



Delft University of Technology

Exploring the effects of interactive interfaces on user search behaviour

Roy, N.

DOI

[10.4233/uuid:9b70ead2-a41c-4657-9ecf-7cd668409fe4](https://doi.org/10.4233/uuid:9b70ead2-a41c-4657-9ecf-7cd668409fe4)

Publication date

2024

Document Version

Final published version

Citation (APA)

Roy, N. (2024). *Exploring the effects of interactive interfaces on user search behaviour*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:9b70ead2-a41c-4657-9ecf-7cd668409fe4>

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.

Exploring the Effects of Interactive Interfaces on User Search Behaviour

Nirmal Roy



EXPLORING THE EFFECTS OF INTERACTIVE INTERFACES ON USER SEARCH BEHAVIOUR

EXPLORING THE EFFECTS OF INTERACTIVE INTERFACES ON USER SEARCH BEHAVIOUR

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 Thursday, 27th of June 2024 at 15:00 o'clock.

by

Nirmal ROY

Master of Science in Computer Science,
Delft University of Technology, the Netherlands,
born in Kolkata, India.

This dissertation has been approved by the promotores.

Composition of the doctoral committee:

Rector Magnificus,
Prof. dr. ir. G.J.P.M Houben
Dr. C. Hauff

voorzitter
Delft University of Technology, promotor
Delft University of Technology, promotor

Independent members:

Prof. dr. A. van Deursen
Prof. dr. C. Eickhoff
Prof. dr. ir. D. Hiemstra
Dr. L. Azzopardi
Prof. dr. K.G. Langendoen

Delft University of Technology
University of Tübingen, Germany
Radboud University
University of Strathclyde, Scotland, UK
Delft University of Technology, reserve member

SIKS Dissertation Series No. 2024-24

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 project SearchX (639.022.722).



Published and distributed by: Nirmal Roy

Keywords: Information Retrieval, Qualitative studies, Natural Language Processing, Search Interfaces, User Modeling, User Behaviour, Search As Learning

Printed by: Gildeprint

Cover design by: Dr. Agathe Balayn and Dr. David Maxwell

Cover: Our search for meaning and knowledge in the universe; image of Kanchenjunga used with permission from [Green Valley Nepal Treks](#); Typeset in [Inter](#) by Rasmus Andersson, under the terms of the [Open Font Licence](#)

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

ISBN: 978-94-6496-158-4

Copyright © 2024 by N. Roy

An electronic version of this dissertation is available at <http://repository.tudelft.nl/>.

The only solutions that are ever worth anything are the solutions that people find themselves.

Satyajit Ray

CONTENTS

Summary	xi
Samenvatting	xiii
Acknowledgments	xv
1 Introduction	1
1.1 (Re)-examining User Interactions	4
1.2 Search As Learning	6
1.3 Modeling User Interaction	8
1.4 Relevance Judgement Collection	10
1.5 Thesis Origins	11
2 (Re)-examining User Interactions	13
2.1 Introduction	14
2.2 Related Work	15
2.2.1 Task Complexity and User Interactions	15
2.2.2 SERP Presentation and User Interactions	16
2.3 Methodology	17
2.3.1 Search Interface Design and System	17
2.3.2 Search Tasks	20
2.3.3 Query Selection and SERP Curation	20
2.3.4 Experimental Procedure	22
2.3.5 Study Participants	22
2.4 Results and discussion	23
2.4.1 RQ1: SERP Type and User Interactions	23
2.4.2 RQ2: Task Complexity and User Interactions	25
2.4.3 RQ3: Task Complexity, SERP Type and User Interactions	27
2.4.4 RQ4: Perceived Experience of SERPs	29
2.5 Conclusion	30
3 Search As Learning	33
3.1 Introduction	34
3.2 Related Work	36
3.3 Highlighting and Note-Taking	38
3.3.1 SearchX Interface	38
3.3.2 SearchX Logging	39
3.3.3 Text Highlighting	39
3.3.4 Note-Taking	40

3.4	User Study Design	40
3.4.1	Experimental Conditions	40
3.4.2	Procedure	41
3.4.3	Topics	42
3.4.4	SearchX Setup	43
3.4.5	Participants	44
3.4.6	Measuring Learning	45
3.5	Results	48
3.5.1	RQ1: Highlighting, Note-Taking and Learning	48
3.5.2	RQ2: Highlighting, Note-Taking and Search Behaviour of Users	50
3.5.3	RQ3: Active Reading Strategies and Learning	52
3.6	Conclusions	55
4	Modeling User Interactions and Widget Positioning	57
4.1	Introduction	58
4.2	Microeconomic Theory and IIR	60
4.3	Considering Widget Positioning	61
4.3.1	Positioning based on <i>Fitts' Law</i>	61
4.3.2	The Query History Widget (QHW)	62
4.4	Hypotheses	65
4.5	User Study Design	66
4.5.1	System, Corpus, Topic and Task	66
4.5.2	Interface and Incentives	66
4.5.3	Operationalising the QHW	68
4.5.4	Post-Task Survey	68
4.5.5	Crowdsourced Participants	68
4.6	Results	69
4.6.1	H1: Query Length	71
4.6.2	H2: Query Positioning in the QHW	71
4.6.3	H3: Distance of the QHW	72
4.6.4	H4: Slow Typing	73
4.6.5	H5: F-Shaped Gaze Pattern	73
4.7	Conclusions	74
5	Voice Modality and Relevance Judgment	77
5.1	Introduction	78
5.2	Related Work	79
5.2.1	Relevance Judgement Collection	79
5.2.2	Voice Modality	80
5.2.3	Cognitive Abilities	81
5.3	Methodology	81
5.3.1	Study Overview	82
5.3.2	Query/Passage Pairings	83
5.3.3	Cognitive Ability Tests	85
5.3.4	Assessor Interface	86
5.3.5	Outcome Measures	87

5.3.6	Participant Demographics	87
5.4	Results and Discussion	88
5.4.1	RQ1: Modality of Passage Presentation	89
5.4.2	RQ2: Passage Length	90
5.4.3	RQ3: Assessor Cognitive Abilities	93
5.5	Conclusions	95
6	Conclusions	97
6.1	(Re)-examining User Interactions	97
6.2	Search As Learning	99
6.3	Modeling User Interactions	100
6.4	Relevance Judgement Collection	102
6.5	Common next steps	103
6.6	Broader Research Directions	104
6.6.1	LLMs and User Interaction	105
6.6.2	Search As Learning	105
6.6.3	Relevance Judgement Collection	106
6.7	Final Remarks	107
	Bibliography	109
	Curriculum Vitæ	137
	SIKS Dissertation Series	139

SUMMARY

Interactive information retrieval (IIR) is a user-centered approach to information seeking and retrieval. In this paradigm, the search process is not confined to a single query and a static set of results. Instead, it emphasises the active involvement of users in refining their information needs, iteratively modifying queries, and exploring retrieved content. IIR studies research how to facilitate a more tailored and practical search experience, adapting to the evolving requirements and preferences of users. In this thesis, we focus on four distinct yet interrelated areas in the domain of IIR to have a better understanding of the interaction between the user and the information retrieval system.

How users interact with a search system depends on several things, including, but not limited to, the device on which they search, the interface, the task at hand, their prior expertise and so on. In Chapter 2, we explore the role of search interface layout and task complexity on user search behaviour and their task effectiveness. We aim to reproduce the setup of two IIR studies conducted a decade back that explored the effect of the search interface and task complexity on user behaviour. As search interfaces have kept on evolving, we ask the question of whether user search behaviour has remained the same. Our goal is to observe to what extent the findings from those two studies still hold today.

Next, we focus on a specific aspect of IIR, called Search as Learning (SAL), where users participate in learning-oriented search tasks. These search tasks are exploratory, involving multiple iterations that require cognitive processing and sensemaking. It often requires the researchers to spend time scanning, viewing, comparing and understanding documents. Prior studies have shown that, in offline classroom learning scenarios, active reading tools like highlighting and note-taking tools help learners better process what they read and consequently help their learning outcomes. In Chapter 3, we explore to what extent highlighting and note-taking tools, when we implement and incorporate them into the interface of a standard search engine, affect search behaviour and users' learning outcomes. We intend to explore if they are also beneficial in the online SAL scenario.

While designing and incorporating widgets (e.g. a note-taking tool) in a search interface, researchers face numerous design decisions regarding where to place the widgets, what they should look like, what functionalities they must have and so on. Due to budget constraints, it is not feasible to run A/B tests on all possible options. Thus, next in Chapter 4, we build a user model leveraging Search Economic Theory (SET), where we, for the first time, incorporate positional information of widgets. SET is based on micro-economic theory that assumes that users are rational agents—they aim to maximise profit and minimise cost. Previous work has utilised SET to develop models for predicting user interaction under various circumstances where widgets on the SERP are typically considered fixed, and their position is not part of the user model definition. Thus, in this thesis, we explore if we can derive a sensible hypothesis of user behaviour using our user model that incorporates positional information of widgets.

Finally, having so far dealt with documents in *text* modality of presentation, in Chapter 5 we look into the *voice* modality of presentation in the context of collecting relevance judgments for building test collections by employing crowdworkers. Previous studies have explored to what extent various factors like document length, topic difficulty, cognitive aspects of crowdworkers, etc., affect their relevance judgement effectiveness. However, none of them considered the presentation modality of the documents to be judged. Audio-only devices are getting popular, and leveraging these devices can increase the scope of collecting relevance judgements. For example, crowdworkers can judge document *on-the-go*, those with visual disabilities can also participate in the judgement task and so on. Thus, we observe how the presentation modality of documents, that is, representing them as text or voice, affects the relevance judgement effectiveness of crowdworkers. We also explore to what extent there is an interplay of document length and cognitive aspects of crowdworkers with the presentation modality.

With the studies conducted in this thesis, we make scientific contributions to the field by providing novel insights covering a breadth of topics and advancing our understanding of the field. We hope our contributions pave the way for further research and exploration in the field of IIR with the ultimate goal of enhancing the web search experience and performance of users.

SAMENVATTING

Interactive Information Retrieval (IIR) is een op de gebruiker gerichte benadering van het zoeken en verkrijgen van informatie. In dit paradigma is het zoekproces niet beperkt tot een enkele zoekopdracht en een statische set resultaten. In plaats daarvan benadrukt het de actieve betrokkenheid van gebruikers bij het verfijnen van hun informatiebehoeften, iteratief aanpassen van zoekopdrachten en verkennen van opgehaalde inhoud. IIR-onderzoeken richten zich op hoe een meer op maat gemaakte en praktische zoekervaring kan worden gefaciliteerd, aangepast aan de evoluerende eisen en voorkeuren van gebruikers. In deze scriptie richten we ons op vier afzonderlijke maar onderling verbonden gebieden binnen het domein van IIR om een beter begrip te krijgen van de interactie tussen de gebruiker en het informatieretrievalsysteem.

Hoe gebruikers omgaan met een zoeksysteem hangt af van verschillende factoren, waaronder, maar niet beperkt tot, het apparaat waarop ze zoeken, de interface, de taak die ze uitvoeren, hun eerdere expertise enzovoort. In Hoofdstuk 2, onderzoeken we de rol van de lay-out van de zoekinterface en de complexiteit van de taak op het zoekgedrag van de gebruiker en hun taakeffectiviteit. We streven ernaar de opstelling van twee IIR-onderzoeken die tien jaar geleden zijn uitgevoerd en die het effect van de zoekinterface en de complexiteit van de taak op het gebruikersgedrag hebben onderzocht, te reproduceren. Omdat zoekinterfaces blijven evolueren, stellen we de vraag of het zoekgedrag van gebruikers nog steeds hetzelfde is. Ons doel is om te observeren in hoeverre de bevindingen van die twee studies vandaag de dag nog steeds geldig zijn.

Vervolgens richten we ons op een specifiek aspect van IIR, genaamd 'Search As Learning' (SAL), waar gebruikers deelnemen aan op leren gerichte zoektaken. Deze zoektaken zijn verkennend, omvatten meerdere iteraties die cognitieve verwerking en betekenisgeving vereisen. Het vereist vaak van de zoekers dat ze tijd besteden aan scannen, bekijken, vergelijken en begrijpen van documenten. Eerdere studies hebben aangetoond dat actieve leesinstrumenten zoals markeren en notities maken leerlingen in offline klaslokaalscenario's helpen om beter te verwerken wat ze lezen en bijgevolg hun leerresultaten verbeteren. In Hoofdstuk 3, onderzoeken we in hoeverre markeren en notities maken, wanneer we ze implementeren en opnemen in de interface van een standaard zoekmachine, van invloed zijn op het zoekgedrag van gebruikers en de leerresultaten van gebruikers. We willen verkennen of ze ook gunstig zijn in de online SAL-scenario's.

