledge acquisition is not simple. There, we propose to help learners using *instructional scaffolding*, a concept borrowed from the field of education [91–94]. Under this setting, instructors (or, in the case of a SAL system, the system itself) guide the learners throughout their learning journey. In that chapter, we proposed three different methods for providing such guidance: (i) automatic

query rewriting (AQE_{SC}); (ii) a curated static topical outline (CURATED_{SC}); and (iii) a curated topical outline with instant feedback on the exploration of the topic space (FEEDBACK_{SC}).

After performing a user study with 126 participants, our findings show that, while none of the proposed methods significantly impacted learning outcomes, they significantly changed user behavior on several metrics. Specifically, we show that learners' dwell time, number of queries, and number of documents clicked significantly increased when provided with explicit visual scaffolding (i.e., the FEEDBACK_{SC} and CURATED_{SC} conditions).

We also observed signs of *gamification* in Chapter 2 under the FEEDBACK_{SC}. Learners can become overwhelmed by excessive feedback and focus more on filling the progress bars instead of acquiring the knowledge in the documents read, reflecting results from the psychology field [106], echoing our conclusion that *too much feedback should be considered harmful*.

6.1.2 Modeling Learner's Knowledge and Behavior

Our second research question focuses on understanding how behavior and knowledge change during a learner's search session. As learners interact with the search system, they submit queries, read documents, and, ideally, reflect on the new information encountered.

We explored this theme in Chapter 3, where we devised RULK, a framework for estimating and tracking learners' knowledge progress as they interact with more documents throughout their search system. Our framework assumes that a "target knowledge state" exists for a given learning topic and that the learner, perhaps unconsciously, tries to move closer to that reference level. As the learner interacts with the search system, we estimate their knowledge state by combining, in real-time, the content of the documents they read. We proposed the use of both a keyword (RULK_{KW}) and a language-model-based approach (RULK_{LM}) to represent the knowledge of the visited documents in fixed-length latent space embeddings. Therefore, the learner's knowledge state is estimated as the distance between a combination of the embeddings of the documents they read and the target knowledge state.

By applying RULK and its variants to the dataset generated by the user study in Chapter 2, we show that RULK has a considerable correlation with learning metrics. Further, by combining its keyword and language model variants into RULK_{KW+LM}, we show that syntactic and semantic concepts are complementary, leading to a more robust estimation of the learner's knowledge state. Thus, answering our second research question, we show that, by considering the content of the documents visited by a learner throughout their search session, it is possible to estimate their knowledge state throughout their learning process reliably. In follow-up work, we also show that a revised version of RULK, including information about entities present in the documents, further increases its reliability [43].

Chapter 4 also addresses **ORQ2**. There, we implement a simulating agent for IIR users, the SACSM, an evolution of the CSM framework proposed by Maxwell and Azzopardi [151]. By running 11,520 simulated user sections, with 144 types of learners over 8 learning topics, we can effectively simulate how learners behave during a SAL search session, mimicking behaviors seen by real-world learners.

We also briefly discuss **ORQ2** in Chapter 5. There, we propose a third model for better understanding learner's behaviors. By proposing KALM, we define latent variables from the learner's search session and discuss how they interact causally. We discuss how these

latent variables interact and how they can be used to understand the learner's behavior better. Our findings in that chapter show that some of the metrics more commonly associated with better learning outcomes (e.g., exploration-related metrics, such as dwell time and number of queries submitted) are not always good predictors for learning.

In summary, we answer **ORQ2** by proposing three markedly distinct approaches to model a learner's behavior and knowledge changes throughout their search session. The first, RULK, considers how the content of the documents read by a learner can be used to track how close they are to their learning goal. With SACSM, on the other hand, we propose to fully simulate how a user behaves during their learning process, simulating their clicking behavior and how their queries change over time as more knowledge is acquired. Finally, KALM shows how to model a learner and their interactions with the search engine using PLS, a tool for modeling causal relationships between latent variables.

6.1.3 Explaining and Predicting Learner's Knowledge Gains

Finally, our third research question aims to understand what *causes* learning during a learner's search session. While most current work in SAL focuses on computing correlations between metrics and learning outcomes, we argue in Chapter 5 that this only provides a partial view of the learning process. Therefore, we propose to use a causal modeling tool, namely, PLS-SEM, to better understand the causal relationships and interplay between latent variables during the learner's session.

We show that incorporating metrics of vocabulary usage on open-ended essays considerably increases the explanatory power for differences in knowledge gains during SAL search sessions compared to multiple-choice-only evaluations. However, given that manual evaluation of essays is expensive and impractical in most real-world scenarios, we propose a *mixed strategy* when assessing learning outcomes in SAL systems by measuring using both a multiple-choice questionnaire and some more straightforward to measure essay metrics, such as the number of key vocabulary terms used by the learner.

In Chapter 5, we also explored how different latent variables, such as the learner's effort on their querying strategy and the amount of knowledge they acquire, interact. We show that the two latent variables with higher explanatory power for learning outcomes are the quality of the documents read by the learner (as measured by indicators such as the textual complexity and density of vocabulary terms in the document) and the effort they put in writing their essays (measured by the number and reuse of terms in the essay). The quality of the queries also plays an important role indirectly, as it is a strong predictor of the quality of the documents retrieved by the search engine. We also show that contrary to most previous SAL research, exploration (e.g., the number of documents clicked by the learner or the maximum depth in a SERP) are not causal explainers for learning outcomes.

Answering **ORQ3**, we show that the quality of the documents retrieved by search systems and the effort learners put into writing their essays are the two main causal explainers for learning outcomes in SAL. We also show that exploration is not a causal explainer for learning outcomes for the three datasets used. Finally, we show that, by incorporating metrics of vocabulary usage in essays, we can achieve a higher explanatory power for the differences in learning outcomes during SAL search sessions.

Our findings in response to **ORQ3** indicate that the quality of documents retrieved by search systems and the effort learners put into writing their essays are the main causal

explainers for differences in learning outcomes in SAL. Additionally, our analysis indicates that exploration is not a significant factor in explaining learning outcomes. Furthermore, we showed that incorporating metrics of vocabulary usage in essays can provide a higher level of explanation for differences in learning outcomes during SAL search sessions.

6.2 Ethical and Societal Implications

People rely on the information available online to make decisions daily. According to research conducted by Turner and Rainie [220], 81% of US adults rely on information from the Internet “a lot” daily. Therefore, it is of utmost importance that the online information is accurate and trustworthy.

Phenomena like cognitive overload, similar to what some learners have experienced in our experiments in Chapter 2, considerably impact how searchers see untrustworthy content online [221]. Additionally, given the omnipresence of search engines, for-profit companies are incentivized to manipulate their content for optimizing for conversion and engagement instead of accuracy and trustworthiness, using tools like search engine optimization (SEO) and search engine marketing (SEM). These practices’ impact on knowledge acquisition is unknown and should be studied [222].

6.3 Moving Forward

The field of SAL is rapidly evolving. Therefore, this section discusses some directions towards which the field may move. Starting from the findings in this thesis, we discuss how future research in SAL (and, consequently, in IIR) should look shortly.

Before we start, however, we need to acknowledge the improvements in the generative capabilities of LLMs [223]¹. The use of LLMs is quickly becoming commonplace in search engines, and their impact on SAL-oriented search engines should not be ignored. Therefore, while the goal of this section is not to directly discuss the impact of LLMs, it is inevitable that some discussion involving these will be present.

Improving Learning-Optimized Search Engines

Learners will frequently recourse to search engines at some point during their learning process. Search systems should be aware of this scenario and provide adequate support throughout a learner’s search journey. As discussed in Chapter 1, this involves changes and optimizations in the search system’s front and back-ends.

The relevance of a passage should not be computed solely based on the content of the last user query, as commonly done in ad-hoc search scenarios [67, 68]. Instead, as discussed in Chapter 3, it should consider how the learner’s knowledge of the topic evolves. While some early works tackled the problem of crafting a learning-aware ranking, such as Syed and Collins-Thompson [67], this problem has not gathered much attention recently, despite its potential impact on learners’ knowledge acquisition.

To properly incorporate the learner’s knowledge into the computation of relevance for a document, we must first accurately track their knowledge throughout their search

¹Here, we use the term LLM when talking about language models with generative capabilities, primarily based on decoder-only models, such as GPT models [224], rather than encoder-only (e.g., BERT) or encoder-decoder language models (e.g., T5 [225])

session. While we explored this in Chapter 3, how to use this information fruitfully is unclear. Some early work exists, such as using one-armed-bandits [226] for recommending content to learners based on their knowledge. However, most current research, specifically in SAL, is focused on describing and predicting user behavior [187, 227, 228] rather than proposing specific interventions to the ranking algorithm itself.

Modern retrieval and ranking models like those based on transformers-based models such as BERT should gain significantly more attention. Modifying already indexed documents as the user's knowledge evolves is currently unfeasible in real-time. However, as we briefly discussed in Chapter 3, embedding-based models can capture the semantics of the documents visited by the learner throughout their search session. Therefore, such models can be used to dynamically modify the embedding of the learner's queries according to their knowledge state, pushing them to different locations within the embedding model's latent space.

Another research direction quickly gaining traction is using generative models, such as LLMs, in multiple places within a search engine [229]. For instance, using synthetic queries or documents for generating training data for retrievers has become commonplace [230–232]. Outside the training step, manipulating the user query by re-writing [233, 234], augmenting [235, 236] or helping the user crafting a better query [237] are some exciting directions for LLM-augmented retrieval. Specifically in the SAL domain, however, there is little exploration of their potential. Here, we envision these models will be used in a “transparent” manner (i.e., without direct interaction with the learner). For instance, such models can be used to (re-) write queries for learners based on their current knowledge level. Another possibility is to use these models for estimating the relevance of documents given the full context of documents previously read by the learner.

Another use of LLMs that is ever more common is using these models to estimate document relevance. Several promising approaches are already tackling this [238–241]. The main roadblocks towards their use in more search settings, especially in SAL, are their limited context length (i.e., how much content can be used at once) and their slow inference speed, given their large size. However, these two issues are quickly being addressed, with novel models with longer context windows [242] and parallel, multi-headed decoding [243] significantly increasing the practicality of employing LLMs for search and, specifically, SAL, with longer contexts.

Defining, Measuring, and Evaluating Learning

As discussed in Section 1.1.3, measuring learning is far from a solved problem. Defining what learning is and at what *level* we want to measure are the first steps toward defining adequate learning metrics.

In Chapter 5, we showed that relying exclusively on multiple-choice questionnaires for assessing learner's knowledge does not provide a complete view of their learning process. However, evaluating open-ended essays is expensive and impractical in most real-world scenarios. We showed that simple metrics, such as the frequency of vocabulary terms in essays, can be used as alternatives to complete essay evaluations. However, this is still a non-comprehensive view, with few opportunities for providing feedback to the learner.

Therefore, LLMs can be essential in evaluating and providing feedback for learners in their learning journey. It has been shown that LLMs can, to a certain extent, evaluate the

quality of written essays [244–246]. However, studying how they can be better utilized in the SAL context remains an unexplored research direction. Longer context windows could, for instance, provide a better overview of the content a learner needs to address and more context on what they have already learned.

Exploring how real-time evaluation of essays can be used to provide feedback to learners is also an exciting direction. High-quality, reliable, and real-time feedback opens a few interesting possibilities. First, evaluating long-form essays in the pre- and post-tests becomes feasible in real-world scenarios. Second, multiple, smaller-scale evaluations are also possible throughout the learner’s journey [59, 227]. However, this more constant, higher-quality feedback loop should be taken in context with the findings from Section 2 that too frequent feedback may be harmful. Therefore, assuming that a high-quality, cheap, and fast evaluation is possible, striking the correct balance between evaluating, providing feedback, and letting the user explore is an exciting direction for future research.