Bij het ontwerpen en opnemen van widgets (bijv. een notitietafel) in een zoekinterface staan onderzoekers voor tal van ontwerpbeslissingen over waar ze de widgets moeten plaatsen, hoe ze eruit moeten zien, welke functionaliteiten ze moeten hebben, enzovoort. Vanwege budgetbeperkingen is het niet haalbaar om A/B-tests uit te voeren voor alle mogelijke opties. Daarom bouwen we in Hoofdstuk 4, een gebruikersmodel op basis van de Search Economic Theory (SET), waarin we voor het eerst positionele informatie van widgets opnemen. SET is gebaseerd op micro-economische theorie die ervan uitgaat dat gebruikers rationele agenten zijn - ze streven naar het maximaliseren van winst en het

minimaliseren van kosten. Eerdere werk heeft SET gebruikt om modellen te ontwikkelen voor het voorspellen van gebruikersinteractie onder verschillende omstandigheden waarbij widgets op de SERP doorgaans als vast worden beschouwd en hun positie geen deel uitmaakt van de definitie van het gebruikersmodel. In deze scriptie onderzoeken we dus of we een zinvolle hypothese van gebruikersgedrag kunnen afleiden met ons gebruikersmodel dat positionele informatie van widgets opneemt.

Tot slot, nadat we tot nu toe met documenten in de *text* modaliteit van presentatie hebben gewerkt, kijken we in Hoofdstuk 5, naar de *voice* modaliteit van presentatie in de context van het verzamelen van relevantieoordelen voor het bouwen van testverzamelingen door gebruik te maken van crowdworkers. Eerdere studies hebben onderzocht in hoeverre verschillende factoren zoals documentlengte, onderwerpsmoeilijkheid, cognitieve aspecten van crowdworkers, enzovoort, van invloed zijn op hun effectiviteit in het beoordelen van relevantie. Echter, geen van hen overwoog de presentatiemodaliteit van de te beoordelen documenten. Audio-only apparaten worden steeds populairder, en het benutten van deze apparaten kan de mogelijkheden voor het verzamelen van relevantieoordelen vergroten. Bijvoorbeeld, crowdworkers kunnen documenten beoordelen terwijl ze onderweg zijn, mensen met visuele beperkingen kunnen ook deelnemen aan de beoordelingstaak, enzovoort. Daarom observeren we hoe de presentatiemodaliteit van documenten, dat wil zeggen, ze voorstellen als tekst of stem, van invloed is op de effectiviteit van crowdworkers bij het beoordelen van relevantie. We onderzoeken ook in hoeverre er een wisselwerking is tussen de lengte van het document en de cognitieve aspecten van crowdworkers met de presentatiemodaliteit.

Met de studies die in deze scriptie zijn uitgevoerd, leveren we wetenschappelijke bijdragen aan het vakgebied door nieuwe inzichten te bieden die een breed scala van onderwerpen beslaan en ons begrip van het vakgebied bevorderen. We hopen dat onze bijdragen de weg effenen voor verder onderzoek en verkenning op het gebied van IIR, met als uiteindelijk doel de webzoekervaring en prestaties.

ACKNOWLEDGMENTS

“The goal of the PhD program is to train you to be an *independent* researcher”—while this sentence often soars as a permanent banner at the beginning of such PhD journeys, those who have completed it (or are close to doing so) know the importance of the *dependencies* that shape the PhD journey. Some might even argue that the goal of the PhD program is actually to train you to find the support systems vital to you (and your research). The dependencies I found in the last four and a half years, most of them in sweet serendipity, made my PhD journey one I will look back on with fondness. While words are not enough to encapsulate my gratitude, they are a decent starting point. This is a *letter* to everyone who formed my support system and helped me undertake and complete this Goliath of a task.

I particularly remember the Slack message Claudia sent confirming that I would indeed start a PhD position with her. I recall a feeling of relief that I would not be jobless anymore (every Master’s student’s dream). But most importantly, I remember feeling proud and excited to join her research group and work with her for four more years. Having already completed my Master’s thesis, I was aware of the learning potential this opportunity provided me. At the end of this journey, I am not surprised that it turned out to be exactly what I had imagined. Thank you, Claudia, for all the Overleaf feedback (I hope if we plot the amount of feedback vs. the number of years, we see a decreasing trend), for the guidance on thinking about research and questions, but most importantly, for providing me the freedom and support to pursue activities beyond my PhD. I come out of this journey as a better writer, a better presenter, and definitely a better researcher. It has been an utmost pleasure working with you!

An important aspect of doing research in academia as a PhD student, which is often overlooked, is the people in your office. You spend a lot of time together whiteboarding, drinking coffee, and venting when the *p*-value turns out to be greater than 0.05. It’s a bonus if they are three Brazilians and one German since now, if you include me, in one room we have perspectives, opinions, and biases from three different continents. I consider myself enormously lucky to have this unique concoction of Cool Promovendi—Felipe, Gustavo, Arthur, and Tim—as my colleagues-turned-friends. I wish we spent more time working in the office together (but I also know we enjoyed working remotely more). Nonetheless, I am glad that we *did* have fun with barbecues, trips, and beaches outside of it! A pandemic PhD is unimaginable without you guys.

To the entire WIS group, thank you for pondering with me the meaning of it all, even when things felt difficult for reasons within and beyond our control. The pizzas, lemon chickens, and beer felt better because we persevered no matter what. So, to Avishek, Ujwal, Jie, Jurek, Alisa, Andra, Gaoile, Garrett, Petros, Peide, Kyriakos, Lijun, Lorenzo, Manuel, Sara, Sepideh, Shahin, Shabnam, and Ziyu, you have helped me navigate this journey in a myriad of ways. Thanks a lot to all of you for that. It would be remiss not to mention (and thank) George for the trips to Maria’s and Christos for a lot of shared pain (driving license,

being rich to poor just by changing the country of residence, bureaucracy, etc.). And thank you, Daphne, for your infinite patience and encouragement, and for making all of our lives easier when we almost drowned in the said bureaucracy required for the internships or even for us to graduate.

Thank you, Kevin, Markus, Scott, Can, Leonardo, Rexhina, and everyone at Amazon who gave me a chance to work on very interesting and cool stuff. I learned a lot in my combined 10 months of internship, and that would not have been possible without your help and collaboration. Life in the USA might not have been as exciting if not for many people—Deeptish, Suchita, Monish, Sourav, Naveen, Sreejita, Jahin, Simran, Arnab, Swastik, Bahnny, Ria, Suvrajit, Souvik Kundu, Debsomit, Shwetasree, and everyone else—thanks for making my life far away from home feel a lot like home. And Prasid, although I wish that some of the miles we hiked and drove were in Europe too (where we, of course, count them in the metric system), it was almost relieving to know that our conversations were still some of the closest things to my heart, even though we were having them more than 180 longitudinal degrees away from where we used to have them. USA, the internship, and all the memories that came with it would not have happened without you there.

I would like to express my utmost gratitude to my committee: Arie van Deursen, Djoerd Hiemstra, Carsten Eickhoff, Leif Azzopardi and Koen Langendoen. Thanks for taking the time to read and comment on my thesis and agreeing to be part of the committee. I have no doubt that it will be a very enjoyable discussion on June 27, 2024. A special thanks to Geert-Jan for always emphasizing the importance of the ‘P’ in PhD. I learned how to dig myself out of the proverbial muck and broaden my perspectives whenever I went too deep into one particular train of thought. I will always remember all our discussions with a lot of reverence.

I have a lot of respect for you too, David and Agathe, for everything you have achieved and for all the skills you possess as researchers and beyond. But most of all, I consider myself extremely lucky to have you as my friends (and paranympths, coauthors, cover designers, etc.). David, none of my papers or my thesis would look the way they do now without your time and effort behind the work. You have been extremely patient and always the helpful shoulder whenever I needed it. Agathe, thanks for always being the reliable friend in need. I barely remember you saying no to any requests (other than the one asking you to come climbing with us). Thank you for being amazing partners in my PhD journey.

To my ever-expanding *Indian* family in the greater Delft area: perhaps you’re not aware of the impact you have had on my overall mental stability over the course of the last seven years. Or perhaps, more realistically, you have contributed much more to my insanity. Nonetheless, without your constant routine presence in my life outside research, my often unstructured research life would have been a mess. And without the happy chaos and laughter that ensue whenever we are together, my mind wouldn’t have been calm when I had to work on my papers. Summi, Batheja, Tavishi, Kanav, Sanjeev, Pulkit, Panda, Shivli, Raju, Sid, Arka, Sasha, Rahul, Apoorva, Gouri, and Ayesha, you probably know that I cannot detail every little highlight that we had in the last few years as it will probably be longer than my thesis, but I hold all those memories very close to my heart. To my more-or-less constant family back in Kolkata, Ma, Baba—I know things have often been difficult in this “long-distance” relationship; I know I have been too preoccupied to pay the required attention some matters demanded, but I hope to have made you smile and proud at

least a few times in the past years and plan to do so more in the future. A reiteration of the importance of family and parents in any endeavor of life is futile, and my acknowledgment section certainly will not provide any ground-breaking information on that, but you should nonetheless know that without your support (and sometimes your *great expectations*), I would not have been motivated to put my head down and work hard when it mattered the most.

The heart seeks out the familiarity and comfort of home as life goes through its usual crests and troughs. And because of you, Meghdipa, I never had to look very far for those feelings. You have always been the grounding influence I badly needed, a reminder not to forget about the small things whenever I had the tendency to frantically chase an elusive end-game. The crests wouldn't make a lot of sense without you and the troughs would feel extremely daunting if you were not there. So there's no better person I can think of to share this moment, a goal I chased for a while, than you and the orange fur ball of happiness you have introduced into my life. Thank you for everything!

*Nirmal Roy
Delft, 2024*

1

INTRODUCTION

Searching for information on computers is an everyday thing. When users have an *information need*, they turn to contemporary commercial **Information Retrieval** (IR) systems such as Google and Bing. The users expect these IR systems to return results relevant to their information needs. Typically, these results are ranked by decreasing order of relevance. When a user searches for information using an IR system, several *interactions* occur between the user and the IR system, where the user's goal is to satisfy their information need. The study of **Interactive Information Retrieval** (IIR), a sub-field of the broader study of IR, is primarily devoted to considering the interactions between the searcher and retrieval system [44]. In this thesis, we aim to deepen our understanding of the interactions between the user and retrieval system by conducting four studies in the space of IIR.

The need to access information effectively has been present throughout human history. Prior to the age of computers, libraries typically housed extensive collections of books and papers. In addition to using these catalogues, an individual seeking out information could also *interact* with trained librarians, a process commonly known as reference interview [48]. The librarians who maintained such collections had a good overview of the inventory and were trained to assist the searchers in expressing their needs and help them find relevant information. The development of the **Internet** and **World Wide Web** (WWW) [38] enabled a massive surge of information to be present *online* and accessible for users looking to satisfy their information needs. How users explore and access this knowledge also differs from how they access information *offline* (e.g., in libraries).

The development of IR systems like commercial web search engines enables information seekers to search the ever-growing space of information available via the internet with minimal effort. Given a **query** that represents the information need of a user, an IR system searches through a collection of unstructured or semi-structured data (such as a collection of web pages or other text documents, or even images or videos, representing multimedia retrieval) before returning potential matches to the searcher. The *matching process* (of documents in a collection and the user query) can be performed using different methodologies. The broader field of IR primarily deals with evaluating *system-sided* aspects

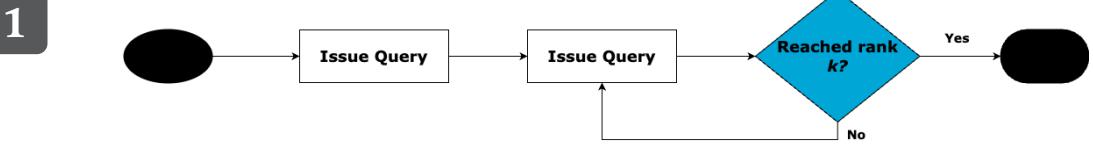


Figure 1.1: Model of how users interact according to the Cranfield paradigm

concerning the methodologies and quality of returned rankings, how efficient the retrieval engine is, etc.

At the core of much of the research conducted in the field of IR lies the **Cranfield paradigm**, a term that signifies a standardized approach to evaluating IR systems [65]. This paradigm is primarily attributed to Cyril Cleverdon of Cranfield University. It revolves around the concept of standardized test collections – structured sets of documents that can be utilized by various researchers, thereby establishing a consistent foundation for experimenting with IR methodologies. Through the utilization of the Cranfield paradigm, we have achieved notable strides in the realm of evaluating IR systems. Nonetheless, there is an argument that this approach may possess certain limitations when viewed from the perspective of IIR as it oversimplifies the intricate interactions occurring between a searcher and a retrieval system [43, 120]. The experimental frameworks that have emerged from the Cranfield paradigm are built upon a set of assumptions of simple interactions between the user and the IR systems that often diverge significantly from the actual dynamics of how they engage with such a system (Figure 1.1). These assumptions can be summarized as follows:

- The searcher will *submit* a single query throughout a search session.
- They will examine documents upto a *predetermined depth* (typically set at around 100 in TREC experiments).
- They will assess *all* documents to that fixed depth.

In other words, the paradigm broadly needs to consider the complexities that arise from the user-sided aspects core to IIR. Inspired by an event in their daily lives (perhaps by observation, reading a book, or through conversation with another human), a searcher will have an information need. This information need can arise from a knowledge gap in the searcher's mind, an internal inconsistency in what they are experiencing, or a conflict of evidence. The searcher will then begin the IIR process to satisfy their (perhaps vague) information need. Upon bringing up the interface of a retrieval system, the searcher starts their so-called search session, which begins with formulating the information need as a **query**. Once the query has been submitted, a series of interactions occur between the system and the searcher [120] as depicted in Figure 1.2. **Results** will be retrieved by the underlying retrieval system and presented to the searcher in the form of a **Search Engine Results Page** (SERP). Depending on the features available on the corresponding SERP, the searcher may decide to examine **snippets** click and read returned results, explore images, videos or other **verticals**, scan **direct answers** or **entity cards**, etc. These interactions, occurring on the SERP, are essential to those studying IIR.

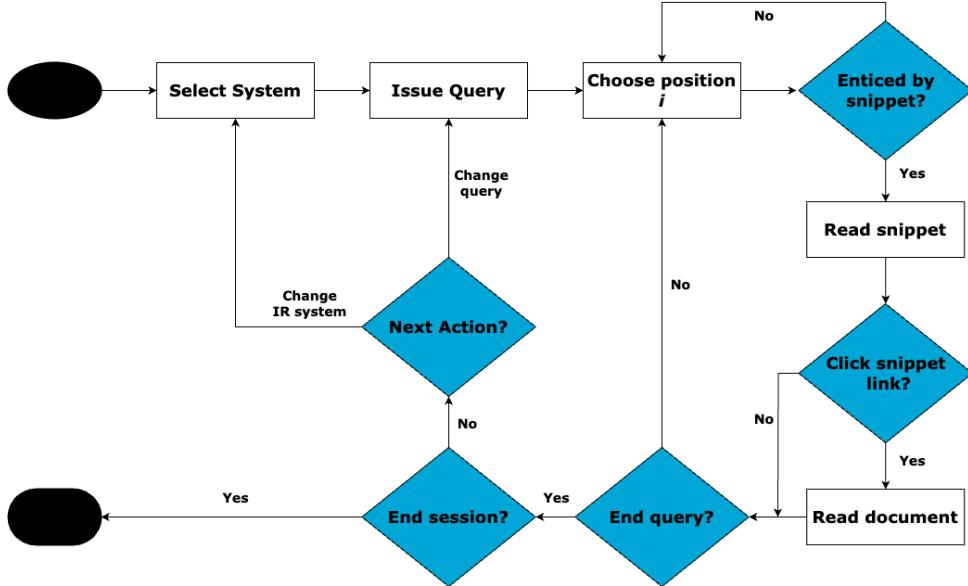


Figure 1.2: A model of a simplified version of the IIR process inspired from the works of [176, 272]. Depending on the specific search system, available search widgets and the task at hand there will be other decision points or actions taken by the user.

Searchers could also issue multiple queries during a search session. Subsequently, they would adapt their interactions based on the perceived quality of presented ranked result lists [188]. The interactions will also depend on the complexity of the task at hand—searchers would expect the first result to be relevant for *navigational* tasks (e.g., finding the homepage of Burton).¹ In contrast, they would be willing to spend more time and effort searching for more complex tasks (e.g., comparing different climbing shoes to buy). Moreover, in the ever-changing landscape of web search, not only how the results being presented have gone through multiple transformations (from ‘ten blue links’ to Bing Chat), but numerous widgets and functionalities tailored to specific search goals (e.g., answer cards, query suggestions, entity cards etc.) have been added (and sometimes removed) from the web search interfaces. The presentation of results, including their order, formatting, and visual elements, plays a crucial role in determining how searchers engage with the information presented [13, 15, 45, 81, 165, 194, 240, 304]. For example, top results and those with rich features like featured snippets or image carousels tend to attract more attention due to their prominence [219].

The study of IIR attempts to understand searcher’s interactions with an IR system and incorporate new findings into the development of retrieval systems [51]. IIR studies can include aspects from both *user-sided* and retrieval *system-sided* research. For example, one might present the results of a user study examining a particular phenomenon of a searcher’s behaviour and provide details of a system-sided evaluation. As discussed by Kelly et al.

¹<https://www.burton.com/>

[135], IIR can trace its roots back to a variety of different disciplines, including traditional IR (i.e. exclusively system-sided research); library and information sciences; psychology; and **Human-Computer Interaction** (HCI). Typically presented as a branch of IR and HCI, arguments also exist to consider IIR as a distinct area of research [237].

Work to improve our understanding of the user-IR system interaction has been undertaken in various aspects of the study of IIR, including (but not limited to) interaction in the presence (or absence) of various SERP components [13, 15, 165, 240, 249], or while undertaking search goals of varying levels of complexity [16, 127, 233], modelling of user interactions [19, 33, 181], etc. These prior researchers have aided in uncovering critical insights related to user interaction during their information-seeking process. For example, researchers have been observed to follow an F-shaped pattern while navigating a SERP [81]. Behavioural metrics like document reading time or number of clicks on SERP have been shown to correlate with the amount of knowledge searchers gain during a search session tailored towards learning [68, 84].

In this thesis, we aim to deepen our understanding of four such aspects of IIR—*(i)* effect of SERP layout and task complexity on user interactions; *(ii)* influence of learning tools on user interaction and learning outcomes during a learning-oriented search process; *(iii)* modelling user interaction; and *(iv)* effect of document modality on collection of relevance judgements. In the following sections, we describe in detail each of these aspects, together with our main research questions and key findings.

1.1. (RE)-EXAMINING USER INTERACTIONS

Prior work in IIR have shown how (and where) content is displayed in a SERP [16, 249, 265] affects user interaction. The incorporation of heterogeneous content like images, videos etc. [13, 15, 45, 81, 165, 194, 240, 304] also affects user interaction. In addition, past research [16, 256, 265, 304] has shown that user behaviour on the SERP does not only depend on the *presentation* of information, but also on the *search task* at hand. For a navigational task such as '*finding the homepage of Burton*', a user—in the ideal case—requires a single query and a single click. Contrast this to an informational task, such as '*good and affordable ski-resorts in Europe*'. Such tasks require the scanning of multiple results and likely result in further query reformulations to learn more about specific suggestions. While contemporary web SERPs maintain the original idea of a *list* of items that are ranked in decreasing order of relevance, alternative presentations such as a *grid* layout—as also recently (again) popularised by *You.com*—have also been explored [128, 193, 256, 305]. Moreover, as the layout of SERPs of commercial web search engines has evolved, users have become accustomed to different types of SERPs (Figure 1.3). From an IIR reproducibility perspective, this begs the question—*do users exhibit similar web search behaviour today compared to 10 years ago?*

Despite numerous research in IIR, there needs to be more effort to reproduce findings from past research. IIR research often involves a combination of qualitative and quantitative methods, and the complexity of studying user interaction can sometimes make it challenging to ensure full reproducibility. Factors such as variations in study participants, differences in experimental setups (e.g., the search system deployed, interfaces of the search system, retrieval algorithms, corpus, etc.), and ethical concerns regarding the release of experimental logs contribute to difficulties in replicating IIR studies. To this extent, in this

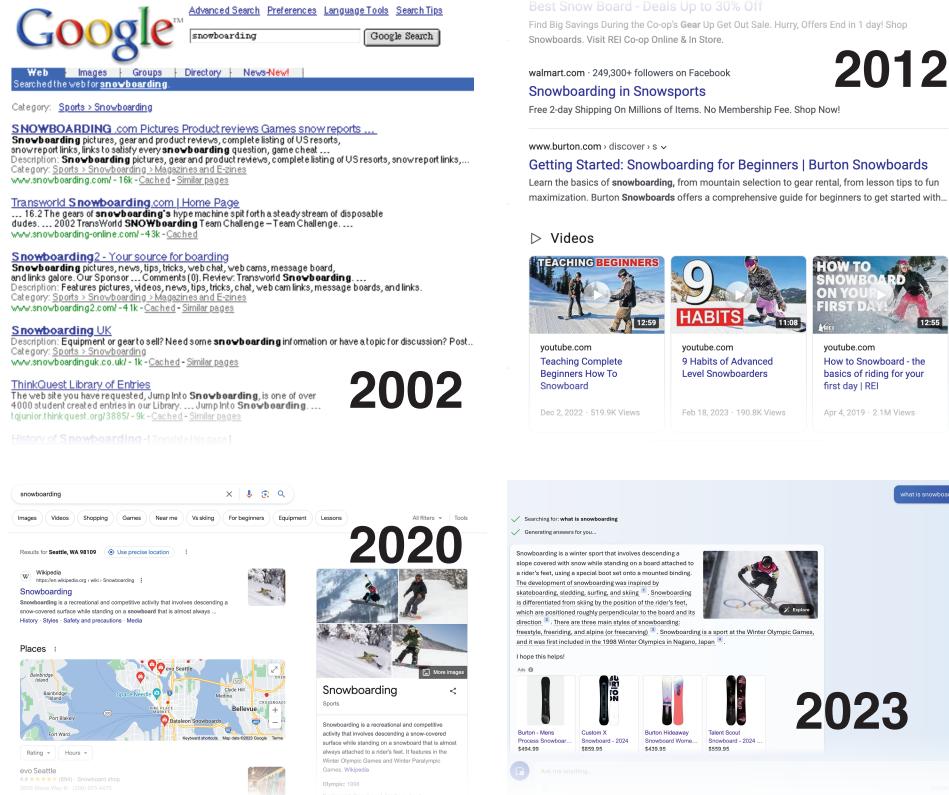


Figure 1.3: Evolution of SERP over time (for the query [snowboarding](#)). **Top-left:** In early 2002, search engines presented results in the *ten blue links* format. **Top-right:** Earlier in the last decade, results from different verticals like images, videos, news etc. started to be interleaved with the web results. The SERP renderings of 2002 and 2012 are obtained from www.theoldweb.com. **Bottom-left:** Widgets like entity cards, direct answers, maps etc., were introduced towards the second half of the last decade. **Bottom-right:** An example from Bing-Chat, LLM-powered conversational agents are considered to be the future of web search. 2020 and 2023 SERPs are screenshots taken from Google and Bing chat in November 2023.

thesis, we focus our attention on reproducing the experimental setup of two prior studies: Arguello et al. [16] (published in 2012) as well as Siu and Chaparro [256] (published in 2014)—these both investigated how user interactions on the SERP are influenced by the presence or absence of heterogeneous content, the layout of the SERP (list vs. grid), and task complexity.

The process of reproducing prior user studies will further enable us to identify potential challenges in reproducibility in an IIR context. Our goal is to pinpoint important factors that future researchers should be aware of while setting up their IIR studies to ensure better reproducibility of their findings. Reproducibility in science ensures that the findings, methods, and conclusions of scientific studies can be independently verified and validated by other researchers. When research results can be replicated by different individuals or groups using the same methods and data, it strengthens the reliability and robustness of scientific knowledge. In the other areas of computer science, there has been a growing recognition of the significance of reproducibility, leading to various initiatives aimed at replicating and validating research outcomes. Efforts like the ‘ACM Artifact Review and Badging’ initiative encourage authors to provide the necessary materials for others to reproduce their results, fostering transparency and accountability. Platforms like GitHub have enabled researchers to openly share their code and data, facilitating the validation of computational findings. Additionally, conferences like SIGIR, NeurIPS, ICML, and others have introduced reproducibility tracks, where researchers can submit papers focusing on reproducing and validating previously published work.

Inspired by the two papers we reproduced, the following *broad* research questions (**B-RQ**) guide our chapter 2.

B-RQ1: How do the layout of the SERP and the complexity of the task at hand affect user interaction? To what extent has user interaction with web search engines changed in the last ten years?