Hallucinations, where an LLM produces lexically correct but factually incorrect results, must be considered when developing automatic evaluation systems [247, 248]. While using retrieval augmentation (i.e., adding the content of search results to the input context of these models) may help with hallucinations, it does not entirely solve the issue [249, 250]. Hallucinations can be especially problematic in learning scenarios. Suppose an LLM provides unreliable feedback not backed by sound sources, or the content is fabricated. In that case, it can hurt not only the trust of the system but also the learner’s knowledge acquisition process. Therefore, while powerful, using LLMs in education should be done carefully considering their quality, with safeguards (such as retrieval augmentation) and frequent supervision by experts to ensure that their advantages are not outweighed by their disadvantages.

In the opposite direction, a fully automated evaluation of a learner’s knowledge, with no explicit tests and essays, is also an enticing direction. We have shown in Chapter 3 that estimating a learner’s knowledge level is possible solely based on the content of visited documents. Incorporating more robust evaluations and evaluations of the quality of visited documents and how much of the content was acquired by the learner could be helpful as a non-invasive but real-time estimation of the learner’s knowledge level.

Conversational Interfaces for Learning

With the popularization of high-quality conversational agents, such as ChatGPT², using conversational interfaces powered by LLMs in educational settings is increasingly common [251–253]. However, while general-purpose search engines have been making considerable efforts to incorporate conversational agents³, their implications for SAL-oriented sessions are not yet clear.

These conversational interfaces are slowly transforming how users interact with search engines. Some users are moving away from traditional search interfaces and prefer the conversational experience of these services⁴. In these settings, users may be disincensed to click on documents but instead read summaries and ask follow-up questions to

²<https://openai.com/blog/chatgpt/>

³At the moment of writing, Bing and You.com, two commercial search engines, have incorporated conversational search agents using GPT-4, and Google has released their Bard experiment as a search-assisted conversational agent

⁴<https://twitter.com/jdkelly/status/1598021488795586561>

these agents. Here, the impact of these agents on the learner's knowledge gain is still unknown. While they may provide more valuable and richer interactions with the content of the documents and the possibility of follow-up questions and clarifications, the impact of hallucinations and the quality of the sources have yet to be studied. Additionally, the closed-source nature of the most powerful models, such as GPT-4, may make them impenetrable to scrutiny and evaluation, making it even harder to evaluate the quality and reliability of the quality of content generated, especially for non-experts.

Therefore, while promising, using LLMs-backed conversational agents for SAL, even when augmented with retrieved results, must be studied in depth, primarily as it is already being used in real-world settings. Studying the impact of conversational-first search system interfaces and how learner behavior may change must be a priority for SAL researchers.

Causality In SAL

Chapter 5 discusses how causality remains a somewhat blind spot of SAL and IIR. Causal methods allow a better understanding of how learners' complex behavior and interactions with a search engine impact their learning outcomes. Causality also enables us to have a more principled approach to proposing interventions in the search engine itself by pointing at specific variables that are more likely to cause an increase in the acquired knowledge of learners.

Our results in Chapter 5 already point out that some commonly held concepts, such as that more exploration causes learners to learn more, do not always hold. Therefore, further research in causal methods in SAL should validate these results in a broader set of settings and explore different variables and interventions and their connections to better learning outcomes. We look forward to future studies using causal methods to describe, propose, and test interventions in SAL-oriented search engines.

Using search engines, be it in a traditional, ten-blue-links setting or as the backbone of a more personalized process, such as in a conversational agent, is and will continue to be one of the most common use cases for such engines. While considerable strides have been made recently towards better learning-oriented search systems, much remains to be done. From a more abstract level, measuring and understanding what causes learning is far from solved. As the goal of a SAL system is to support learners when searching, a clearer view of these is crucial to inform better interventions in such systems. In a more practical sense, LLMs have been causing perceptible changes in the whole IR field. These models' impacts on learning, especially in SAL, should not be underestimated. Rather, understanding these shifts and how to better use these tools in the search process, from query writing to ranking, question-answering, and evaluation, is of utmost importance for the Search-as-Learning field.

Bibliography

References

- [1] Robert M. Gagne. Contributions of learning to human development. *Psychological Review*, 75(3):177–191, 1968. ISSN 1939-1471. doi: 10.1037/h0025664.
- [2] A. H. Maslow. A theory of human motivation. *Psychological Review*, 50(4):370–396, 1943. ISSN 1939-1471. doi: 10.1037/h0054346.
- [3] Todd Bridgman, Stephen Cummings, and John Ballard. Who Built Maslow’s Pyramid? A History of the Creation of Management Studies’ Most Famous Symbol and Its Implications for Management Education. *Academy of Management Learning & Education*, 18(1):81–98, March 2019. ISSN 1537-260X. doi: 10.5465/amle.2017.0351.
- [4] Steven L. Kuhn. Signaling Theory and Technologies of Communication in the Paleolithic. *Biological Theory*, 9(1):42–50, March 2014. ISSN 1555-5550. doi: 10.1007/s13752-013-0156-5.
- [5] Lionel Casson. *Libraries in the Ancient World*. Yale University Press, 2002.
- [6] H. P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4):309–317, October 1957. ISSN 0018-8646. doi: 10.1147/rd.14.0309.
- [7] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham, 2022. doi: 10.1007/978-3-031-02181-7.
- [8] 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. *Aslib Proceedings*, 60(4):290–310, January 2008. ISSN 0001-253X. doi: 10.1108/00012530810887953.
- [9] Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, April 2006. ISSN 0001-0782. doi: 10.1145/1121949.1121979.
- [10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008. doi: 10.1017/CBO9780511809071.

-
- [11] Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. Supporting Complex Search Tasks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14*, pages 829–838, New York, NY, USA, November 2014. Association for Computing Machinery. doi: 10.1145/2661829.2661912.
- [12] Bernard J. Jansen, Amanda Spink, and Vinish Kathuria. How to Define Searching Sessions on Web Search Engines. In Olfa Nasraoui, Myra Spiliopoulou, Jaideep Srivastava, Bamshad Mobasher, and Brij Masand, editors, *Advances in Web Mining and Web Usage Analysis*, Lecture Notes in Computer Science, pages 92–109, Berlin, Heidelberg, 2007. Springer. doi: 10.1007/978-3-540-77485-3_6.
- [13] Christian Otto, Markus Rokicki, Georg Pardi, Wolfgang Gritz, Daniel Hienert, Ran Yu, Johannes von Hoyer, Anett Hoppe, Stefan Dietze, Peter Holtz, Yvonne Kammerer, and Ralph Ewerth. SaL-Lightning Dataset: Search and Eye Gaze Behavior, Resource Interactions and Knowledge Gain during Web Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR '22*, pages 347–352, New York, NY, USA, March 2022. Association for Computing Machinery. doi: 10.1145/3498366.3505835.
- [14] Gene Golovchinsky, Abdigani Diriye, and Tony Dunnigan. The future is in the past: Designing for exploratory search. In *Proceedings of the 4th Information Interaction in Context Symposium, IIX '12*, pages 52–61, New York, NY, USA, August 2012. Association for Computing Machinery. doi: 10.1145/2362724.2362738.
- [15] Nicholas J. Belkin, Robert N Oddy, and H M Brooks. ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2):61–71, January 1982. ISSN 0022-0418. doi: 10.1108/eb026722.
- [16] Janette R. Hill and Michael J. Hannafin. Teaching and learning in digital environments: The resurgence of resource-based learning. *Educational Technology Research and Development*, 49(3):37–52, September 2001. ISSN 1556-6501. doi: 10.1007/BF02504914.
- [17] Pertti Vakkari, Mikko Pennanen, and Sami Serola. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39(3):445–463, May 2003. ISSN 0306-4573. doi: 10.1016/S0306-4573(02)00031-6.
- [18] Barbara M. Wildemuth. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, 55(3): 246–258, 2004. ISSN 1532-2890. doi: 10.1002/asi.10367.
- [19] Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Information Retrieval Series*. Springer-Verlag, Berlin/Heidelberg, 2005. doi: 10.1007/1-4020-3851-8.
- [20] J. Patrick Biddix, Chung Joo Chung, and Han Woo Park. Convenience or credibility? A study of college student online research behaviors. *The Internet and Higher Education*, 14(3):175–182, July 2011. ISSN 10967516. doi: 10.1016/j.iheduc.2011.01.003.

- [21] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. Predicting users' domain knowledge from search behaviors. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1225–1226, New York, NY, USA, July 2011. Association for Computing Machinery. doi: 10.1145/2009916.2010131.
- [22] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. *ACM SIGIR Forum*, 46(1):2–32, May 2012. ISSN 0163-5840. doi: 10.1145/2215676.2215678.
- [23] Jingjing Liu, Nicholas J. Belkin, Xiangmin Zhang, and Xiaojun Yuan. Examining users' knowledge change in the task completion process. *Information Processing & Management*, 49(5):1058–1074, September 2013. ISSN 0306-4573. doi: 10.1016/j.ipm.2012.08.006.
- [24] Mathew J. Wilson and Max L. Wilson. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology*, 64(2):291–306, 2013. ISSN 1532-2890. doi: 10.1002/asi.22758.
- [25] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: A query log analysis of within-session learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 223–232, New York, NY, USA, February 2014. Association for Computing Machinery. doi: 10.1145/2556195.2556217.
- [26] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, 42(1):19–34, February 2016. ISSN 0165-5515. doi: 10.1177/0165551515615841.
- [27] Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, February 2016. ISSN 0165-5515. doi: 10.1177/0165551515615833.
- [28] 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*, CHIIR '16, pages 163–172, New York, NY, USA, March 2016. Association for Computing Machinery. doi: 10.1145/2854946.2854972.
- [29] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. Search as Learning (Dagstuhl Seminar 17092). *Dagstuhl Reports*, 7(2):135–162, 2017. ISSN 2192-5283. doi: 10.4230/DagRep.7.2.135.
- [30] Carsten Eickhoff, Jacek Gwizdka, Claudia Hauff, and Jiyin He. Introduction to the special issue on search as learning. *Information Retrieval Journal*, 20(5):399–402, October 2017. ISSN 1573-7659. doi: 10.1007/s10791-017-9315-9.

-
- [31] Rohail Syed and Kevyn Collins-Thompson. Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 191–200, New York, NY, USA, March 2018. Association for Computing Machinery. doi: 10.1145/3176349.3176397.
- [32] Souvick Ghosh, Manasa Rath, and Chirag Shah. Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-related Tasks. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 22–31, New York, NY, USA, March 2018. Association for Computing Machinery. doi: 10.1145/3176349.3176386.
- [33] Johannes von Hoyer, Georg Pardi, Yvonne Kammerer, and Peter Holtz. Metacognitive Judgments in Searching as Learning (SAL) Tasks: Insights on (Mis-) Calibration, Multimedia Usage, and Confidence. In *Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information*, SALMM '19, pages 3–10, New York, NY, USA, October 2019. Association for Computing Machinery. doi: 10.1145/3347451.3356730.
- [34] Nilavra Bhattacharya and Jacek Gwizdka. Measuring Learning During Search: Differences in Interactions, Eye-Gaze, and Semantic Similarity to Expert Knowledge. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, pages 63–71, New York, NY, USA, March 2019. Association for Computing Machinery. doi: 10.1145/3295750.3298926.
- [35] Hanrui Liu, Chang Liu, and Nicholas J. Belkin. Investigation of users' knowledge change process in learning-related search tasks. *Proceedings of the Association for Information Science and Technology*, 56(1):166–175, 2019. ISSN 2373-9231. doi: 10.1002/pr2.63.
- [36] Heather L. O'Brien, Andrea Kampen, Amelia W. Cole, and Kathleen Brennan. The Role of Domain Knowledge in Search as Learning. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, pages 313–317, New York, NY, USA, March 2020. Association for Computing Machinery. doi: 10.1145/3343413.3377989.
- [37] Georg Pardi, Johannes von Hoyer, Peter Holtz, and Yvonne Kammerer. The Role of Cognitive Abilities and Time Spent on Texts and Videos in a Multimodal Searching as Learning Task. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, pages 378–382, New York, NY, USA, March 2020. Association for Computing Machinery. doi: 10.1145/3343413.3378001.
- [38] Nilavra Bhattacharya and Jacek Gwizdka. Visualizing and Quantifying Vocabulary Learning During Search. In *Proceedings of the First International Workshop on Investigating Learning During Web Search*, page 4, Galway, Ireland, October 2020.
- [39] Cecilia di Sciascio, Eduardo Veas, Jordan Barria-Pineda, and Colleen Culley. Understanding the effects of control and transparency in searching as learning. In

- Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 498–509, Cagliari Italy, March 2020. ACM. doi: 10.1145/3377325.3377524.
- [40] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. How Do Active Reading Strategies Affect Learning Outcomes in Web Search? In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 368–375, Cham, 2021. Springer International Publishing. doi: 10.1007/978-3-030-72240-1_37.
- [41] Sara Salimzadeh, David Maxwell, and Claudia Hauff. The Impact of Entity Cards on Learning-Oriented Search Tasks. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, pages 63–72, New York, NY, USA, August 2021. Association for Computing Machinery. doi: 10.1145/3471158.3472255.
- [42] Kelsey Urgo and Jaime Arguello. Learning assessments in search-as-learning: A survey of prior work and opportunities for future research. *Information Processing & Management*, 59(2):102821, March 2022. ISSN 0306-4573. doi: 10.1016/j.ipm.2021.102821.
- [43] Dima El Zein, Arthur Câmara, Célia Da Costa Pereira, and Andrea Tettamanzi. RULKNE: Representing User Knowledge State in Search-as-Learning with Named Entities. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, pages 388–393, New York, NY, USA, March 2023. Association for Computing Machinery. doi: 10.1145/3576840.3578330.
- [44] Ranking Results – How Google Search Works. <https://archive.is/XsBDP>, August 2021.
- [45] L. Anderson, D. Krathwohl, and B. Bloom. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives. December 2000.
- [46] Kelsey Urgo, Jaime Arguello, and Robert Capra. Anderson and Krathwohl’s Two-Dimensional Taxonomy Applied to Task Creation and Learning Assessment. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, pages 117–124, New York, NY, USA, September 2019. Association for Computing Machinery. doi: 10.1145/3341981.3344226.
- [47] Andrei Broder. A taxonomy of web search. *ACM SIGIR Forum*, 36(2):3–10, September 2002. ISSN 0163-5840. doi: 10.1145/792550.792552.
- [48] Neil Selwyn. An investigation of differences in undergraduates’ academic use of the internet. *Active Learning in Higher Education*, 9(1):11–22, March 2008. ISSN 1469-7874. doi: 10.1177/1469787407086744.
- [49] Nicholas J. Belkin. The cognitive viewpoint in information science. *Journal of Information Science*, 16(1):11–15, February 1990. ISSN 0165-5515. doi: 10.1177/016555159001600104.

-
- [50] Marcia J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, January 1989. ISSN 0309-314X. doi: 10.1108/eb024320.
- [51] Peter Ingwersen. *Information Retrieval Interaction*. Taylor Graham, London, 1992.
- [52] Qi Guo, Dmitry Lagun, and Eugene Agichtein. Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 2050–2054, New York, NY, USA, October 2012. Association for Computing Machinery. doi: 10.1145/2396761.2398570.
- [53] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting User Knowledge Gain in Informational Search Sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, pages 75–84, New York, NY, USA, June 2018. Association for Computing Machinery. doi: 10.1145/3209978.3210064.
- [54] Christian Otto, Ran Yu, Georg Pardi, Johannes von Hoyer, Markus Rokicki, Anett Hoppe, Peter Holtz, Yvonne Kammerer, Stefan Dietze, and Ralph Ewerth. Predicting Knowledge Gain During Web Search Based on Multimedia Resource Consumption. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education, Lecture Notes in Computer Science*, pages 318–330, Cham, 2021. Springer International Publishing. doi: 10.1007/978-3-030-78292-4_26.
- [55] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. Predicting Retrieval Success Based on Information Use for Writing Tasks. In Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes, editors, *Digital Libraries for Open Knowledge, Lecture Notes in Computer Science*, pages 161–173, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-030-00066-0_14.
- [56] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. Predicting essay quality from search and writing behavior. *Journal of the Association for Information Science and Technology*, 72(7):839–852, 2021. ISSN 2330-1643. doi: 10.1002/asi.24451.
- [57] Dima El Zein and Célia da Costa Pereira. A Cognitive Agent Framework in Information Retrieval: Using User Beliefs to Customize Results. In Takahiro Uchiya, Quan Bai, and Iván Marsá Maestre, editors, *PRIMA 2020: Principles and Practice of Multi-Agent Systems, Lecture Notes in Computer Science*, pages 325–333, Cham, 2021. Springer International Publishing. doi: 10.1007/978-3-030-69322-0_21.
- [58] Yao Zhang and Chang Liu. Users’ Knowledge Use and Change during Information Searching Process: A Perspective of Vocabulary Usage. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 47–56, Virtual Event China, August 2020. ACM. doi: 10.1145/3383583.3398532.

- [59] Nirmal Roy, Felipe Moraes, and Claudia Hauff. Exploring Users' Learning Gains within Search Sessions. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, pages 432–436, New York, NY, USA, March 2020. Association for Computing Machinery. doi: 10.1145/3343413.3378012.
- [60] Arthur Câmara and Dima El-Zein. RULK: A Framework for Representing User Knowledge in Search-as-Learning. In *Design of Experimental Search & Information REtrieval Systems*, DESIRES '22, San Jose, CA, August 2022.
- [61] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. ScentBar: A Query Suggestion Interface Visualizing the Amount of Missed Relevant Information for Intrinsically Diverse Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 405–414, New York, NY, USA, July 2016. Association for Computing Machinery. doi: 10.1145/2911451.2911546.
- [62] Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. Searching to Learn with Instructional Scaffolding. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 209–218, New York, NY, USA, March 2021. Association for Computing Machinery. doi: 10.1145/3406522.3446012.
- [63] Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H. Chi. Signpost from the masses: Learning effects in an exploratory social tag search browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 625–634, New York, NY, USA, April 2009. Association for Computing Machinery. doi: 10.1145/1518701.1518797.
- [64] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, pages 229–238, New York, NY, USA, March 2021. Association for Computing Machinery. doi: 10.1145/3406522.3446025.
- [65] Luanne Freund, Rick Kopak, and Heather O'Brien. The Effects of textual environment on reading comprehension: Implications for searching as learning. *Journal of Information Science*, 42(1):79–93, February 2016. ISSN 0165-5515. doi: 10.1177/0165551515614472.
- [66] C. Liu, X. Song, H. Liu, and N.J. Belkin. Modeling Knowledge Change Behaviors in Learning-related Tasks. In *CEUR Workshop Proceedings*, volume 2699, 2020.
- [67] Rohail Syed and Kevyn Collins-Thompson. Optimizing search results for human learning goals. *Information Retrieval Journal*, 20(5):506–523, October 2017. ISSN 1573-7659. doi: 10.1007/s10791-017-9303-0.
- [68] Rohail Syed and Kevyn Collins-Thompson. Retrieval Algorithms Optimized for Human Learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 555–564,

- New York, NY, USA, August 2017. Association for Computing Machinery. doi: 10.1145/3077136.3080835.
- [69] Peter Bailey. User task understanding: A web search engine perspective, October 2012.
- [70] Marti A. Hearst. *Search User Interfaces*. Cambridge University Press, Cambridge, 2009. doi: 10.1017/CBO9781139644082.
- [71] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13*, pages 953–964, New York, NY, USA, May 2013. Association for Computing Machinery. doi: 10.1145/2488388.2488471.
- [72] Jingjing Liu and Nicholas J. Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10*, pages 26–33, New York, NY, USA, July 2010. Association for Computing Machinery. doi: 10.1145/1835449.1835457.
- [73] Max L. Wilson, Paul André, and mc schraefel. Backward highlighting: Enhancing faceted search. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, UIST '08*, pages 235–238, New York, NY, USA, October 2008. Association for Computing Machinery. doi: 10.1145/1449715.1449754.
- [74] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Towards Memorable Information Retrieval. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 69–76, Virtual Event Norway, September 2020. ACM. doi: 10.1145/3409256.3409830.
- [75] David J. Bell and Ian Ruthven. Searcher’s Assessments of Task Complexity for Web Searching. In Sharon McDonald and John Tait, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 57–71, Berlin, Heidelberg, 2004. Springer. doi: 10.1007/978-3-540-24752-4_5.
- [76] Yuelin Li and Nicholas J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, November 2008. ISSN 0306-4573. doi: 10.1016/j.ipm.2008.07.005.
- [77] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium, IIX '12*, pages 254–257, New York, NY, USA, August 2012. Association for Computing Machinery. doi: 10.1145/2362724.2362768.
- [78] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126, January 1995.

- [79] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
- [80] Rodrigo Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT, April 2020.
- [81] Darwin P. Hunt. The concept of knowledge and how to measure it. *Journal of Intellectual Capital*, 4(1):100–113, January 2003. ISSN 1469-1930. doi: 10.1108/14691930310455414.
- [82] Marjorie Wesche and T. Sima Paribakht. Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *The Canadian Modern Language Review*, 53(1): 13–40, October 1996. ISSN 0008-4506. doi: 10.3138/cmlr.53.1.13.
- [83] Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. Contrasting Search as a Learning Activity with Instructor-designed Learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, pages 167–176, New York, NY, USA, October 2018. Association for Computing Machinery. doi: 10.1145/3269206.3271676.
- [84] Rishita Kalyani and Ujwal Gadiraju. Understanding User Search Behavior Across Varying Cognitive Levels. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media, HT '19*, pages 123–132, New York, NY, USA, September 2019. Association for Computing Machinery. doi: 10.1145/3342220.3343643.
- [85] Dima El Zein and Célia da Costa Pereira. User’s Knowledge and Information Needs in Information Retrieval Evaluation. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22*, pages 170–178, New York, NY, USA, July 2022. Association for Computing Machinery. doi: 10.1145/3503252.3531325.
- [86] Wolfgang Gritz, Anett Hoppe, and Ralph Ewerth. On the Impact of Features and Classifiers for Measuring Knowledge Gain during Web Search - A Case Study. In *Proceedings of the Second International Workshop on Investigating Learning During Web Search (IWILDS'21) Co-Located with 29th ACM International Conference on Information and Knowledge Management (CIKM 2021)*, IWILDS'21, page 10, Gold Coast, Queensland, Australia, November 2021. CEUR-WS.
- [87] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, CHIIR '18*, pages 2–11, New York, NY, USA, March 2018. Association for Computing Machinery. doi: 10.1145/3176349.3176381.
- [88] Markus Kattenbeck and David Elswailer. Estimating Models Combining Latent and Measured Variables: A Tutorial on Basics, Applications and Current Developments

- in Structural Equation Models and their Estimation using PLS Path Modeling. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 375–377, New York, NY, USA, March 2018. Association for Computing Machinery. doi: 10.1145/3176349.3176899.
- [89] Pertti Vakkari and Salla Huuskonen. Search effort degrades search output but improves task outcome. *Journal of the American Society for Information Science and Technology*, 63(4):657–670, 2012. ISSN 1532-2890. doi: 10.1002/asi.21683.
- [90] Arthur Câmara, David Maxwell, and Claudia Hauff. Searching, Learning, and Subtopic Ordering: A Simulation-Based Analysis. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 142–156, Cham, 2022. Springer International Publishing. doi: 10.1007/978-3-030-99736-6_10.
- [91] Douglas C. Merrill, Brian J. Reiser, Michael Ranney, and J. Gregory Trafton. Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. *The Journal of the Learning Sciences*, 2(3):277–305, 1992. ISSN 1050-8406.
- [92] Ann L. Brown and Annemarie S. Palincsar. Guided, cooperative learning and individual knowledge acquisition. In *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, pages 393–451. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1989.
- [93] Barbara Rogoff. Adult assistance of children’s learning. *The contexts of school-based literacy*, 1986.
- [94] L. S. Vygotsky. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press, 1978. doi: 10.2307/j.ctvjf9vz4.
- [95] Catherine L. Smith and Soo Young Rieh. Knowledge-Context in Search Systems: Toward Information-Literate Actions. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, pages 55–62, New York, NY, USA, March 2019. Association for Computing Machinery. doi: 10.1145/3295750.3298940.
- [96] Sindunuraga Rikarno Putra, Felipe Moraes, and Claudia Hauff. SearchX: Empowering Collaborative Search Research. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1265–1268, New York, NY, USA, June 2018. Association for Computing Machinery. doi: 10.1145/3209978.3210163.
- [97] Rohail Syed, Kevyn Collins-Thompson, Paul N. Bennett, Mengqiu Teng, Shane Williams, Dr. Wendy W. Tay, and Shamsi Iqbal. Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation. In *Proceedings of The Web Conference 2020*, WWW '20, pages 1693–1703, New York, NY, USA, April 2020. Association for Computing Machinery. doi: 10.1145/3366423.3380240.