Findings and Contribution: To answer **B-RQ1**, in Chapter 2, we conduct a user study with 41 participants where the layout of search results on the SERP and task complexity are the primary dependent variables. We experiment with four different SERP layouts and tasks belonging to 4 different levels of task complexity. Specifically, we explore whether eight observations from [16, 256] about users and their interactions with list vs. grid layouts across different task complexities hold today. We find that both results layouts on SERP and task complexity significantly affect various aspects of user interaction with web search results. Regarding reproducibility, we find evidence to confirm two, with partial evidence for four further hypotheses.

1.2. SEARCH AS LEARNING

The seminal paper of Marchionini [174] defines, what is known as **Search As Learning** (SAL) [67] as search activities whose ultimate goal is human learning. Over the last decades, people have turned to web search engines not only to access information quickly but also to learn, discover and ingest information on topics of interest and gain knowledge on those topics [42, 197, 229]. Searching for information to learn about how to get better at snowboarding is an inherently different process than searching for the homepage of Burton.

The former is a more complex process dependent on the searcher's prior knowledge of the topic. In this example, information relevant to a beginner snowboarding enthusiast might not be relevant to someone more advanced. For the latter scenario, we can safely assume that most searchers are looking for one specific website irrespective of the context from where their information needs originated. Secondly, the former is a more complex and iterative process where the searcher must read multiple documents, issue multiple queries and spend more cognitive resources to process the information and gain knowledge on the topic. In contrast, the latter is a more straightforward process where the searchers expect to find the homepage of Burton as the first result returned by their search engine. However, modern web search engines are tailored more towards the latter kind of search goal (quick access to information) as compared to the former—there is not much support built to help searchers scan, compare, evaluate and analyse the information they find [18, 99].

Previous research within the SAL domain has focused on: (i) understanding user behaviours when undertaking a learning-oriented search task; [47, 84, 127, 160, 163, 189]; (ii) exploring different types of users and their behaviours (e.g., novices vs experts) [40, 84, 202, 233]; and (iii) the optimisation of retrieval systems for learning [267–269]. These studies have shown that search behaviour and user characteristics affect learning outcomes during the SAL process. In terms of scalable behavioural metrics as proxies for measuring knowledge gained across a search session, the document dwell time is a good indicator for learning [68, 84] as well as the number of SERP clicks [68] and the number of unique domains present among the top-ranked search results [84]. In terms of user characteristics, there have been contrasting findings on how their prior knowledge on a topic affects their learning outcomes—while some found users with low prior knowledge achieved higher learning gains than learners with at least some knowledge a priori [95], others did not find a difference in learning outcomes between the two cohorts [162, 202].

Outside of the web search scenario, active reading strategies such as annotating content (highlighting, note-taking, etc.) have been shown to have multiple benefits when engaging in long and complex learning task [175, 214, 297, 314]. These tools enable learners to limit their working memory load, as well as articulate and reformulate their thoughts. In turn, this can lead to substantial improvements in the understanding and retention of knowledge [140, 175]. Despite the apparent benefits of active reading tools within a learning context, highlighting and note-taking tools are not found in contemporary web search engines. Efforts have, however, been made to develop information organisational tools (e.g., a note-taking interface allowing users to keep track of their search context) [39, 80]. However, none of these studies explicitly measured the effect these tools had on learning. Therefore, in this thesis, we explore the impact of users' searching behaviour and their learning process if, during a learning-oriented search process, they had access to highlighting and note-taking affordances—tools that have been shown to aid in traditional classroom learning. This research gap motivates our second research question:

B-RQ2: How do active reading tools affect user interactions and learning outcomes during a learning-oriented search (SAL) process?

Findings and Contribution: To answer B-RQ2, in Chapter 3, we conduct another user study with 115 participants observing the effect of two active reading tools (highlighting and note-taking) on their learning outcomes and search behaviour. We measure their learning

over two tasks: a recall-oriented vocabulary learning task [189, 233]; and a cognitively demanding essay writing task [162, 260]. We observe that neither the highlighting nor the note-taking tool helped participants in the vocabulary learning tasks. However, access to only the highlighting or note-taking tool allows them to write better essays than participants without the tools. We also observe that access to the highlighting tool leads participants to submit fewer queries and spend more time examining documents. On the other hand, note-taking leads to participants spending less time reading documents and taking more notes. We also explore how different highlighting and note-taking strategies help with their learning outcomes by investigating whether five hypotheses, inspired by the education literature, hold up in our SAL setup, too. We confirmed three of those hypotheses and showed that while engaging in complex learning-oriented search tasks on the web, merely highlighting and note-taking may not benefit learners. Instead, how these tools change the way the learners scan and process text is more important for learning while searching.

1.3. MODELING USER INTERACTION

One challenge of conducting user studies is that the experiment conditions must be determined. An IIR practitioner or a designer of a retrieval system/SERP has to make numerous design decisions regarding positioning a particular widget (e.g., a note-taking tool) on the SERP. In theory, there are innumerable design choices, as the widget in question can be positioned anywhere on the SERP. The number of design choices keeps growing exponentially as we include more than one design feature of the widget (e.g., location and size of the note-taking widget, offered functionalities like text formatting, etc.). As it is not possible to test all possible experimental conditions, they are usually decided based on *best guesses*. In the previous study, we positioned the note-taking widget on the right side of the SERP as that area is typically empty. However, one can argue about positioning it at a different position. One way of overcoming this limitation is by the use of simulation.

Simulation is defined as the imitation of the operation of a real-world process or system over time [32]. Such an approach allows one to gain insight into the functioning of some real-world phenomenon, such as the interactions that take place during the IIR process. Running simulations by modelling user interactions can help us rapidly explore different scenarios (e.g., where to position the note-taking widget on the SERP), all at a low cost and without needing to consider issues such as subject fatigue (within a user study, for example). Ultimately, the goal is to only run user studies or A/B tests on interface designs that have shown promise from prior simulations. Many models of user interaction in the context of IIR have been defined in the past and can generally be categorised into two groups: **descriptive** models [34, 36, 85, 120, 143, 302] and **formal** (mathematical) models [19, 92, 217]. The former provides us with intuitions and a holistic view of a user's search behaviour (e.g., with the **Berrypicking model** [34], users *pick* through information patches—analogous to people collecting berries). While they provide us with explanations of why searchers behave in a particular manner, they do not allow us to *predict* how a user's search behaviour will change in response to changes to the SERP, the quality of the results, etc. For this step, formal models such as **Search Economic Theory (SET)** [19, 20], *Information Foraging Theory (IFT)* [216] or the **Interactive Probability Ranking Principle (iPRP)** [92] are required. Of particular interest in this thesis is the SET proposed by Azzopardi [19].

SET is a theory explaining the search process in terms of economics – in particular, microeconomic theory. Microeconomic theory assumes that individuals aim to optimise their profits within the confines of budgetary or other limitations [288]. This framework can also be an intuitive method for modelling interactions between humans and computers. When presented with a demand (which might arise due to factors like contextual elements, the core task, or the utilised system), individuals will invest *effort* in engaging with the system, utilising internal resources like cognitive capacity, attention, and energy. Moreover, users will face a *cost* incurred by expending external resources such as time, money, or physical exertion (such as manipulating a mouse or typing on a keyboard) [184]. In the sphere of IR, the interactions between users and systems can yield *benefits* in terms of acquired information or the fulfilment of information requirements [26, 28]. Rational users who seek to optimise the benefits from their interactions have two options: they can either maximise the benefits they receive or minimise the costs and efforts they expend. In this way, they align with the Principle of Least Effort [317].

Thus, with SET, we can relate changing costs (e.g., the cost of querying or the cost of examining a search result snippet) to changing search behaviours. Prior work in this area have focused on how users interact with ranked list [57, 187], their stopping behaviours [180, 303], the trade-off between querying and assessing [19, 20, 22], and browsing costs [27, 133]. However, in terms of the layout of the SERP, all these models typically assume simplicity where interface components or *widgets* are usually fixed and not part of the formal model. However, contemporary SERPs are complex, and widgets can appear at various positions on the SERP. While prior work [177, 204, 291] have successfully employed formal models to derive testable hypotheses of search behaviours, to the best of our knowledge, none of them have, however, considered the position of a user interface widget as important enough to include in the derived model. Hence, in this thesis, we employ the formal model of Search Economic Theory to predict via simulation how the positioning of a search interface widget impacts the search behaviour of users. With this focus, we selected one specific SERP widget, a **Query History Widget** (QHW) to provide an initial exploration of how to incorporate widget positioning into a SET-based model. It allows a user to view and thus reflect upon their recently issued queries during a search session. The widget is easy to understand for users. It involves only a small number of interactions—making it ideal as the first widget to employ for our exploration and formulate our third broad research question:

B-RQ3: How can we utilise the Search Economic Theory model of user interaction to refine the design hypothesis space for widget positioning?

Findings and Contribution: To this end, in Chapter 4, we derive a position-aware interaction model of search behaviour. We focus on the QHW and formulate a model that can predict search behaviour related to the reissuing of queries from the same search session. We use Fitts' Law to approximate the cost of finding the widget based on its five different positions on the screen. Based on our model and prior work, we develop five testable hypotheses. We conduct a between-subjects user study with $n = 120$ participants. We evaluate the impact of the position of QHW on search behaviour. We find partial support for three out of the five hypotheses based on our study. We did find that widget positioning

plays a role and changes a user’s search behaviour, and thus, position matters—and should be incorporated into formal interaction models.

1.4. RELEVANCE JUDGEMENT COLLECTION

So far in our thesis, we have primarily dealt with user interactions with *text* documents. We focus on a different modality—*audio*. Thanks to the development of voice-based conversational search systems, users have become accustomed to being presented with search results that are read out to them, an approach that is very different from the presentation of text on-screen. In this thesis, we observe the effect of representing documents in audio clips on an essential aspect of IR—relevance judgment collection.

The methodology behind most classical IR research has focused on the Cranfield paradigm. The goal of the paradigm is to measure a given retrieval system’s effectiveness using a set of documents, information needs or queries and standard IR measures, *precision* and *recall*. At the core of the experiments lie the concept of **test collections** consisting of three components—(i) the corpus (collection of documents) to be used; (ii) the statements of different information needs hereafter referred to as topics; and (iii) a set of relevance judgements – a list of relevant documents that the retrieval system should retrieve in evaluating each topic.

Several different evaluation forums have been derived from the Cranfield experimental paradigm to develop improvements in the various retrieval models and other retrieval system components. One of the most well-known evaluation forums is the U.S. government-funded, NIST-sponsored **Text REtrieval Conference** (TREC). Each year, a series of TREC *tracks* are defined, with each consisting of a test collection, in turn consisting of the three components defined above. These tasks are used in conjunction with the relevance judgements provided by assessors. Assessors are usually employees of NIST [225], who were, in turn, previously employed as news analysts by various U.S. security agencies. A series of documents ($\text{top} - k$) are extracted from the document collection using a simple query (a process called pooling). Due to the potentially large document collections, pooling is an acceptable solution to reducing the number of documents to be examined. As an example, given the topic of wildlife extinction, the query *wildlife extinction* is issued over several different retrieval systems. Documents returned are pooled together and then judged by the assessors. For many TREC tracks, judgements are binary, with 0 denoting non-relevance and 1 denoting relevance. Graded relevance judgements can also be used.

The traditional method of employing assessors is typically costly and does not scale up [8] once the number of information needs or k increases. In the last decade, creating test collections using crowdsourcing via platforms like Prolific or Amazon Mechanical Turk (AMT) is a less costly yet reliable alternative [144, 318]. Nevertheless, *how accurate are crowdworkers in their relevance judgement*, this question has been explored by many studies where they have found that the relevance judgement effectiveness of crowdworkers is dependent on several factors including (and not limited to) difficulty of the topic [74], document length [58], their cognitive abilities [245] etc. In this thesis, we focus on an aspect that has received little attention so far: the presentation modality of the documents during the judging process. In this thesis, we posit that by utilising such audio-based devices, we can increase the scope for collecting relevance judgements for text documents in many ways. For example, crowdworkers can contribute by judging documents on their

smartphones [7, 287], if they have visual impairments [224, 290, 319], or if they come from a low-resource background [9, 224]. Although prior work have investigated crowdworker quality and behaviour for the relevance judgement task [74, 106], and tooling to support them in their task, we have few insights into the impact of a document’s presentation modality on assessment efficiency and effectiveness. Hence, in this thesis, we investigate whether it is feasible for crowdworkers to judge the relevance of text documents via a voice-based interface as compared to the traditional way of reading them on a screen.

B-RQ4: How does the presentation modality of documents, text (reading on screen) vs. voice (listening to audio clips) affect the relevance judgement process of assessors? How do the cognitive abilities of assessors and their interplay with presentation modality affect the effectiveness of relevance judgment?

Findings and Contribution: To answer **B-RQ4**, we conduct our last user study (Chapter 5) with 49 crowdworkers where we measure their relevance judgements effectiveness in terms of accuracy, time taken and perceived workload. We also explore the effect of assessors’ cognitive abilities on their judgement effectiveness. Each crowdworker had to judge the relevance of query-passage pairs either by reading the passages on-screen or listening to audio clips. Relevance judgement accuracy was equivalent between crowdworkers reading the passages and those listening. However, as passage length increases, it takes participants significantly longer to make relevance judgements when they listen to them than those reading the passages. Our results suggest that we can leverage the voice modality for this task and the possibility of designing hybrid tasks, where we can use the voice modality for judging shorter passages and text for longer passages.

Our research encompasses four distinct yet interrelated studies, presented in the four chapters of our thesis. We delve into various aspects of user behavior, system design, and the interactions between them. The chapters of this thesis collectively aim to enhance the effectiveness, efficiency, and user experience in information retrieval scenarios.

1.5. THESIS ORIGINS

We now list the publications on which the research chapters are based on.

Chapter 2 is based on the conference paper:

- ❑ Nirmal Roy, David Maxwell and Claudia Hauff. 2022. *Users and Contemporary SERPs: A (Re-) Investigation*. In SIGIR. 2765-2775 [232]

Chapter 3 is based on the conference papers:

- ❑ Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, Claudia Hauff. 2021. *Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment*. In CHIIR. 229-238 [235].

- ❑ Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, Claudia Hauff. 2021. *How Do Active Reading Strategies Affect Learning Outcomes in Web Search?* In ECIR. 368-375 [234].

Chapter 4 is based on the conference paper:

1

- ❑ *Nirmal Roy, Arthur Câmara, David Maxwell, Claudia Hauff. 2022. Incorporating Widget Positioning in Interaction Models of Search Behaviour. In ICTIR. 53-62 [231].*

Chapter 5 is based on the conference paper:

- ❑ *Nirmal Roy, Agathe Balayn, David Maxwell, Claudia Hauff. 2023. Hear Me Out: A Study on the Use of the Voice Modality for Crowdsourced Relevance Assessments. In SIGIR. 718-728 [230].*

2

(RE)-EXAMINING USER INTERACTIONS

The Search Engine Results Page (SERP) has evolved significantly over the last two decades, moving away from the simple ten blue links paradigm to considerably more complex presentations that contain results from multiple verticals and granularities of textual information. Prior work investigated how the presence or absence of heterogeneous content (e.g., images, videos, or news content), the layout of the SERP (list vs. grid layout), and task complexity influenced user interactions on the SERP. In this chapter, we reproduced the user studies conducted in prior work—specifically those of Arguello et al. [16] and Siu and Chaparro [256]—to explore to what extent the findings from research conducted in 2012 and 2014 still hold today as the average web user has become accustomed to SERPs with ever-increasing presentational complexity. To this end, we designed and ran a user study with four different SERP interfaces: (i) a heterogeneous grid; (ii) a heterogeneous list; (iii) a simple grid; and (iv) a simple list. We collected the interactions of 41 study participants over 12 search tasks for our analyses. SERP types and task complexity affected user interactions with search results. We also found evidence to support most (6 out of 8) observations from [16, 256] indicating that user interactions with different interfaces and solving tasks of varying complexity have remained mostly similar over time.

This chapter is based on the following paper:

- ✉ Nirmal Roy, David Maxwell and Claudia Hauff. 2022. *Users and Contemporary SERPs: A (Re-) Investigation*. In SIGIR. 2765-2775 [[232](#)].

2.1. INTRODUCTION

The SERP has evolved significantly over the last two decades, moving away from the *ten blue links* paradigm, to considerably more complex presentations that contain results from multiple verticals and multiple granularities of textual information (snippets, direct answers, entity cards, etc.)—all interleaved within one page. The incorporation of heterogeneous content in a SERP has been shown to change how users interact with web results [13, 15, 45, 81, 165, 194, 240, 304]. How (and where) content is displayed in a SERP affects user interactions as well [16, 249, 265]. While contemporary web SERPs maintain the original idea of a *list* of items that are ranked in decreasing order of relevance, alternative presentations such as a *grid* layout—as also recently (again) popularised by *You.com*—have also been explored [128, 193, 256, 305].

In addition, past research [16, 256, 265, 304] has shown that user behaviour on the SERP does not only depend on the *presentation* of information, but also on the *search task* at hand. For a navigational task such as ‘*find and access the homepage of SIGIR 2022*’, a user—in the ideal case—requires a single query and a single click. Contrast this to an informational task, such as ‘*good restaurants near the venue of SIGIR 2022*’. This requires the scanning of multiple results, and likely results in further query reformulations to learn more about specific suggestions.

As commercial web search engine SERPs have evolved over time (and thus end users have become accustomed to different types of SERPs), we explore in this chapter to what extent user study findings from 2012 and 2014 still hold today. Specifically, we focus our attention on reproducing the experimental setup of two prior studies: Arguello et al. [16] (published in 2012) as well as Siu and Chaparro [256] (published in 2014)—these both investigated how user interactions on the SERP are influenced by the presence or absence of heterogeneous content, the layout of the SERP (list vs. grid), and task complexity. Inspired by the two papers we reproduce, our study is guided by the following research questions.

RQ1 *How does a user’s interactions with a SERP differ when results are presented in a list and grid layout?*

RQ2 *How does task complexity affect user interactions with a SERP?*

RQ3 *What is the interplay between task complexity and SERP layout on user interactions?*

RQ4 *How do users perceive the different SERP layouts?*

To this end, we conduct a user study with $n = 41$ participants that each were given 12 search tasks of varying complexity (ranging from search tasks of type **Remember** to **Analyse**) to solve with one of four different SERP interfaces: (i) *heterogeneous grid*; (ii) *heterogeneous list*; (iii) *simple grid*; and (iv) *simple list*. We explore whether the following eight observations from [16, 256] about users and their interactions with list vs. grid layouts (and heterogeneous vs. simple results) across different task complexities hold today.

O1 Users fixated significantly more on the grid layout SERP compared to the list layout SERP for completing more complex tasks [256].

- O2** On the grid layout SERP, users fixated on search results significantly more for completing more complex tasks compared to simple tasks. A similar observation was found for the list layout SERP [256].
- O3** On the list layout SERP, users fixated significantly longer for completing more complex tasks compared to simple tasks. For the grid SERP, there were no significant differences in fixation duration between varying task complexities [256].
- O4** In the list layout SERP, more complex tasks required significantly greater levels of search interaction: longer search sessions, more clicks on SERP, and more web pages visited [16].
- O5** In a SERP where web results are arranged in a list layout, users clicked on significantly more vertical results when they were present on the main page of the SERP (blended, heterogeneous display) compared to when they were only present as tabs (non-blended, simple display) [16].
- O6** Task complexity did not have a significant effect on user interaction with vertical results in the list layout SERP [16].
- O7** The interplay between task complexity and display of verticals (blended, heterogeneous display vs. non-blended, simple display) did not have a significant effect on user interaction with vertical results in the list layout SERP [16].
- O8** Neither study [16, 256] found significant differences in user evaluation of the different SERP types, list vs grid layout for the former and blended vs non-blended display for the latter, in their experiments.

2

In our user study, we observe that SERP types and task complexity affect user interactions with search results. We also find evidence to support most—6 out of 8—observations from [16, 256].

2.2. RELATED WORK

2.2.1. TASK COMPLEXITY AND USER INTERACTIONS

A number of work have focused on the effect of task types on user interactions on SERPs. Buscher et al. [50] performed a large-scale analysis using query logs to understand how individual and task differences might affect search behaviour. Their findings show that there are cohorts of users who examine search results in a similar way. They also showed that the type of task has a pronounced impact on how users engage with the SERP. Arguello et al. [16] observed that the more complex the task, the more users would interact with various components on the SERP whereas Thomas et al. [274] found that users tended to examine the result list deeper and more quickly when facing complex tasks. Jiang et al. [122] compared user interactions in relatively long search sessions (10 minutes; about 5 queries) for search tasks of four different types. Wu et al. [304] also observed differences in user interactions with the SERP based on whether they had to look for answers to a factoid question or a non-factoid question. In these studies, the SERPs were composed of web search results in the *de facto* list format.

2.2.2. SERP PRESENTATION AND USER INTERACTIONS

Sushmita et al. [265] observed that positioning (top, middle, bottom) of different verticals on a SERP affects clickthrough rates of users when the verticals (news, image and video) were presented in a blended manner with the web search results. Arguello et al. [16] also looked into how task complexity affects user interactions and usage of aggregated vertical results when they are interleaved with web results, versus when they are presented as tabs. On a similar note, they observed that for more complex tasks, users clicked on more vertical results when they were interleaved with web search results. Bota et al. [45] conducted a crowdsourced online user study to investigate the effects of entity cards given ambiguous search topics. They found that the presence of entity cards has a strong effect on both the way users interact with search results and their perceived task workload. Furthermore, Levi et al. [157] performed a comprehensive analysis of the presentation of results from seven different verticals (including a community question answering vertical) based on the logs of a commercial web search engine. They observed that the community question answering vertical receives on average the highest number of clicks compared to other verticals. Wu et al. [304] studied how the presence of answer modules on SERPs affect user behaviour and whether that varies with question types (factoid vs. non-factoid). They found that the answer module helps users complete search tasks more quickly, and reduces user effort. In the presence of answer modules, users' clicks on web search results were significantly reduced while answering factoid questions. Shao et al. [249] conducted a user study to understand how user interaction is affected by the presence of results in the right rail of a heterogeneous SERP in addition to the traditional web results in the left-rail. They found that users have more interactions with the SERPs, appear to struggle more, and feel less satisfied if they examine the right-rail results. Overall, findings observed that the presence of verticals and other heterogeneous modalities of results and their position on the SERP affect user interactions. In these studies, results were also presented in the list format.

Kammerer and Gerjets [128] observed that when web search results are presented in a grid layout, the impact of search result positioning on selecting trustworthy sources is drastically reduced in comparison to the more traditional list approach. Users typically follow a top-down approach when scanning lists, and are more susceptible to select untrustworthy sources if they appear high up in the list. This effect is reduced for a grid-based presentation. However, the authors do not compare different types of tasks, nor do they explore user behaviours when results from various vertical features of the search engines are present on the SERP. Siu and Chaparro [256] compared the eye-tracking data of grid and list SERP layouts with two types of tasks (informational vs. navigational), and investigated potential differences in gaze patterns. The '*F-shaped*' pattern was less prominent on the grid in comparison to the list layout. These two studies explore how user interactions change when web results are presented in a grid vs. a list layout. They do not, however, include vertical results in their study.

2.3. METHODOLOGY

To address our four overarching research questions as outlined in §2.1, we conducted a user study with $n = 41$ participants. Each participant was assigned to one of four experimental search interface conditions (**interfaces**: *between-subjects*), and completed 12 search tasks (**tasks**: *within-subjects*).

Our four experimental search interface conditions considered the **layout type** (*list*-vs. *grid-based* layout) and the **verticals present** on the SERP (*heterogeneous content* vs. *homogeneous content*). These combinations result in the interface conditions outlined below, with examples of the two layout types presented in Figure 2.1, with further details provided in §2.3.1.

SL Simple List Considered as our baseline interface condition (the standard and widely used *ten blue links* [110]), this interface presents results in a list, with each result presented one under the other. All results are *web results*, and as such are *homogeneous* in terms of presented content.

SG Simple Grid The same homogeneous approach to content is taken as for **SL**, but with results presented in a *grid-based* approach. Instead of scrolling along the vertical, participants subjected to this interface scroll along the *horizontal*.

HL Heterogeneous List Similar in approach to **SL**, **HL** presents results in a list. However, different verticals are mixed in with the standard web results. Beyond web results, heterogeneous content used in this study includes *image* and *video* results.

HG Heterogeneous Grid Similar to **HL** but now the content is displayed in grid form, with *web-based* results appearing in a grid, before additional image and video content.

2.3.1. SEARCH INTERFACE DESIGN AND SYSTEM

Given the above, our goal is to find out to what extent the observations from prior studies by both Arguello et al. [16] and Siu and Chaparro [256] are valid after almost a decade of SERP design evolutions (and additions). To operationalise our four experimental interface conditions, we first needed to create a SERP template design that closely mirrors the design of a contemporary web search engine.