- [98] Shawn M. Glynn and Francis J. di Vesta. Outline and hierarchical organization as aids for study and retrieval. *Journal of Educational Psychology*, 69(2):89–95, 1977. ISSN 1939-2176. doi: 10.1037/0022-0663.69.2.89.
- [99] Priya Sharma and Michael J. Hannafin. Scaffolding in technology-enhanced learning environments. *Interactive Learning Environments*, 15(1):27–46, April 2007. ISSN 1049-4820. doi: 10.1080/10494820600996972.
- [100] Brian R. Belland. Instructional Scaffolding: Foundations and Evolving Definition. In Brian R. Belland, editor, *Instructional Scaffolding in STEM Education: Strategies and Efficacy Evidence*, pages 17–53. Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-02565-0_2.
- [101] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. Outline Generation: Understanding the Inherent Content Structure of Documents. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR’19, pages 745–754, New York, NY, USA, July 2019. Association for Computing Machinery. doi: 10.1145/3331184.3331208.
- [102] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. TREC Complex Answer Retrieval Overview. In *TREC*, (NIST) Special Publication. National Institute of Standards and Technology {NIST}, 2017.
- [103] Brendan Luyt. The inclusivity of Wikipedia and the drawing of expert boundaries: An examination of talk pages and reference lists. *Journal of the American Society for Information Science and Technology*, 63(9):1868–1878, 2012. ISSN 1532-2890. doi: 10.1002/asi.22671.
- [104] Henri G. Colt, Mohsen Davoudi, Septimiu Murgu, and Nazanin Zamanian Rohani. Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical Endoscopy*, 25(1):207–216, January 2011. ISSN 1432-2218. doi: 10.1007/s00464-010-1161-4.
- [105] John L. Shefelbine. Student Factors Related to Variability in Learning Word Meanings from Context. *Journal of Reading Behavior*, 22(1):71–97, March 1990. ISSN 0022-4111. doi: 10.1080/10862969009547695.
- [106] Stephan P. Swinnen, Richard A. Schmidt, Diane E. Nicholson, and Diane C. Shapiro. Information feedback for skill acquisition: Instantaneous knowledge of results degrades learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):706–716, 1990. ISSN 1939-1285. doi: 10.1037/0278-7393.16.4.706.
- [107] Richard E. Mayer, Emily Griffith, Ilana T. N. Jurkowitz, and Daniel Rothman. Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. *Journal of Experimental Psychology: Applied*, 14(4):329–339, December 2008. ISSN 1076-898X. doi: 10.1037/a0013835.
- [108] Lynne Anderson-Inman and Leigh Zeitz. Computer-Based Concept Mapping: Active Studying for Active Learners. *Computing Teacher*, 21(1), 1993. ISSN 0278-9175.

-
- [109] Katriina Byström and Kalervo Järvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, March 1995. ISSN 0306-4573. doi: 10.1016/0306-4573(95)80035-R.
- [110] Lydia Harbarth, Svea Delsing, Florian Richtscheid, Volkan Yücepur, Florian Feldmann, Milad Akhavanfar, Sven Manske, Julia Othlinghaus, and H. Ulrich Hoppe. *Learning by Tagging – Supporting Constructive Learning in Video-Based Environments*. Gesellschaft für Informatik e.V., 2018.
- [111] Nirmal Roy, Arthur Câmara, David Maxwell, and Claudia Hauff. Incorporating Widget Positioning in Interaction Models of Search Behaviour. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR ’21*, pages 53–62, New York, NY, USA, August 2021. Association for Computing Machinery. doi: 10.1145/3471158.3472243.
- [112] Ran Yu, Rui Tang, Markus Rokicki, Ujwal Gadiraju, and Stefan Dietze. Topic-independent modeling of user knowledge in informational search sessions. *Information Retrieval Journal*, 24(3):240–268, June 2021. ISSN 1386-4564, 1573-7659. doi: 10.1007/s10791-021-09391-7.
- [113] Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. Knowledge Tracing: A Survey. *ACM Computing Surveys*, 55(11):224:1–224:37, February 2023. ISSN 0360-0300. doi: 10.1145/3569576.
- [114] Benjamin S. Bloom. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. David McKay Company, New York, first edition edition, 1956.
- [115] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, January 2020. ISSN 0020-0255. doi: 10.1016/j.ins.2019.09.013.
- [116] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [117] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. CEDR: Contextualized Embeddings for Document Ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’19*, pages 1101–1104, New York, NY, USA, July 2019. Association for Computing Machinery. doi: 10.1145/3331184.3331317.
- [118] Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM ’18*, pages 497–506, New York, NY, USA, October 2018. Association for Computing Machinery. doi: 10.1145/3269206.3271800.

- [119] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. Multi-Stage Conversational Passage Retrieval: An Approach to Fusing Term Importance Estimation and Neural Query Rewriting. *ACM Transactions on Information Systems*, 39(4):48:1–48:29, September 2021. ISSN 1046-8188. doi: 10.1145/3446426.
- [120] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, October 2021.
- [121] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.
- [122] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MINLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, pages 5776–5788, Red Hook, NY, USA, December 2020. Curran Associates Inc.
- [123] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. MS MARCO: Benchmarking Ranking Models in the Large-Data Regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*, pages 1566–1576, New York, NY, USA, July 2021. Association for Computing Machinery. doi: 10.1145/3404835.3462804.
- [124] Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. Applying BERT to Document Retrieval with Birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 19–24, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3004.
- [125] Jianwei Shi, Christian Otto, Anett Hoppe, Peter Holtz, and Ralph Ewerth. Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality. In *Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information, SALMM ’19*, pages 11–19, New York, NY, USA, October 2019. Association for Computing Machinery. doi: 10.1145/3347451.3356731.
- [126] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning Video Representations From Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023.
- [127] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding, October 2023.

-
- [128] Chang Liu and Xiaoxuan Song. How do Information Source Selection Strategies Influence Users' Learning Outcomes'. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 257–260, New York, NY, USA, March 2018. Association for Computing Machinery. doi: 10.1145/3176349.3176876.
- [129] Yihan Lu and I-Han Hsiao. Personalized Information Seeking Assistant (PiSA): From programming information seeking to learning. *Information Retrieval Journal*, 20(5): 433–455, October 2017. ISSN 1573-7659. doi: 10.1007/s10791-017-9305-y.
- [130] Heather O'Brien, Amelia Cole, Andrea Kampen, and Kathy Brennan. The Effects of Domain and Search Expertise on Learning Outcomes in Digital Library Use. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 202–210, New York, NY, USA, March 2022. Association for Computing Machinery. doi: 10.1145/3498366.3505761.
- [131] Jingjing Liu, Chang Liu, and Nicholas J. Belkin. Predicting information searchers' topic knowledge at different search stages. *Journal of the Association for Information Science and Technology*, 67(11):2652–2666, 2016. ISSN 2330-1643. doi: 10.1002/asi.23606.
- [132] Michael Villano. Probabilistic student models: Bayesian Belief Networks and Knowledge Space Theory. In Claude Frasson, Gilles Gauthier, and Gordon I. McCalla, editors, *Intelligent Tutoring Systems*, Lecture Notes in Computer Science, pages 491–498, Berlin, Heidelberg, 1992. Springer. doi: 10.1007/3-540-55606-0_58.
- [133] Shuanghong Shen, Qi Liu, Enhong Chen, Han Wu, Zhenya Huang, Weihao Zhao, Yu Su, Haiping Ma, and Shijin Wang. Convolutional Knowledge Tracing: Modeling Individualization in Student Learning Process. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1857–1860, New York, NY, USA, July 2020. Association for Computing Machinery. doi: 10.1145/3397271.3401288.
- [134] Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. Towards an Appropriate Query, Key, and Value Computation for Knowledge Tracing. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, L@S '20, pages 341–344, New York, NY, USA, August 2020. Association for Computing Machinery. doi: 10.1145/3386527.3405945.
- [135] Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. Context-Aware Attentive Knowledge Tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 2330–2339, New York, NY, USA, August 2020. Association for Computing Machinery. doi: 10.1145/3394486.3403282.
- [136] Shalini Pandey and Jaideep Srivastava. RKT: Relation-Aware Self-Attention for Knowledge Tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 1205–1214, New York, NY,

- USA, October 2020. Association for Computing Machinery. doi: 10.1145/3340531.3411994.
- [137] Dongmin Shin, Yugeun Shim, Hangyeol Yu, Seewoo Lee, Byungsoo Kim, and Youngduck Choi. SAINT+: Integrating Temporal Features for EdNet Correctness Prediction. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, pages 490–496, New York, NY, USA, April 2021. Association for Computing Machinery. doi: 10.1145/3448139.3448188.
- [138] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. Graph-based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. In *IEEE/WIC/ACM International Conference on Web Intelligence, WI '19*, pages 156–163, New York, NY, USA, October 2019. Association for Computing Machinery. doi: 10.1145/3350546.3352513.
- [139] Shiwei Tong, Qi Liu, Wei Huang, Zhenya Hunag, Enhong Chen, Chuanren Liu, Haiping Ma, and Shijin Wang. Structure-Based Knowledge Tracing: An Influence Propagation View. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 541–550, November 2020. doi: 10.1109/ICDM50108.2020.00063.
- [140] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3):243–266, August 2009. ISSN 1573-1391. doi: 10.1007/s11257-009-9063-7.
- [141] Zach A Pardos, Ryan S.J.D Baker, Maria San Pedro, Sujith M Gowda, and Supreeth M Gowda. Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics*, 1(1):107–128, May 2014. ISSN 1929-7750. doi: 10.18608/jla.2014.11.6.
- [142] Carol Collier Kuhlthau. Developing a Model of the Library Search Process: Cognitive and Affective Aspects. *RQ*, 28(2):232–242, 1988. ISSN 0033-7072.
- [143] Leif Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 15–24, New York, NY, USA, July 2011. Association for Computing Machinery. doi: 10.1145/2009916.2009923.
- [144] Leif Azzopardi. Modelling interaction with economic models of search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 3–12, New York, NY, USA, July 2014. Association for Computing Machinery. doi: 10.1145/2600428.2609574.
- [145] Norbert Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, June 2008. ISSN 1573-7659. doi: 10.1007/s10791-008-9045-0.