For this study, we selected the *Google* SERP as it presents information recognisably, and commands approximately 92% market share.¹ A replica template was created with particular attention paid to the colour schemes, fonts, width and height of components—as well as the spacing between them. The end result was a highly realistic template of a contemporary SERP, on which we based all of our study’s results pages.²

SERP Template Overview Figure 2.1 presents the SERP template used, presenting results for the query ‘[how do dams generate electricity?](#)’. Present are examples for both list- (*Figure 2.1(a)*) and grid-based (*Figure 2.1(b)*) interfaces.

¹<https://gs.statcounter.com/search-engine-market-share>. All URLs in this chapter were last checked on 2022-02-14.

²Templates are released for future user studies, available at <https://github.com/roynirmal/sigir2022-serp-reproducibility>.

2

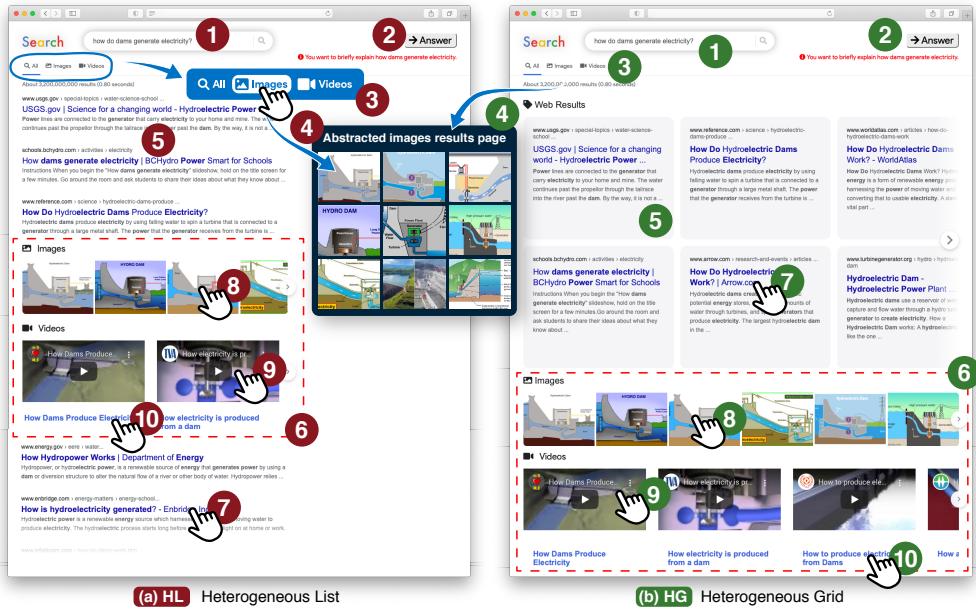


Figure 2.1: Examples of both the (a) list-based and (b) grid-based interfaces trialled. Note the inclusion of links for the separate **All** (as shown), **Images**, and **Videos** result pages. Heterogeneous content is displayed in red boxes, and is *not* present in the two homogeneous content interface conditions (**SG** and **SL**). Circled numbers correspond to the narrative of Section 2.3.1.

A query box is provided ①. However, this is disabled and provides no functionality for this study. It does however display the query terms that were used to derive the presented results *a priori* (see §2.3.2). This is presented next to the *information need* for a given task ②, alongside which there is a button to take the participant to the next stage of the experiment. The SERP template also provides links to additional results pages, namely **All** ③, **Images** ④, and **Videos** ⑤, emulating the setup of the study by Arguello et al. [16]. As shown at ⑥, a grid-based layout is shown for both image and video pages, as is the norm in commercial web search engines such as Google and *Bing*. For images, a total of 16 were displayed (in a 4×4 grid); for videos, a total of nine were shown (in a 3×3 grid).

On the SERP, standard web results are presented ⑦, with the 10 *results per page* (*RPP*) provided. For list-based interfaces ⑧ (SL and **HL**), web results are displayed in the standard way, with one result following the other down the left rail of the SERP. Grid-based interfaces ⑨ (SG and **HG**) present the results in a *carousel* user interface component (emulating the setup of Siu and Chaparro [256]), where the 10 results are arranged in the form of a 5×2 grid. A total of six (3×2) results were visible *above-the-(vertical)-fold*; access to the remaining four results (2×2) was made available through use of a button to scroll across.

As denoted by the red dashed boxes in Figure 2.1 ⑩, heterogeneous content is also added to the SERP template. Present only in experimental search interface conditions **HL** and **HG**, these components provided inline image and video results to the participants. Like web results in the grid-based interfaces, these were also scrollable, mimicking the behaviour of contemporary web search engine SERPs. Sufficient content was placed within

these components to ensure two complete scrolls could be completed; the number of images displayed varied as their widths were variable. On interface condition **SG**, image and video components were placed under the *third* web result; on interface condition **HG**, they were placed directly underneath the web results grid.

The SERP template was also fully interactive—participants could click on links of web results ⑦, with a new browser tab then opening to present the page at the linked URL. In addition, images and videos within the SERP could also be interacted with. Clicking on an image ⑧ took the participant to the webpage containing the image (again, in a new tab). Videos, all sourced from *YouTube*, could be played on the SERP itself ⑨, with the necessary infrastructure in place to enable such functionality. If the participant wished to view the video on the YouTube website itself, they could click the link underneath ⑩ to do so. Again, YouTube links opened a new tab in the participant’s browser.

SERP Definition Note that for each query, there are three unique results pages to replicate the study of Arguello et al. [16]. These are the results ‘*landing*’, or ① **All** page, containing the web results (and additional components, for interface conditions **HL** and **HG**)—as shown in Figure 2.1, as well as the ② **Images** and ③ **Videos** pages. From hereon in, we refer to a SERP as the ‘*landing*’ page, containing web results. To replicate the study of Siu and Chaparro [256], the ‘*landing*’ page itself is sufficient.

Capturing Interactions and Experiences Integrated with the SERP template was **LogUI** [182], a framework-agnostic JavaScript library for capturing different interactions and other events within a web-based environment. **LogUI** was configured to capture a series of mouse events (including hovers, clicks, and scrolls) over the various components of the SERP. Interactions on components included (but were not limited to) web results, images and video contents (including the capturing of the playback, pausing, and completion of YouTube videos). We also recorded interactions to (and on) the supplementary ④ **Images** and ⑤ **Videos** pages. Browser-wide events were also captured, and included the ability for us to compute the time spent *away* from the SERP—when participants would click on a document/image/video link, which would open a new tab.

Experience data was captured via a number of *Qualtrics* surveys;³ a pre- and post-experiment survey were completed by each participant, in addition to the small post-task summary the participants had to write. More details on these questions and the flow of the experiment can be found in §2.3.4. Our setup ensured that participants would jump between the Qualtrics surveys and SERPs as and when required.

Static SERPs As alluded to with the disabled query box ⑥, our experimental setup featured no programmable backend or search functionality. This meant that there was *no* additional querying functionality. We served manually curated SERPs that we produced *a priori* for each of the 12 search tasks we asked participants to undertake. This setup ensured that all study participants viewed the same results (a setup also chosen in prior studies, such as those by Sushmita et al. [265] and Wu et al. [304]). While making the search experience somewhat less realistic, it did provide us with the benefit of not having to deal with participants submitting diverse queries. The design also removed a confounding variable, the query proficiency of the user, and allowed us to address our four RQs by calculating the user interaction measures on a fixed set of web results, images, and videos.

³<https://www.qualtrics.com/>

2.3.2. SEARCH TASKS

For this study, we used four different types of information need:

- **Navigation**, where individuals seek to find particular websites,
- **Remember**, involving the retrieval, recognition, and recalling of relevant knowledge,
- **Understand**, constructing meaning from information sources, and
- **Analyse**, involving the breakdown of information into constituent parts, and determining how they related to one another

For each category, we produced three unique information needs. This led to a total of 12 information needs which are listed in Table 2.1. Particular attention was paid to designing tasks that enticed participants to not only look at web results but also at image and video search results as well.

The choice of our information needs is based on the study designs used by both Arguello et al. [16] and Siu and Chaparro [256]. More specifically, Arguello et al. [16] designed a series of tasks that required different levels of diversity of *information* to complete—as well as different amounts of search effort. Tasks were grounded in the revised *Bloom taxonomy*, as outlined by Anderson and Krathwohl [10].⁴ Search tasks were informational and belong to the **Remember**, **Understand**, and **Analyse** categories. In addition, Siu and Chaparro [256] employed just two categories of tasks for their study—navigational and informational. Our design thus combines the setups of both prior work that we wish to examine.

2.3.3. QUERY SELECTION AND SERP CURATION

To ensure that participants received helpful search results, we required a common search query for each. To this end, we ran a small crowdsourced pilot study on the *Prolific* platform.⁵ This pilot had $n = 25$ workers, with the design largely inspired by the study reported by Bailey et al. [30]. Workers took approximately 10 minutes to complete the task and were paid at the hourly rate of GBP8.00 for their time. All 12 information needs were presented to the workers. They were instructed to type the query terms that they would issue to their web search engine of choice if they were seeking information to address the information need. Collected queries were then normalised (case normalisation, stripped punctuation, whitespace cleanup) and passed through the *Bing Spell Check API* to generate a final canonical form of each submitted query. Subsequently, we determined the most frequently occurring query variation for the 12 tasks, taking this query forward as the one to use for the next stage of our study. These are listed in the parentheses in Table 2.1.⁶

Curating SERPs We then used a combination of the *Bing Web Search API*, *Bing Image Search API*, and *Bing Video Search API* to curate a collection of: *web* (title, snippet text, and target URL); *image* (source image and document URL); and *video* (video source URL) results for each of the 12 queries. Snippet text was truncated to the equivalent of two

⁴The Bloom taxonomy is typically used to create educational materials.

⁵<https://www.prolific.co/>

⁶All query variations belonging to *solar panels installation* information need of **Analyse** tasks were slightly different from each other, as crowdworkers tended to submit natural language queries for this information need. We manually picked the query that we deemed to be the best one for this particular information need.

Table 2.1: Overview of information needs and their type. The rightmost column shows the most popular query obtained from our query selection pilot study, outlined in Section 2.3.3. Numbers in parentheses indicate how many crowdworkers ($n = 25$) submitted the most popular query variation.

Type	Information Need	Most Popular Query Variation
Navigation	You want to find the homepage of Andrew Zimmern, the chef.	andrew zimmern chef (14)
	You want to find the page of Air Jordan on the Nike website.	nike air jordan (14)
	You want to find the page displaying the Flixbus route map in Europe.	flixbus europe route map (6)
Remember	You want to know where is the pituitary gland located in the body.	where is the pituitary gland (9)
	You want to find out what clothes the famous cartoon character Mickey Mouse typically wears.	what clothes does mickey mouse wear (5)
	You want to find out how to calculate the volume of an ellipsoid.	ellipsoid volume formula (4)
Understand	You want to find out the steps required to make a paper airplane.	how to make a paper airplane (10)
	You want to briefly explain how dams generate electricity.	how do dams generate electricity (17)
	You want to find out how to prevent shower mirrors from fogging.	stop shower mirror fogging (3)
Analyse	You want to get into martial arts, but you have no fighting experience. Which form of martial arts is more suitable for beginners?	best martial arts for beginners (3)
	You want to find out the main things to look for while installing solar panels on the roof of a house.	things to consider before installing rooftop solar panels (1)
	You want to buy a new camera lens for taking professional pictures of your friend. Which camera lenses are best for portrait photography?	best camera lenses for portrait photography (3)

2

sentences/lines, as this has been previously shown to be a good trade-off in terms of providing a sufficient *information scent* and encouraging interaction (i.e., clicks) [181]. Video links were filtered to YouTube only, as utilising only one video content provider reduced complexity for playback on our SERPs. Any URLs that proved non-functional or redirected to a 404 page were also removed. The content was then placed on our SERP templates, allowing us to construct SERPs, an image results page, and a video results page for each query. SERP variations for all four search interface conditions were produced.

2.3.4. EXPERIMENTAL PROCEDURE

The 12 search tasks undertaken by each participant were preceded and followed by pre- and post-experiment surveys. We first performed screen and browser viewport resolution checks, requiring that all participants use a maximised browser window with a resolution of 1920×1080 or greater. This ensured that we could guarantee the SERPs displayed to the participants could be viewed without scrollbars along the horizontal. If the checks were successful, participants began the experiment by providing basic demographic information and were also asked minor questions on their search engine usage, specifically on what components on a contemporary SERP they often make use of. In addition, we asked what their preferred search engine is. They were then randomly assigned to one of the four search interface conditions (**SL**, **SG**, **HL**, or **HG**).