-
- [146] Alistair Moffat, Falk Scholer, and Paul Thomas. Models and metrics: IR evaluation as a user process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, ADCS '12, pages 47–54, New York, NY, USA, December 2012. Association for Computing Machinery. doi: 10.1145/2407085.2407092.
- [147] Leif Azzopardi and Guido Zuccon. Two Scrolls or One Click: A Cost Model for Browsing Search Results. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauß, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 696–702, Cham, 2016. Springer International Publishing. doi: 10.1007/978-3-319-30671-1_55.
- [148] Abhijith Kashyap, Vagelis Hristidis, and Michalis Petropoulos. FACeTOR: Cost-driven exploration of faceted query results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 719–728, New York, NY, USA, October 2010. Association for Computing Machinery. doi: 10.1145/1871437.1871530.
- [149] David Maxwell and Leif Azzopardi. Information Scent, Searching and Stopping. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 210–222, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-319-76941-7_16.
- [150] Wan-Ching Wu, Diane Kelly, and Avneesh Sud. Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 557–566, New York, NY, USA, July 2014. Association for Computing Machinery. doi: 10.1145/2600428.2609626.
- [151] David Maxwell and Leif Azzopardi. Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1141–1144, New York, NY, USA, July 2016. Association for Computing Machinery. doi: 10.1145/2911451.2911469.
- [152] David Maxwell and Leif Azzopardi. Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 731–740, New York, NY, USA, October 2016. Association for Computing Machinery. doi: 10.1145/2983323.2983805.
- [153] Peter Pirolli and Stuart Card. Information foraging. *Psychological Review*, 106(4): 643–675, 1999. ISSN 1939-1471. doi: 10.1037/0033-295X.106.4.643.
- [154] David Ellis. Modeling the Information-Seeking Patterns of Academic Researchers: A Grounded Theory Approach. *The Library Quarterly: Information, Community, Policy*, 63(4):469–486, 1993. ISSN 0024-2519.

- [155] Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 115–122, New York, NY, USA, July 2009. Association for Computing Machinery. doi: 10.1145/1571941.1571963.
- [156] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 611–620, New York, NY, USA, October 2011. Association for Computing Machinery. doi: 10.1145/2063576.2063668.
- [157] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Modeling behavioral factors in interactive information retrieval. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 2297–2302, New York, NY, USA, October 2013. Association for Computing Machinery. doi: 10.1145/2505515.2505660.
- [158] Paul Thomas, Alistair Moffat, Peter Bailey, and Falk Scholer. Modeling decision points in user search behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*, IliX '14, pages 239–242, New York, NY, USA, August 2014. Association for Computing Machinery. doi: 10.1145/2637002.2637032.
- [159] Leif Azzopardi, Kalervo Järvelin, Jaap Kamps, and Mark D. Smucker. Report on the SIGIR 2010 workshop on the simulation of interaction. *ACM SIGIR Forum*, 44(2): 35–47, January 2011. ISSN 0163-5840. doi: 10.1145/1924475.1924484.
- [160] Andy Peytchev, Mick P. Couper, Sean Esteban McCabe, and Scott D. Crawford. Web Survey Design: Paging versus Scrolling. *The Public Opinion Quarterly*, 70(4):596–607, 2006. ISSN 0033-362X.
- [161] Michael Albers and Loel Kim. Information design for the small-screen interface: An overview of web design issues for personal digital assistants. *Technical Communication*, 49(1):45–60, 2002.
- [162] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk Scholer. QWERTY: The Effects of Typing on Web Search Behavior. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, pages 281–284, New York, NY, USA, March 2018. Association for Computing Machinery. doi: 10.1145/3176349.3176872.
- [163] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Time Pressure and System Delays in Information Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 767–770, New York, NY, USA, August 2015. Association for Computing Machinery. doi: 10.1145/2766462.2767817.
- [164] David Maxwell and Leif Azzopardi. Stuck in traffic: How temporal delays affect search behaviour. In *Proceedings of the 5th Information Interaction in Context Symposium*, IliX '14, pages 155–164, New York, NY, USA, August 2014. Association for Computing Machinery. doi: 10.1145/2637002.2637021.

-
- [165] Eric Schurman and Jake Brutlag. Performance related changes and their user impact. In *O'Reilly Velocity Conference*, 2009.
- [166] ChengXiang Zhai, William W. Cohen, and John Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. *ACM SIGIR Forum*, 49(1):2–9, June 2015. ISSN 0163-5840. doi: 10.1145/2795403.2795405.
- [167] Zhengbao Jiang, Ji-Rong Wen, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, and Ming Yue. Learning to Diversify Search Results via Subtopic Attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 545–554, New York, NY, USA, August 2017. Association for Computing Machinery. doi: 10.1145/3077136.3080805.
- [168] Wei Dai and Rohini Srihari. Minimal document set retrieval. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 752–759, New York, NY, USA, October 2005. Association for Computing Machinery. doi: 10.1145/1099554.1099735.
- [169] Tu Ngoc Nguyen and Nattiya Kanhabua. Leveraging Dynamic Query Subtopics for Time-Aware Search Result Diversification. In Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 222–234, Cham, 2014. Springer International Publishing. doi: 10.1007/978-3-319-06028-6_19.
- [170] Guido Zuccon, Leif Azzopardi, Claudia Hauff, and C.J. Keith van Rijsbergen. Estimating interference in the QPRP for subtopic retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 741–742, New York, NY, USA, July 2010. Association for Computing Machinery. doi: 10.1145/1835449.1835593.
- [171] Toru Takaki, Atsushi Fujii, and Tetsuya Ishikawa. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 399–405, New York, NY, USA, November 2004. Association for Computing Machinery. doi: 10.1145/1031171.1031251.
- [172] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 59–68, New York, NY, USA, July 1993. Association for Computing Machinery. doi: 10.1145/160688.160695.
- [173] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pages 313–322, New York, NY, USA, October 2015. Association for Computing Machinery. doi: 10.1145/2806416.2806476.

- [174] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. An Initial Investigation into Fixed and Adaptive Stopping Strategies. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 903–906, New York, NY, USA, August 2015. Association for Computing Machinery. doi: 10.1145/2766462.2767802.
- [175] David Martin Maxwell. *Modelling Search and Stopping in Interactive Information Retrieval*. PhD thesis, University of Glasgow, 2019.
- [176] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. A query model based on normalized log-likelihood. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1903–1906, New York, NY, USA, November 2009. Association for Computing Machinery. doi: 10.1145/1645953.1646261.
- [177] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990, December 2012. ISSN 1554-3528. doi: 10.3758/s13428-012-0210-4.
- [178] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. Differences in the Use of Search Assistance for Tasks of Varying Complexity. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 23–32, New York, NY, USA, August 2015. Association for Computing Machinery. doi: 10.1145/2766462.2767741.
- [179] Bernard J. Jansen, Danielle Booth, and Brian Smith. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, 45(6): 643–663, November 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2009.05.004.
- [180] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. Development and Evaluation of Search Tasks for IIR Experiments using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 101–110, New York, NY, USA, September 2015. Association for Computing Machinery. doi: 10.1145/2808194.2809465.
- [181] Joe F. Hair, Christian M. Ringle, and Marko Sarstedt. PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, 19(2):139–152, April 2011. ISSN 1069-6679. doi: 10.2753/MTP1069-6679190202.
- [182] Adamantios Diamantopoulos and Heidi M. Winklhofer. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, 38(2):269–277, 2001. ISSN 0022-2437.
- [183] Ned Kock and Pierre Hadaya. Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal*, 28(1):227–261, 2018. ISSN 1365-2575. doi: 10.1111/isj.12131.
- [184] Marko Sarstedt, Christian M. Ringle, Jörg Henseler, and Joseph F. Hair. On the Emancipation of PLS-SEM: A Commentary on Rigdon (2012). *Long Range Planning*, 47(3):154–160, June 2014. ISSN 0024-6301. doi: 10.1016/j.lrp.2014.02.007.

-
- [185] Claes Cassel, Peter Hackl, and Anders Westlund. Robustness of Partial Least-Squares Method for Estimating Latent Variable Quality Structures. *Journal of Applied Statistics*, 26:435–446, February 1999. doi: 10.1080/02664769922322.
- [186] Jacek Gwizdka and Xueshu Chen. Towards Observable Indicators of Learning on Search. *Proceedings of the Second International Workshop on Search as Learning*, page 3, July 2016.
- [187] Kelsey Urgo and Jaime Arguello. Understanding the “Pathway” Towards a Searcher’s Learning Objective. *ACM Transactions on Information Systems*, 40(4): 77:1–77:43, January 2022. ISSN 1046-8188. doi: 10.1145/3495222.
- [188] Yu Chi, Shuguang Han, Daqing He, and Rui Meng. Exploring Knowledge Learning in Collaborative Information Seeking Process. In *Proceedings of the Second International Workshop on Search as Learning*, July 2016.
- [189] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi, Alan Juffs, and Lois Wilson. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20(1):73–98, 2010. ISSN 1560-4306.
- [190] Diego Demaree, Halszka Jarodzka, Saskia Brand-Gruwel, and Yvonne Kammerer. The Influence of Device Type on Querying Behavior and Learning Outcomes in a Searching as Learning Task with a Laptop or Smartphone. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR ’20, pages 373–377, New York, NY, USA, March 2020. Association for Computing Machinery. doi: 10.1145/3343413.3378000.
- [191] Srishti Palani, Zijian Ding, Stephen MacNeil, and Steven P. Dow. The “Active Search” Hypothesis: How Search Strategies Relate to Creative Learning. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR ’21, pages 325–329, New York, NY, USA, March 2021. Association for Computing Machinery. doi: 10.1145/3406522.3446046.
- [192] Edward Cutrell and Zhiwei Guan. What are you looking for? an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 407–416. Association for Computing Machinery, New York, NY, USA, April 2007.
- [193] Nilavra Bhattacharya and Jacek Gwizdka. Relating eye-tracking measures with changes in knowledge on search tasks. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA ’18, pages 1–5, New York, NY, USA, June 2018. Association for Computing Machinery. doi: 10.1145/3204493.3204579.
- [194] Michael J. Cole, Jacek Gwizdka, Chang Liu, Nicholas J. Belkin, and Xiangmin Zhang. Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, 49(5):1075–1091, September 2013. ISSN 0306-4573. doi: 10.1016/j.ipm.2012.08.004.

- [195] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09*, pages 132–141, New York, NY, USA, February 2009. Association for Computing Machinery. doi: 10.1145/1498759.1498819.
- [196] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. Modeling the usefulness of search results as measured by information use. *Information Processing & Management*, 56(3):879–894, May 2019. ISSN 0306-4573. doi: 10.1016/j.ipm.2019.02.001.
- [197] Kumaripaba Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces, UII '16*, pages 359–369, New York, NY, USA, March 2016. Association for Computing Machinery. doi: 10.1145/2856767.2856786.
- [198] Jaime Arguello. Predicting Search Task Difficulty. In Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval, Lecture Notes in Computer Science*, pages 88–99, Cham, 2014. Springer International Publishing. doi: 10.1007/978-3-319-06028-6_8.
- [199] Joseph F. Hair, William C. Black, and Barry J. Babin. *Multivariate Data Analysis: A Global Perspective*. Pearson Education, 2010.
- [200] Joseph Hair, G. Tomas M. Hult, Christian Ringle, and Marko Sarstedt. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. January 2022. doi: 10.1007/978-3-030-80519-7.
- [201] Edward E. Rigdon. Rethinking Partial Least Squares Path Modeling: In Praise of Simple Methods. *Long Range Planning*, 45(5):341–358, October 2012. ISSN 0024-6301. doi: 10.1016/j.lrp.2012.09.010.
- [202] Pertti Vakkari. The Usefulness of Search Results: A Systematization of Types and Predictors. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20*, pages 243–252, New York, NY, USA, March 2020. Association for Computing Machinery. doi: 10.1145/3343413.3377955.
- [203] Jiaxin Mao, Yiqun Liu, Huanbo Luan, Min Zhang, Shaoping Ma, Hengliang Luo, and Yuntao Zhang. Understanding and Predicting Usefulness Judgment in Web Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*, pages 1169–1172, New York, NY, USA, August 2017. Association for Computing Machinery. doi: 10.1145/3077136.3080750.
- [204] Kenneth A. Bollen. *Structural Equations with Latent Variables*. Structural Equations with Latent Variables. John Wiley & Sons, Oxford, England, 1989. doi: 10.1002/9781118619179.

-
- [205] Claes Fornell and Fred L. Bookstein. Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *JMR, Journal of Marketing Research (pre-1986)*, 19(000004):440, November 1982. ISSN 00222437.
- [206] Christian Nitzl. The use of partial least squares structural equation modelling (PLS-SEM) in management accounting research: Directions for future theory development. *Journal of Accounting Literature*, 37:19–35, December 2016. ISSN 0737-4607. doi: 10.1016/j.acclit.2016.09.003.
- [207] Jesse Chandler, Gabriele Paolacci, and Pam Mueller. Risks and Rewards of Crowdsourcing Marketplaces. In Pietro Michelucci, editor, *Handbook of Human Computation*, pages 377–392. Springer, New York, NY, 2013. doi: 10.1007/978-1-4614-8806-4_30.
- [208] Florian Schmidt-Weigand and Katharina Scheiter. The role of spatial descriptions in learning from multimedia. *Computers in Human Behavior*, 27(1):22–28, January 2011. ISSN 0747-5632. doi: 10.1016/j.chb.2010.05.007.
- [209] Christian M. Ringle, Sven Wende, and Jan-Michael Becker. SmartPLS 4. Technical report, 2022.
- [210] Joseph F. Hair, Jeffrey J. Risher, Marko Sarstedt, and Christian M. Ringle. When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1):2–24, January 2019. ISSN 0955-534X. doi: 10.1108/EBR-11-2018-0203.
- [211] K. G. Jöreskog. Simultaneous factor analysis in several populations. *Psychometrika*, 36(4):409–426, December 1971. ISSN 1860-0980. doi: 10.1007/BF02291366.
- [212] Adamantios Diamantopoulos, Marko Sarstedt, Christoph Fuchs, Petra Wilczynski, and Sebastian Kaiser. Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, 40(3):434–449, May 2012. ISSN 1552-7824. doi: 10.1007/s11747-011-0300-3.
- [213] Aimee L. Drolet and Donald G. Morrison. Do We Really Need Multiple-Item Measures in Service Research? *Journal of Service Research*, 3(3):196–204, February 2001. ISSN 1094-6705. doi: 10.1177/109467050133001.
- [214] Jörg Henseler, Christian M. Ringle, and Rudolf R. Sinkovics. The use of partial least squares path modeling in international marketing. In Rudolf R. Sinkovics and Pervez N. Ghauri, editors, *New Challenges to International Marketing*, volume 20 of *Advances in International Marketing*, pages 277–319. Emerald Group Publishing Limited, January 2009. doi: 10.1108/S1474-7979(2009)0000020014.
- [215] Sascha Raithel, Marko Sarstedt, Sebastian Scharf, and Manfred Schwaiger. On the value relevance of customer satisfaction. Multiple drivers and multiple markets. *Journal of the Academy of Marketing Science*, 40(4):509–525, July 2012. ISSN 1552-7824. doi: 10.1007/s11747-011-0247-4.

- [216] D. I. Newble, A. Baxter, and R. G. Elmslie. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education*, 13(4):263–268, July 1979. ISSN 0308-0110. doi: 10.1111/j.1365-2923.1979.tb01511.x.
- [217] James W. Pellegrino. A Learning Sciences Perspective on the Design and Use of Assessment in Education. In R. Keith Sawyer, editor, *The Cambridge Handbook of the Learning Sciences*, Cambridge Handbooks in Psychology, pages 233–252. Cambridge University Press, Cambridge, 2 edition, 2014. doi: 10.1017/CBO9781139519526.015.
- [218] Marc Bron, Jasmijn van Gorp, Frank Nack, Maarten de Rijke, Andrei Vishneuski, and Sonja de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’12, pages 425–434, New York, NY, USA, August 2012. Association for Computing Machinery. doi: 10.1145/2348283.2348342.
- [219] Jingjing Liu and Nicholas J. Belkin. Searching vs. writing: Factors affecting information use task performance. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012. ISSN 1550-8390. doi: 10.1002/meet.14504901127.
- [220] Erica Turner and Lee Rainie. Most Americans rely on their own research to make big decisions, and that often means online searches, March 2020.
- [221] Li Shi, Nilavra Bhattacharya, Anubrata Das, and Jacek Gwizdka. True or false? Cognitive load when reading COVID-19 news headlines: An eye-tracking study. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR ’23, pages 107–116, New York, NY, USA, March 2023. Association for Computing Machinery. doi: 10.1145/3576840.3578290.
- [222] Sebastian Schultheiß. How search engine marketing influences user knowledge gain: Development and empirical testing of an information search behavior model. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR ’23, pages 475–478, New York, NY, USA, March 2023. Association for Computing Machinery. doi: 10.1145/3576840.3578297.
- [223] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020.
- [224] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.
- [225] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document Ranking with a Pre-trained Sequence-to-Sequence Model, March 2020.

-
- [226] Ethan Prihar, Adam Sales, and Neil Heffernan. A Bandit You Can Trust. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 106–115, Limassol Cyprus, June 2023. ACM. doi: 10.1145/3565472.3592955.
- [227] Dima El Zein and Célia Da Costa Pereira. The Evolution of User Knowledge during Search-as-Learning Sessions: A Benchmark and Baseline. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR '23*, pages 454–458, New York, NY, USA, March 2023. Association for Computing Machinery. doi: 10.1145/3576840.3578273.
- [228] Bogeum Choi, Jaime Arguello, and Robert Capra. Understanding Procedural Search Tasks “in the Wild”. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR '23*, pages 24–33, New York, NY, USA, March 2023. Association for Computing Machinery. doi: 10.1145/3576840.3578302.
- [229] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Large Search Model: Redefining Search Stack in the Era of LLMs, October 2023.
- [230] Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval, May 2023.
- [231] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. Promptagator: Few-shot Dense Retrieval From 8 Examples, September 2022.
- [232] Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. A Test Collection of Synthetic Documents for Training Rankers: ChatGPT vs. Human Experts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23*, pages 5311–5315, New York, NY, USA, October 2023. Association for Computing Machinery. doi: 10.1145/3583780.3615111.
- [233] Zhiyu Chen, Jason Choi, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. Generate-then-Retrieve: Intent-Aware FAQ Retrieval in Product Search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 763–771, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.73.
- [234] Pedro Faustini, Zhiyu Chen, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. Answering Unanswered Questions through Semantic Reformulations in Spoken QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 729–743, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.70.
- [235] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.99.

- [236] Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James Glass. Expand, Rerank, and Retrieve: Query Reranking for Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12131–12147, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.768.
- [237] Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware Retrieval with Instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.225.
- [238] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '23*, pages 39–50, New York, NY, USA, August 2023. Association for Computing Machinery. doi: 10.1145/3578337.3605136.
- [239] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences, September 2023.
- [240] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models, September 2023.
- [241] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent, April 2023.
- [242] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madihan Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective Long-Context Scaling of Foundation Models. <https://arxiv.org/abs/2309.16039v1>, September 2023.
- [243] Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, and Tri Dao. Medusa, September 2023.
- [244] Cheng-Han Chiang and Hung-yi Lee. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870.
- [245] Kathrin Seßler, Tao Xiang, Lukas Bogenrieder, and Enkelejda Kasneci. PEER: Empowering Writing with Large Language Models. In Olga Viberg, Ioana Jivet, Pedro J.

- Muñoz-Merino, Maria Perifanou, and Tina Papathoma, editors, *Responsive and Sustainable Educational Futures*, Lecture Notes in Computer Science, pages 755–761, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-42682-7_73.
- [246] Alaa Abd-alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Pdraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Medical Education*, 9:e48291, June 2023. ISSN 2369-3762. doi: 10.2196/48291.
- [247] Yijun Xiao and William Yang Wang. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.236.
- [248] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248:1–248:38, March 2023. ISSN 0360-0300. doi: 10.1145/3571730.
- [249] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.320.
- [250] Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2436, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.136.
- [251] Hyangeun Ji, Insook Han, and Yujung Ko. A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 55(1):48–63, January 2023. ISSN 1539-1523. doi: 10.1080/15391523.2022.2142873.
- [252] Lucas Kohnke, Benjamin Luke Moorhouse, and Di Zou. ChatGPT for Language Teaching and Learning. *RELC Journal*, 54(2):537–550, August 2023. ISSN 0033-6882. doi: 10.1177/00336882231162868.
- [253] Rubén Pérez-Mercado, Antonio Balderas, Andrés Muñoz, Juan Francisco Cabrera, Manuel Palomo-Duarte, and Juan Manuel Dodero. ChatbotSQL: Conversational agent to support relational database query language learning. *SoftwareX*, 22:101346, May 2023. ISSN 2352-7110. doi: 10.1016/j.softx.2023.101346.
- [254] Arthur Câmara and Craig Macdonald. Dockerising Terrier for The Open-Source IR Replicability Challenge (OSIRRC 2019). *Open-Source IR Replicability Challenge 2019*, 2019.

- [255] Sarah Ibrahimi, Shuo Chen, Devanshu Arya, Arthur Câmara, Yunlu Chen, Tanja Crijns, Maurits van der Goes, Thomas Mensink, Emiel van Miltenburg, Daan Odijk, William Thong, Jiaojiao Zhao, and Pascal Mettes. Interactive Exploration of Journalistic Video Footage through Multimodal Semantic Matching. In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, pages 2196–2198, New York, NY, USA, October 2019. Association for Computing Machinery. doi: 10.1145/3343031.3350597.
- [256] Arthur Câmara and Claudia Hauff. Diagnosing BERT with Retrieval Heuristics. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 605–618, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-45439-5_40.
- [257] Arthur Câmara and Claudia Hauff. Moving Stuff Around: A study on the efficiency of moving documents into memory for Neural IR models. In *Workshop on Reaching Efficiency in Neural Information Retrieval, the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid Spain, July 2022.
- [258] Gustavo Penha, Arthur Câmara, and Claudia Hauff. Evaluating the robustness of retrieval pipelines with query variation generators. In *European Conference on Information Retrieval*, pages 397–412. Springer, 2022.
- [259] Saher Esmeir, Arthur Câmara, and Edgar Meij. Entity Retrieval from Multilingual Knowledge Graphs. In *Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 1–15, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.mrl-1.1.
- [260] On the effects of automatically generated adjunct questions for search as learning, March 2024.

Summary

At the core of Search-as-Learning (SAL), as a sub-field of Interactive Information Retrieval (IIR), is exploring how people use search engines to acquire knowledge. Inherently interactive, the knowledge acquisition process via a search system involves learners posing queries, analyzing content, and incorporating novel knowledge. Being this iterative process, learning-oriented systems should be designed to support learners in their search process. In this thesis, we study how to design such systems by leveraging concepts from Natural Language Processing (NLP), Information Retrieval (IR), and the learning sciences.

We begin this thesis by proposing interventions in a search system to better support learners in their journey with ideas based on instructional scaffolding from the learning sciences. The data collected from this user study gives us insights into how learners interact with search systems and provides us with a rich dataset for the rest of the thesis. With that data available, the rest of this thesis focuses on modeling learner behavior using different paradigms and frameworks. We begin by simulating learners' behavior, proposing a novel searcher model based on the idea that learning topics are usually subdivided into subtopics. Then, we propose a framework for tracking and predicting a learner's knowledge state throughout their search session. By using only the information in the documents read by the learner, this framework can accurately predict the learner's knowledge at the end of their session.

Finally, we look into SAL using the underexplored perspective of causality. While most previous works, including our own, mainly look into the correlations between behavior and learning, we take PLS-SEM, a causal modeling technique, to study the causal relationships between the different variables involved in the learning process and how this intricate and complex interactive process unfolds. From this analysis, we not only identify some interesting causal relationships, such as the indirect impact of query quality on learning but also show that the common practice of assessing learning by multiple-choice questions does not fully capture the variability of the learning process.