Participants were primed to summarise their findings after each search task (in no more than 50 words). Upon acceptance of this instruction, the first search task began, with a SERP similar to the one presented in Figure 2.1(a) or Figure 2.1(b). With the selected query ① and information need ② present, participants then began to examine the content. Participants were *not* given a minimum or maximum amount of time to search. We reiterate that they were also *not* given the opportunity to issue their queries. Once they were satisfied with what they had found, they clicked the →Answer button at the top of the SERP ②, and entered their summary. Once complete, the next task began. This process was repeated for the remaining 11 tasks which were displayed to them in random order. Other researchers have also employed randomisation for condition allocation to minimise topic ordering effects [145, 304].

After the search tasks had been completed, participants then moved on to the post-experiment survey. We used the sub-scales from O'Brien's *Engagement Scale* [203, 208] as was done by Arguello et al. [16]. These are aimed at eliciting their evaluation of the interface they used on the following aspects of engagement: *focused attention*; *perceived usability*; *experience*; *aesthetics*; and *felt involvement*. The engagement scale was originally designed to evaluate shopping websites, and hence we modified/removed the statements pertaining to shopping to suit our needs. For example, we changed the original statement (belonging to *aesthetics* sub-scale) "*This shopping website was aesthetically appealing*" to "*The layout of the results page is aesthetically appealing*". For all statements in the sub-scales, participants indicated their level of agreement (1=strongly agree; 5=strongly agree). We also used the *search effectiveness* sub-scale used by Arguello et al. [16] to evaluate how effective the interfaces were in helping participants find information. In total, we used 26 statements from the six sub-scales to elicit user evaluation of the search interfaces. The reliability scores (Cronbach's Alpha) for the sub-scales are reported in Table 2.3. They were also asked to rate the perceived usefulness of web, image and video results.

2.3.5. STUDY PARTICIPANTS

Like our pilot, we recruited participants from the Prolific platform. Our $n = 41$ participants were native English speakers from the United Kingdom, with a 95% approval rate on the platform, and had a minimum of 250 prior successful task submissions. From our participants, 32.5% identified as female, and 67.5% as male. The mean age of our participants was 36.5 ± 9.7 , with a minimum age of 22 and a maximum of 68. 92% of participants listed *Google* as their preferred search engine, with the remaining 8% identified as *DuckDuckGo* users. With respect to the highest completed education level, 51.2% possess a Bachelors

(or equivalent), 24.4% have a Masters (or equivalent), 19.5% have a high school degree, and 4.9% have an Associate (or equivalent). 95% of participants cited using web results on a contemporary SERP, 78% made use of image results, and 37% used video results.

In our random assignment, 11 participants were assigned to **HG**, with ten participants each assigned to **HL**, **SL**, and **SG**. The experiment lasted on average 40 minutes for the 41 participants. Like our pilot participants, they were compensated at the rate of GBP 8.00 per hour. All participants who registered completed the study; post-hoc checks confirmed that they had provided sensible answers for each task, and as such we approved all who took part for payment. As such, our base analyses are reported over $41 * 12 = 492$ search sessions and their corresponding interaction logs.

2.4. RESULTS AND DISCUSSION

We evaluate if observations **O4-O7** also hold in grid SERPs, **HG** and **SG**.

2.4.1. RQ1: SERP TYPE AND USER INTERACTIONS

Table 2.2 presents results that are relevant to our first three research questions. Here, a ✓ indicates a significant effect ($p < 0.05$) on the particular user interaction, and a ✗ indicates no significant effect.

As seen in Tables 2.2 and 2.3, different SERP types do *not* have a significant effect on user interactions except for: (i) the number of web results clicked (row I, Table 2.3); and (ii) the number of hovers on videos present in the video results page (XV, Table 2.3). Post-hoc tests reveal that participants in the **HL** condition have significantly more web result clicks than their **HG** and **SG** counterparts (I, Table 2.3). **HL** participants also have longer web result reading times compared to participants in any of the other SERP conditions (II, Table 2.3)—albeit not significant. Furthermore, participants with the list interfaces (**HL** and **SL**) have a greater number of hovers over web results compared to **HG** and **SG**. As a result, we cannot confirm **O1** where Siu and Chaparro [256] found significantly more fixation counts on the grid interface than on the list interface. We note that, since we did not record eye gaze data, we are approximating fixation counts by user interactions such as web result clicks and snippet text hovers, as mouse position has been shown to correlate with gaze positions in prior studies [194, 226, 227]. One of the possible reasons for the difference in observation with **O1** can be that our participants are more familiar with the standard list layout of web results, as a majority use Google as their main search engine.

Arguello et al. [16] do not compare user interactions with web results on heterogeneous SERPs vs. simple SERPs. However, we observe that participants using the simple SERP interfaces (**SG** and **SL**) scan web search results to lower depths than those of their heterogeneous interface counterparts (IV, Table 2.3). The lack of information (i.e., fewer verticals) on the SERP requires participants to scan web results to a greater depth in the ranked list.

Based on Table 2.3 (VII-X), we find that on average **HL** participants interact more with image and video results that are present on the SERP compared to their **HG** counterparts. **SG** and **SL** participants interact more with vertical results present in the image and video results page than those of their heterogeneous counterparts (XI-XIV, Table 2.3). Post-hoc tests also reveal that **SL** participants have significantly more hovers on video results on the video results page than participants in the other SERP conditions (XIV, Table 2.3). The lack

Table 2.2: Results of a factorial mixed ANOVA, where interface is between-subjects, and task is within-subjects variable. A ✓ indicates significant effect ($p < 0.05$) on the particular user interaction and ✗ indicates no significance.

User Interactions	SERP Main Effect	Task Main Effect	B/W SERP & Task
Web results clicks	✓ (F = 4.27, $p = 0.01$)	✓ (F = 4.18, $p = 0.01$)	✗
Mean web result reading time (s)	✗	✓ (F = 3.97, $p = 0.01$)	✗
Mean session duration (s)	✗	✓ (F = 12.72, $p < 0.0001$)	✗
Mean web result hover duration (s)	✗	✗	✗
Image clicks (SERP)	✗	✓ (F = 7.24, $p = 0.004$)	✗
Video clicks (SERP)	✗	✗	✗
Image hovers (SERP)	✗	✓ (F = 6.98, $p = 0.009$)	✗
Video hovers (SERP)	✗	✗	✗
Image clicks (image results page)	✗	✓ (F = 4.66, $p = 0.01$)	✗
Video clicks (video results page)	✗	✗	✗
Image hovers (image results page)	✗	✓ (F = 5.39, $p = 0.01$)	✗
Video hovers (video results page)	✓ (F = 3.36, $p = 0.02$)	✗	✗

of vertical results on the SERP makes the participants interact with them in the respective vertical results pages which shows that our informational needs indeed require participants to seek out image and video search results as well. **HG** and **HL** participants seem to be satisfied with vertical results present on the SERP and the former barely interacted with vertical results present in the respective results pages (XI-XIV, Table 2.3). Looking at overall interactions with vertical results (adding interactions with vertical results present on the SERP and the vertical results pages for **HG** and **HL**), we see that **HG** and **HL** have slightly more interactions than **SG** and **SL** respectively. This difference is not significant, but we do see a trend in the line of **O5** where Arguello et al. [16] observed a higher number of vertical result clicks when they were blended with the web results in the SERP. This is compared to when they were only present on the respective vertical results page. On a side note, the higher interactions with vertical results present on the SERP by **HL** participants compared

to **HG** participants (VII-X, Table 2.3, also depicted in Figure 2.2(c)) can be attributed to the fact that images and videos in **HL** SERPs appear in the middle of the web results (between rank 3 and 4) whereas they appear below the web results in **HG** SERPs. Participants in the latter interface expend comparatively more effort to access the vertical results, thereby reducing their interaction. We leave further analysis on the effect of positioning of vertical results on user interaction for future work.

Addressing **RQ1**, we found that the interface has a significant main effect on the clicks on web results and hovers on videos on the video results page but not on other user interactions.

2.4.2. RQ2: TASK COMPLEXITY AND USER INTERACTIONS

Table 2.4 shows that the information needs of the **Analyse** type, which are the most complex among our information needs, warrant most web result clicks (**I**), web result dwell time (**II**) and session duration from participants (**III**). Table 2.2 shows that the main effect of task complexity on these interactions is significant. Participants reach greater web result click depth (**IV**, Table 2.4) for **Analyse** tasks, albeit not significant. Post-hoc tests reveal that (*i*) **Analyse** tasks receive significantly more web result clicks than **Remember** tasks; (*ii*) **Analyse** and **Understand** tasks lead to significantly higher web result dwell times than **Navigational** tasks; and (*iii*) the session duration for **Analyse** and **Understand** tasks are significantly higher than for **Navigational** tasks, while the session duration for **Analyse** tasks is also significantly greater than that for **Remember** tasks. Overall, we find that user interactions on web search results increase as the complexity of information needs increase which is inline with the observations of Arguello et al. [16] and we can partially confirm **O4**.

Arguello et al. [16] did not include **Navigational** tasks in their experiments. We argue that they can be considered as tasks requiring the lowest level of cognition, and as such follow the trend of **O4**—they receive the least interaction among all task categories. The only exception to this was web result clicks—the nature of the task requires participants to click web result links to ascertain that they found the correct page.

We approximate fixation duration in **O3** by observing hover duration over the web results, akin to fixation count in §2.4.1. Although participants hover longer over web results (**V**, Table 2.4) and snippet text (**VI**) for **Remember** tasks compared to other tasks, the difference across tasks is not significant. Moreover, the mean hover duration on web results (snippet and title) for participants belonging to the grid SERP types (**HG** and **SG**) is longer than for those belonging to the list SERP types (**VI**, Table 2.3). As seen from Table 2.2, the interplay between SERP type and task complexity do not have a significant effect on hover duration over web results. As a result, we can only partially confirm **O3** where Siu and Chaparro [256] also do not find significant differences in fixation duration for grid layout for the tasks but they *did* find significantly longer fixation duration on the list layout for more complex tasks.

Among interactions with vertical results present on the SERP (VII-X, Table 2.4), we observe that **Remember** tasks receive the most interactions on average. Post-hoc tests reveal that: (*i*) participants click significantly more on images present on the SERP (**VIII**, Table 2.4) for **Remember** and **Navigational** tasks compared to **Analyse**; and (*ii*) they hover significantly more on images present on the SERP (**X**, Table 2.4) for **Remember** tasks compared to all other task categories. For images present on image results page, we again

Table 2.3: User interactions for different interfaces across all tasks. \dagger indicates that there is a significant main effect of SERP layout on that particular user interaction. HG , HL , SG , SL indicate significant difference with **HG**, **HL**, **SG** and **SI** respectively. Maximum values for each interaction is highlighted in **bold**. Rows VII-X indicate interactions on SERP. r.p. is short for results page.