This thesis, therefore, contributes to the field of Search-as-Learning by making all of our datasets and code for simulating and modeling learner behavior available. We also provide several insights into how learners behave while searching and the impact of these behaviors on their learning outcomes. We hope the findings in this thesis can be used as foundations for more principled improvements in learning-oriented search systems.

Samenvatting

De kern van Search-as-Learning (SAL), als subgebied van Interactive Information Retrieval (IIR), is het onderzoeken hoe mensen zoekmachines gebruiken om kennis te verwerven. Het kennisverwervingsproces via een zoekstelsel is inherent interactief en houdt in dat leerlingen vragen stellen, inhoud analyseren en nieuwe kennis integreren. Aangezien dit een iteratief proces is, moeten leergerichte systemen worden ontworpen om leerlingen te ondersteunen in hun zoekproces.

We beginnen dit proefschrift met het voorstellen van interventies in een zoekstelsel om leerlingen beter te ondersteunen op hun reis met ideeën gebaseerd op instructie steigers uit de leerwetenschappen. De gegevens die uit dit gebruikersonderzoek zijn verzameld, geven ons inzicht in de manier waarop leerlingen omgaan met zoeksystemen en bieden ons een rijke dataset voor de rest van het proefschrift. Nu deze gegevens beschikbaar zijn, richt de rest van dit proefschrift zich op het modelleren van het gedrag van leerlingen met behulp van verschillende paradigma's en raamwerken. We beginnen met het simuleren van het gedrag van leerlingen, waarbij we een nieuw zoekmodel voorstellen, gebaseerd op het idee dat leeronderwerpen gewoonlijk onderverdeeld zijn in subonderwerpen. Vervolgens stellen we een raamwerk voor het volgen en voorspellen van de kennisstatus van een leerling tijdens zijn zoeksessie. Door alleen de informatie te gebruiken in de documenten die de leerling leest, kan dit raamwerk de kennis van de leerling aan het einde van de sessie nauwkeurig voorspellen.

Ten slotte onderzoeken we SAL vanuit het onderbelichte perspectief van causaliteit. In tegenstelling tot de meeste eerdere werken, waaronder die van ons, kijken we vooral naar de correlaties tussen gedrag en leren, gebruiken we PLS-SEM, een causale modeleringstechniek, om de causale relaties te bestuderen tussen de verschillende variabelen die betrokken zijn bij het leerproces en hoe deze complexe en complexe interactief proces zich ontvouwt. Uit deze analyse identificeren we niet alleen een aantal interessante causale relaties, zoals de indirecte impact van de kwaliteit van de vragen op het leren, maar laten we ook zien dat de gangbare praktijk van het beoordelen van leren aan de hand van meerkeuzevragen de variabiliteit van het leerproces niet volledig weergeeft.

Dit proefschrift draagt daarom bij aan het vakgebied Search-as-Learning door al onze datasets en code voor het simuleren en modelleren van leergedrag beschikbaar te stellen. We bieden ook verschillende inzichten in hoe leerlingen zich gedragen tijdens het zoeken en de impact van dit gedrag op hun leerresultaten. We hopen dat de bevindingen in dit proefschrift kunnen worden gebruikt als basis voor meer principiële verbeteringen in leergerichte zoeksystemen.

IR Information Retrieval

SAL Search-as-Learning

LLM Large Language Model

NLP Natural Language Processing

SERP Search Engine Results Page

SERP search engine results page

BERT Bidirectional Encoder Representations from Transformers

SEO search engine optimization

SEM search engine marketing

YAKE Yet Another Keyword Extractor

ASK Anomalous State of Knowledge

KT Knowledge Tracing

SEM structural equation modeling

IIR Interactive Information Retrieval

PLS partial least squares

PLS-SEM partial least squares structural equation modeling

CB-SEM covariance-based structural equation modeling

KALM Knowledge Acquisition Learner Model

ALG absolute learning gain

RPL realized potential learning

MLG maximum learning gain

VKS vocabulary knowledge scale

EQM Essay Quality Model

LV latent variable

MV measured variable

SERP search engine results page

QE Query Effort

QQ Query Quality

QX Query Exploration

DQ Document Quality

DQ_R Document Quality (reflective)

QQ_R Query Quality (reflective)

SX Session Exploration

WE Writing Effort

KG Knowledge Gain

l_i outer loading

w_i outer weight

AVE average variance extracted

ρ_a reliability coefficient

VIF variance inflation factor

CSM Complex Searcher Model

SACSM Subtopic-Aware Complex Searcher Model

USM User State Model

Curriculum Vitæ

Arthur Barbosa Câmara

Experience









Jun/2023—Now	Research Engineer at Zeta Alpha Vector	Amsterdam (NL)
Jul/2022—Oct/2022	Research Intern at Naver Labs Europe	Grenoble (FR)
Sep/2021—Jan/2022	AI Research Intern at Bloomberg LP	London (UK)
Feb/2018—Nov 2018	Lead Data Scientist at dti Digital	Belo Horizonte (BR)
May/2018—Feb/2018	Lead Data Scientist at Mediar Solutions	Belo Horizonte (BR)
Sep/2012—Sep/2015	Research Assistant at the LaTIn—UFMG	Belo Horizonte (BR)


Education


2016–2018	M.Sc. in Computer Science, UFMG	Belo Horizonte (BR)
2010–2015	B.Sc. in Computer Science, UFMG	Belo Horizonte (BR)

Publications

1. **Arthur Câmara**, *Craig Macdonald*. 2019. Dockerising Terrier for The Open-Source IR Replicability Challenge (OSIRRC 2019). In OSIRRC 2019, co-located with SIGIR 2019 [254].
2. *Sarah Ibrahimi, Shuo Chen, Arya Devanshu, Arthur Câmara, Yunlu Chen, Tanja Crijns, Maurits van der Goes, Thomas Mensink, Emiel van Miltenburg, Daan Odijk, William Thong, Jiaojiao Zhao, Pascal Mettes*. 2019. Interactive Exploration of Journalistic Video Footage through Multimodal Semantic Matching. In ACM Multimedia 2019 [255].
3. **Arthur Câmara**, *Claudia Hauff*. 2020. Diagnosing BERT with Retrieval Heuristics. In ECIR 2020 [256].
4. **Arthur Câmara**, *Nirmal Roy, David Maxwell, Claudia Hauff*. 2021. Searching to learn with instructional scaffolding. In CHIIR 2021 [62]. 🏆
5. *Nirmal Roy, Arthur Câmara, David Maxwell, Claudia Hauff*. 2021. Incorporating Widget Positioning in Interaction Models of Search Behaviour. In ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR) 2021 [111].
6. **Arthur Câmara**, *Claudia Hauff*. 2022. Moving Stuff Around: A study on the efficiency of moving documents into memory for Neural IR models. In RENeuIR at SIGIR 2022 [257].

-  7. **Arthur Câmara**, Dima El-Zein, da-Costa-Pereira, Célia. 2022. RULK: A Framework for Representing User Knowledge in Search-as-Learning. In DESIRES 2022 [60].
-  8. **Arthur Câmara**, David Maxwell, Claudia Hauff. 2022. Searching, learning, and subtopic ordering: A simulation-based analysis. In ECIR 2022 [90].
-  9. Gustavo Penha, **Arthur Câmara**, Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In ECIR 2022 [258]. 
-  10. Saher Esmeir, **Arthur Câmara**, Edgar Meij. 2022. Entity Retrieval from Multilingual Knowledge Graphs. In MRL at EMNLP 2022 [259].
-  11. Dima El-Zein, **Arthur Câmara**, Célia da-Costa-Pereira, Andrea Tettamanzi. 2023. RULKNE: Representing User Knowledge State in Search-as-Learning with Named Entities. In CHIIR 2023 [43].
-  12. **Arthur Câmara**, Claudia Hauff. 2023. Keep KALM and Search On: Unveiling Causal Connections in Search-as-Learning. Under Review.
-  13. Peide Zhu, **Arthur Câmara**, Nirmal Roy, David Maxwell, Claudia Hauff. 2024. On the Effects of Automatically Generated Adjunct Questions for Search as Learning. In CHIIR 2024 [260].

 Included in this thesis.

 Won a best paper award.

List of Figures

1.1	The process a traditional user of a search engine may take from translating an information need ① into a query and submitting it to a search engine ②, analyzing and clicking in a retrieved document ③, and finally being satisfied by the information found ④.	2
1.2	The process a <i>learner</i> may take from translating an information need ① into a query, submitting that query to the search system ②, clicking and reading a retrieved document ③, processing the newly acquired information ④ formulating new queries based on this new knowledge ⑤ and switching to another subtopic ⑥.	3
1.3	Tracking and predicting a learner’s knowledge state requires inferring hidden variables, such as their knowledge level before and after their search session (depicted as solid boxes), using proxy measurable variables, such as their scores in questionnaires and changes in their querying vocabulary (depicted as dashed boxes).	6
1.4	Measuring and predicting a learner’s behavior while searching requires inferring the value of latent variables related to their internal state (depicted as solid boxes) using proxy observable variables captured by the search system (depicted as dashed boxes).	8
2.1	The SearchX interface: the eight annotated interface components are described in Section 2.3. Note that the scaffolding component (displaying the Ethics topic) shows the FEEDBACK _{SC} scaffolding variant—complete with yellow progress gradients.	14
2.2	Hierarchical topic structure for the topic <i>Subprime Mortgage Crisis</i> . Topic and structure derived from TREC CAR 2017 [102]. Note that third-level subtopics (and deeper) and footnotes/references are excluded (illustrated in the figure by use of strikethroughs).	15
2.3	RPL examples: \triangle represents νks^{pre} and ∇ represents νks^{post} . Here, $n = 10$. Note that MLG is dependent only on νks^{pre} , while ALG is the difference between νks^{post} and νks^{pre} . RPL is defined by the ratio between ALG and MLG.	19
2.4	Overview of the flow of our user study. From seven topics, learners take a pre-test on two random topics and “sports” as a sanity check topic. The topic on which the learner scores the least is selected for their search session. A session lasts at least 30 minutes, using the SearchX platform with the Bing search API. Pages from Wikipedia and mirrors are filtered. At the end of the session, learners take a post-test and are asked to write a summary of what they learned.	20

2.5	RPL over the four different conditions.	22
2.6	Fraction of change in answers between the pre— and post—test VKS questionnaires.	22
2.7	For each row: (<i>top</i>) the fraction of query terms taken from topic outlines; (<i>middle</i>) the fraction of topic outline terms used for querying; and (<i>bottom</i>) the mean query length, over 5-minute blocks (x axes) of the 30-minute search session, considering: CONTROL (<i>left</i>); CURATED _{SC} (<i>center</i>); and FEEDBACK _{SC} (<i>right</i>). Here, we consider the first query instance as the start of the first interval.	23
3.1	The RULK framework and its main components. First, a clicked document d is transformed into \vec{v}_d by γ . Next, σ updates the learner’s current knowledge state \vec{c}_{ks} with \vec{v}_d . Finally, θ compares \vec{c}_{ks} to a vector \vec{t}_{ks} , generated from a target knowledge document, to get an estimation of a learner’s knowledge level (\tilde{G}).	32
3.2	Scatter plot between the self-reported knowledge of a learner, measured by their post-test VKS score and the knowledge level as estimated by RULK _{KW} , RULK _{LM} and RULK _{KW+LM}	38
3.3	t-SNE visualization of the evolution of the \vec{c}_{ks} vector for a learner learning about the topic “Noise-induced hearing loss” and the \vec{t}_{ks} (depicted as a red star ★) in both RULK _{KW} and RULK _{LM} vector spaces	39
3.4	Pearson’s correlations between estimated and measured (post-test scores) knowledge level, stratified by number of queries, clicks, and session duration.	39
4.1	The <i>Subtopic-Aware Complex Searcher Model</i> (SACSM). Changes from the CSM are highlighted in blue. Refer to Section 4.3 for more about the sequence and shapes.	47
4.2	Overview of the four variable parameters for instantiating simulated agents.	50
4.3	Accumulated percentage of the keywords seen for two different agents (averaged over all topics, weighted by the number of keywords) with varying ξ , λ and τ	52
4.4	Fraction of fully explored (i.e., agent reached ξ value) for two different agents (averaged over all topics, weighted by the number of keywords) with varying ξ , λ and τ	53
4.5	Fraction of keywords properly ‘ <i>learned</i> ’ by two different agents (averaged over all topics, weighted by the number of keywords) with varying ξ , λ and τ	54
5.1	A toy example of a traditional path model (left) and a SEM model (right). A SEM model consists of both a <i>measurement</i> and a <i>structural</i> model, while a path model only contains the structural part. In a SEM model, the measurement part describes how indicators (i.e., variables that are directly measured, here drawn as ellipses ●) are related to constructs (e.g., quality of a document, here drawn as squares ■). The structural model describes how these latent variables are causally related. The direction of an arrow $A \rightarrow B$ can be interpreted as “A causes B”.	62

5.2 Overview of the structural model of KALM. Arrows between latent variables denote a causal relationship between them. Boxes in cyan and azure indicate reflective and formative latent variables, respectively. 66

5.3 Path model with coefficients for KALM. Numbers after SC, LG and SW are the path values between two LVs in the Scaffolding, Lightning and SearchWell datasets, respectively. Values in green and red are positive and negative path coefficients. Superscripts *, ** and *** indicates p-values bellow 0.1, 0.01 and 0.001. 79

List of Tables

2.1	Overview of the ten concepts per topic 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.	18
2.2	Overview of the topics used in our study, with associated statistics. Two-way ANOVA tests revealed no significant differences in the average number of queries between topics ($F(6, 99) = 2.01, p = 0.07$) or between the average number of bookmarks ($F(6, 99) = 0.41, p = 0.87$).	19
2.3	Mean (\pm standard deviations) of RPL and search behavior metrics across all participants in each condition. [†] indicates two-way Anova significance, while \mathcal{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.	22
3.1	Statistics, per user, extracted from the dataset used in Chapter 2.	34
3.2	Top-10 keywords extracted by YAKE for each topic from their respective Wikipedia article. The article's keywords are sorted by frequency, and the stemming is manually reversed for clarity.	36
3.3	Pearson's correlation between estimated learning gains using a given implementation of RULK and reported learner's learning. Values in bold indicate the best correlation against a learning metric. All correlations are statistically significant, $p < 0.001$	38
4.1	Interaction costs grounding our agents, as derived from the data from Chapter 2.	49
4.2	# of subtopics and distinct keywords (KW) for each topic. We determine the ten KWs with higher TF-IDF for each subtopic on the respective subtopic section on Wikipedia. A KW may appear in the top ranks of several subtopics. KW difficulty is given by the age-of-acquisition, as proposed-Kuperman et al. [177].	51
4.3	Overview of (average) measures across agents and subtopic switching strategies, and real learners extracted from the FEEDBACK _{SC} cohort from Chapter 2.	52
5.1	Comparison between the three datasets analyzed in this study. Statistics are computed <i>after</i> filtering of participants with no queries, clicks, visits, or that interacted with documents not in English (for Scaffolding and SearchWell) or German (for Lightning).	68

5.2	Indicators, outer loadings (OL), reliability coefficient (ρ_a) and average variance extracted (AVE) for the reflective LVs for all datasets. Indicators for QX and QQ are averaged over all queries. Superscript * mean statistically significant results. Cells with dashes (—) are values that are not possible to compute on the given dataset. Indicators that are grayed-out are not considered in further analysis. Rows removed in some datasets but not others are marked with an indicator \blacklozenge	73
5.3	Indicators, outer weights (OW) and outer loadings (OL) for the indicators of the formatively measured LVs. Superscripts * indicates a value significantly different from 0. Indicators that are grayed-out are removed before further analysis. Rows removed in some datasets but not others are marked with an indicator \blacklozenge	75
5.4	Outer loadings OL , reliability coefficient (ρ_a) and average variance extracted (AVE) for reflectively measured version of the formative LVs for all datasets. Indicators are considered for all queries submitted by each learner. Superscript * mean statistically significant outer loadings.	76
5.5	Path coefficients, Total Effect and f^2 score for each latent variable (LV) into Knowledge Gain (KG). Values with a superscript* indicate statistical significance.	76
5.6	Summary of the indicators	77
5.7	Differences in R^2 when changing indicators for the KG LV. Values of AVE and ρ_a are still within the acceptable ranges defined in Section 5.5.1	81
5.8	Comparison between our main findings and existing literature. References marked with ✓ agree with our findings, while those marked with ✗ disagree.	83
5.9	Total effects between all LVs. Values with a superscript * indicate statistical significance.	84

SIKS Dissertation Series

Since 1998, all dissertations written by PhD. students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series.

- 2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
- 02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
- 03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
- 04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
- 05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
- 06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
- 07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
- 08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
- 09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
- 10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
- 11 Anne Schuth (UvA), Search Engines that Learn from Their Users
- 12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
- 13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
- 14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
- 15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
- 16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
- 17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
- 18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
- 19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
- 20 Daan Odijk (UvA), Context & Semantics in News & Web Search
- 21 Alejandro Moreno Celleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
- 22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
- 23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

- 24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
- 25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
- 26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
- 27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
- 28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
- 29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
- 30 Ruud Mattheij (TiU), The Eyes Have It
- 31 Mohammad Khelghati (UT), Deep web content monitoring
- 32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
- 33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
- 34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
- 35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
- 36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
- 37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
- 38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
- 39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
- 40 Christian Detweiler (TUD), Accounting for Values in Design
- 41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
- 42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
- 43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
- 44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
- 45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
- 46 Jorge Gallego Perez (UT), Robots to Make you Happy
- 47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
- 48 Tanja Buttler (TUD), Collecting Lessons Learned
- 49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

-
- 50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
-
- 2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
 02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
 03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
 04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
 05 Mahdieh Shadi (UvA), Collaboration Behavior
 06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
 07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
 08 Rob Konijn (VUA), Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
 09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
 10 Robby van Delden (UT), (Steering) Interactive Play Behavior
 11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
 12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
 13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
 14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
 15 Peter Berck (RUN), Memory-Based Text Correction
 16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
 17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
 18 Ridho Reinanda (UvA), Entity Associations for Search
 19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
 20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
 21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
 22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
 23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
 24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
 25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
 26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
 27 Michiel Jooisse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
 28 John Klein (VUA), Architecture Practices for Complex Contexts

-
- 29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT”
 - 30 Wilma Latuny (TiU), The Power of Facial Expressions
 - 31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
 - 32 Thaer Samar (RUN), Access to and Retrieval of Content in Web Archives
 - 33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
 - 34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
 - 35 Martine de Vos (VUA), Interpreting natural science spreadsheets
 - 36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
 - 37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
 - 38 Alex Kayal (TUD), Normative Social Applications
 - 39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
 - 40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
 - 41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
 - 42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
 - 43 Maaïke de Boer (RUN), Semantic Mapping in Video Retrieval
 - 44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
 - 45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
 - 46 Jan Schneider (OU), Sensor-based Learning Support
 - 47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
 - 48 Angel Suarez (OU), Collaborative inquiry-based learning
-
- 2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
 - 02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
 - 03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
 - 04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
 - 05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
 - 06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
 - 07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
 - 08 Rick Smetters (RUN), Advances in Model Learning for Software Systems
 - 09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

- 10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
 - 11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
 - 12 Xixi Lu (TU/e), Using behavioral context in process mining
 - 13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
 - 14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
 - 15 Naser Davarzani (UM), Biomarker discovery in heart failure
 - 16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
 - 17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
 - 18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
 - 19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
 - 20 Manxia Liu (RUN), Time and Bayesian Networks
 - 21 Aad Sloomaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
 - 22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
 - 23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
 - 24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
 - 25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
 - 26 Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
 - 27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
 - 28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
 - 29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
 - 30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
-
- 2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
 - 02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
 - 03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
 - 04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
 - 05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
 - 06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
 - 07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
 - 08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

- 09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
- 10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
- 11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
- 12 Jacqueline Heinerma (VUA), Better Together
- 13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
- 14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
- 15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
- 16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
- 17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
- 18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
- 19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
- 20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
- 21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
- 22 Martin van den Berg (VUA), Improving IT Decisions with Enterprise Architecture
- 23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
- 24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
- 25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
- 26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
- 27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
- 28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
- 29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
- 30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
- 31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
- 32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
- 33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
- 34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

-
- 35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
 - 36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
 - 37 Jian Fang (TUD), Database Acceleration on FPGAs
 - 38 Akos Kadar (OU), Learning visually grounded and multilingual representations
-
- 2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
 - 02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
 - 03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
 - 04 Maarten van Gompel (RUN), Context as Linguistic Bridges
 - 05 Yulong Pei (TU/e), On local and global structure mining
 - 06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
 - 07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
 - 08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search
 - 09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
 - 10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
 - 11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation Methods for Long-Tail Entity Recognition Models
 - 12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
 - 13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
 - 14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
 - 15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
 - 16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
 - 17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
 - 18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
 - 19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
 - 20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
 - 21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
 - 22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
 - 23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
 - 24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

- 25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
 - 26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
 - 27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
 - 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
 - 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
 - 30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
 - 31 Gongjin Lan (VUA), Learning better – From Baby to Better
 - 32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
 - 33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
 - 34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
 - 35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
-
- 2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
 - 02 Rijk Mercur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
 - 03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
 - 04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
 - 05 Davide Dell’Anna (UU), Data-Driven Supervision of Autonomous Systems
 - 06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
 - 07 Armel Lefebvre (UU), Research data management for open science
 - 08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
 - 09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children’s Collaboration Through Play
 - 10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
 - 11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
 - 12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
 - 13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
 - 14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
 - 15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
 - 16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
 - 17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

-
- 18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
 - 19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
 - 20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
 - 21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
 - 22 Sihang Qiu (TUD), Conversational Crowdsourcing
 - 23 Hugo Manuel Proença (UL), Robust rules for prediction and description
 - 24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
 - 25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
 - 26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
 - 27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
 - 28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
-
- 2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
 - 02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
 - 03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
 - 04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
 - 05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
 - 06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
 - 07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
 - 08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
 - 09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
 - 10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
 - 11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
 - 12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
 - 13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
 - 14 Michiel Overeem (UU), Evolution of Low-Code Platforms

- 15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
 - 16 Pieter Gijbbers (TU/e), Systems for AutoML Research
 - 17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
 - 18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
 - 19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
 - 20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
 - 21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
 - 22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
 - 23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
 - 24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
 - 25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
 - 26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
 - 27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
 - 28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
 - 29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
 - 30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
 - 31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
 - 32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
 - 33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
 - 34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
 - 35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
-
- 2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
 - 02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
 - 03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
 - 04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

-
- 05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
 - 06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
 - 07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
 - 08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
 - 09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
 - 10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
 - 11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
 - 12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
 - 13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
 - 14 Selma Čaušević (TUD), Energy resilience through self-organization
 - 15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
 - 16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
 - 17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
 - 18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
 - 19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
 - 20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
 - 21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
 - 22 Alireza Shojaiifar (UU), Volitional Cybersecurity
 - 23 Theo Theunissen (UU), Documentation in Continuous Software Development
 - 24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
 - 25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
 - 26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
 - 27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
 - 28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
 - 29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
-

- 2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
- 02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
- 03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
- 04 Mike Huisman (UL), Understanding Deep Meta-Learning
- 05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
- 06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
- 07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
- 08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
- 09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
- 10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
- 11 Mahmoud Shokrollahi-Far (TiU), Computational Reliability of Quranic Grammar
- 12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
- 13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
- 14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
- 15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health