Row	Interaction		Interface Condition		
		HG	HL	SG	
				SL	
I	Web result clicks \dagger	11.27 (± 8.4) HL	21.30 (± 9.0) ${}^{HG, SG}$	18.20 (± 9.9) HL	18.70 (± 11.9)
II	Avg. web result read time (s)	17.96 (± 12.7)	27.00 (± 28.8)	16.82 (± 9.0)	25.09 (± 10.6)
III	Avg. session duration (s)	94.89 (± 56.4)	106.96 (± 46.8)	98.35 (± 47.5)	109.24 (± 64.5)
IV	Max. web result click depth	3.36 (± 2.6)	3.62 (± 1.4)	4.22 (± 1.9)	4.15 (± 1.6)
V	Total web result hovers	66.55 (± 29.8)	83.40 (± 61.4)	122.70 (± 87.1)	124.40 (± 100.8)
VI	Avg. web result hover dur. (s)	2.91 (± 6.9)	2.49 (± 4.7)	2.20 (± 5.0)	0.80 (± 0.7)
VII	Image clicks	0.82(± 1.2)	2.10 (± 2.4)	-	-
VIII	Video clicks	1.55(± 2.8)	11.20 (± 34.0)	-	-
IX	Image hovers	13.27(± 14.9)	16.30 (± 16.7)	-	-
X	Video hovers	6.18(± 9.7)	13.40 (± 22.62)	-	-
XI	Image clicks (image r.p.)	0.00(± 0.0)	0.40 (± 0.7)	0.70(± 1.0)	1.10 (± 1.4)
XII	Video clicks (video r.p.)	0.00(± 0.0)	0.00(± 0.0)	0.00(± 0.0)	0.70 (± 1.5)
XIII	Image hovers (image r.p.)	0.00(± 0.0)	4.10 (± 10.7)	8.80(± 18.6)	13.00 (± 19.9)
XIV	Video hovers (video r.p.) \dagger	0.00(± 0.0) SL	0.00(± 0.0) SG	0.10(± 0.3) SG	2.80 (± 4.8) ${}^{HG, HL, SG}$
XV	Usefulness of image results	2.45 (± 0.9)	2.60 (± 0.8)	2.40 (± 0.8)	2.90 (± 1.4)
XVI	Usefulness of video results	2.27(± 1.1)	2.30(± 1.1)	2.40 (± 0.8)	2.10(± 0.7)
XVII	Usefulness of web results	4.64(± 0.5)	4.70 (± 0.7)	4.50(± 0.7)	4.70 (± 0.5)
XVIII	Focused attention ($\alpha = 0.85$)	3.45 (± 0.8)	4.33 (± 0.5)	3.70 (± 1.2)	3.40 (± 0.9)
XIX	Experience ($\alpha = 0.8$)	4.39 (± 0.4)	4.12(± 0.8)	3.98(± 0.5)	4.15(± 0.7)
XX	Aesthetics ($\alpha = 0.94$)	3.45(± 0.8)	3.62 (± 1.1)	3.33(± 0.9)	3.50(± 0.7)
XXI	Felt involved ($\alpha = 0.65$)	4.00 (± 0.6)	4.00 (± 0.7)	3.80(± 0.7)	3.77(± 0.5)
XXII	Effectiveness ($\alpha = 0.73$)	4.35(± 0.4)	4.30(± 0.4)	3.92(± 0.4)	4.12(± 0.6)
XXIII	Usability ($\alpha = 0.891$)	3.18(± 0.4)	2.68(± 1.28)	2.87(± 0.7)	3.22 (± 0.7)

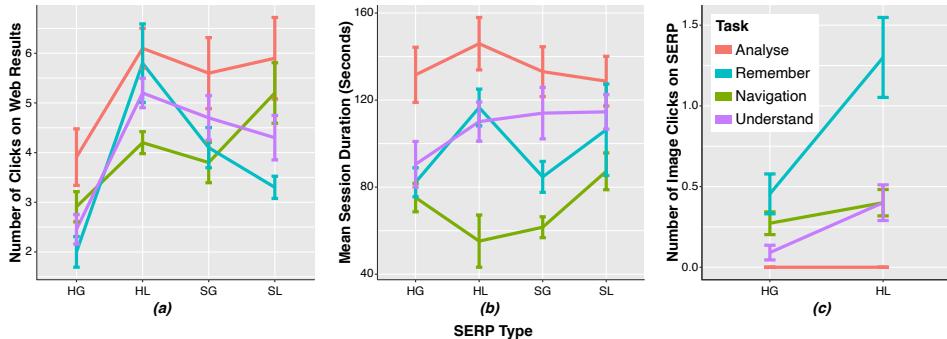


Figure 2.2: Interaction plots, showing effects of SERP types and task complexity over: (a) clicking on web results; (b) the mean session duration (in seconds); and (c) clicks on images presented on the SERP.⁷

observe significantly more image clicks (XII, Table 2.4) and hovers (XIV, Table 2.4) for **Remember** tasks compared to the more complex **Understand** or **Analyse** tasks. Findings regarding user interactions with vertical results (present on the SERP and the vertical results pages) and their relationship with task complexity is contrary to the observations of Arguello et al. [16], and hence we cannot confirm O6. The high interaction with vertical results for **Remember** tasks together with the fact that participants hover over web results and snippet text longer (on average) for the same task (V & VI, Table 2.4) shows that participants prefer to address information needs of the **Remember** type by either hovering over web results and interacting with verticals rather than clicking the link. Arguello et al. [16] do not observe hover duration in their analysis.

To answer RQ2, we find that task complexity does have a significant effect on several user interactions. With participants interacting more with web results as tasks get more complex, we observe significantly more interactions with image results for **Remember** tasks compared to more complex **Analyse/Understand** tasks.

2.4.3. RQ3: TASK COMPLEXITY, SERP TYPE AND USER INTERACTIONS

As seen in Table 2.2, we do not observe a significant effect of the interplay between SERP types and task complexity on user interaction with web results or verticals which is similar to what Arguello et al. [16] found. Hence, we can confirm O7.

From Figure 2.2(a), we observe that participants across all SERP types click the most web results for **Analyse** tasks (in line with O4). For each task type, HG participants click the least number of web results and for most tasks participants with grid SERPs click lower ranked web results than those with list SERPs (in contrary to O1). Approximating fixation count by web result clicks, as done in §2.4.2, we see for each SERP type, the complex **Analyse** tasks receive more interaction than the less complex **Remember** or **Understand** tasks. Although pairwise comparisons do not show a significant difference in web result clicks between different tasks for each SERP, we observe a trend similar to O2—more complex tasks requiring higher document clicks. From Figure 2.2(b), we see that participants across all SERP types take longest to finish **Analyse** tasks and least amount of

⁷Figure 2.2 in the SIGIR proceedings version has a mistake—the legends of **Remember** and **Navigation** are flipped. This is the correct version of the plot.

Table 2.4: User interactions for different task complexity across all search interfaces. † indicates that there is a significant main effect of task complexity on that particular user interaction. \mathcal{N} , \mathcal{R} , \mathcal{U} , \mathcal{A} indicate significant difference with navigational, remember, understand and analyse tasks. Maximum values for each interaction is highlighted in **bold**. Rows VII-X indicate interactions on SERP.

	Interactions	Navigational	Remember	Understand	Analyze
I	Web result clicks †	4.00(± 2.6)	3.76 $^{\mathcal{A}}$ (± 3.2)	4.12(± 2.5)	5.34$^{\mathcal{R}}$(± 4.1)
II	Mean web result reading time (s)†	14.99 $^{\mathcal{A},\mathcal{U}}$ (± 11.6)	20.57(± 21.2)	24.05 $^{\mathcal{N}}$ (± 22.6)	26.91$^{\mathcal{N}}$ (± 29.2)
III	Mean session duration (s)†	69.95 $^{\mathcal{A},\mathcal{U}}$ (± 52.7)	97.11 $^{\mathcal{A}}$ (± 76.0)	106.90 $^{\mathcal{N}}$ (± 62.2)	134.75$^{\mathcal{N},\mathcal{R}}$ (± 74.0)
IV	Maximum web result click depth	3.32(± 2.6)	3.88(± 2.3)	3.85(± 2.4)	4.27(± 2.8)
V	Mean web result hover duration (s)	0.86(± 2.0)	3.26(± 14.4)	2.05(± 8.3)	2.31(± 9.9)
VI	Mean snip. text hover duration (s)	0.08(± 0.2)	0.12(± 0.2)	0.07(± 0.1)	0.08(± 0.1)
VII	Image clicks †	0.17 $^{\mathcal{A}}$ (± 0.4)	0.44$^{\mathcal{A}}$(± 1.0)	0.12(± 0.4)	0.00$^{\mathcal{N},\mathcal{R}}$ (± 0.0)
VIII	Video clicks	0.00(± 0.0)	1.15(± 5.6)	0.12(± 0.6)	1.88(± 11.4)
IX	Image hovers†	1.10 $^{\mathcal{R}}$ (± 2.4)	4.83$^{\mathcal{N},\mathcal{U},\mathcal{A}}$ (± 10.2)	1.07 $^{\mathcal{R}}$ (± 2.5)	0.54$^{\mathcal{R}}$(± 1.9)
X	Video hovers	0.88(± 2.6)	4.49(± 15.7)	1.90(± 5.5)	1.29(± 4.5)
XI	Image clicks (image results page) †	0.07(± 0.3)	0.32$^{\mathcal{U},\mathcal{A}}$ (± 0.6)	0.12 $^{\mathcal{R}}$ (± 0.4)	0.02$^{\mathcal{R}}$(± 0.2)
XII	Video clicks (video results page)	0.00(± 0.0)	0.00(± 0.0)	0.15(± 0.7)	0.02(± 0.2)
XIII	Image hovers (image results page)†	1.15(± 4.1)	4.37$^{\mathcal{A}}$(± 10.9)	0.73(± 2.5)	0.07$^{\mathcal{R}}$(± 0.3)
XIV	Video hovers (video results page)	0.02(± 0.2)	0.00(± 0.0)	0.56(± 2.2)	0.12(± 0.6)

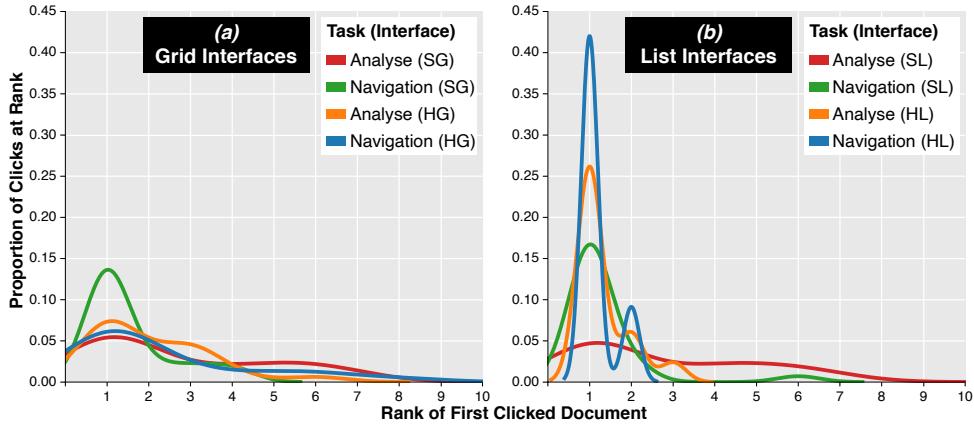


Figure 2.3: Distribution of ranks of the first clicked web results for participants over both grid-based interfaces SL and HL (a), and list-based interfaces SG and HG (b).

time to finish *Navigational* tasks (also in line with O4). Finally, Figure 2.2(c) corroborates our findings from §2.4.2, as we see that participants across both SERP types interact most with image results for **Remember** tasks compared to other tasks. As mentioned earlier, this observation is contrary to what Arguello et al. [16] observed in their study (O6).

In Figure 2.3, we plot the distribution of where (which rank) participants made their first click of web results for **Navigation** and **Analyse** tasks. The **HL** SERP is the only one where the web results are “broken” by vertical results at rank three, and as a result, we observe that most of the first clicks for both tasks appear before rank four (subplot (b) of Figure 2.3). For the other SERPs, the first click distribution for the tasks is more uniform. This is especially prominent for **Analyse** tasks, where we see due to the absence of verticals on **SL** SERP (comparing *Analyse HL* and *Analyse SL* in subplot (b) of Figure 2.3) participants are willing to go further down the list before their first click. We also expect a peak around the first result for *Navigation* tasks, which is true for all SERP types except *Navigation HG* in subplot (a). Either the participants using that SERP type prefer to not click a lot as is evident from Table 2.3 (fewest web result clicks by **HG** participants), or they chose to explore more before their first click. It has been observed in earlier studies [123, 128] that participants have a trust bias for list SERPs (they click on web results appearing higher up the ranked order). The trust bias had been found previously to be less prevalent in grid SERPs [128]. We also find evidence of similar user interaction in subplot (b) compared to subplot (a) where participants are more open to exploration before their first click. To conclude, we find for RQ3, that the interplay between SERP types and task complexity does not have a significant effect on user interactions.

2.4.4. RQ4: PERCEIVED EXPERIENCE OF SERPs

Turning our attention to the post-experiment surveys, we observe little difference in participant ratings of the systems (XVIII-XXIII, Table 2.3). This is in line with both Arguello et al. [16] and Siu and Chaparro [256], who also did not find significant differences in user ratings for different interfaces. Therefore, we can confirm O8.

We also observe that web search results on average are perceived to be more useful (XVII, Table 2.3) than image or video results (XV-XVI, Table 2.3). This is in line with the click behaviour of participants. Across all SERP types, they clicked on more web results than they did on images or videos. Arguello et al. [16] also found the overall number of vertical clicks to be lower than that on web results. Image results were perceived to be more useful by **SL** participants followed by their **HL** counterparts (XV, Table 2.3), which is reflected in their behaviour as well. While the former has the most interactions with images present on the image results page (XI-XIII, Table 2.3) compared to participants in other cohorts, the latter interacted most with images present on the SERP (VII-IX, Table 2.3).

2.5. CONCLUSION

Summary In this chapter, we set out to answer the question of how four different types of SERP and four different types of tasks of varying levels of complexity affect user interaction with web, image and video results. We also explore whether observations about users and their interactions from the studies of Arguello et al. [16] and Siu and Chaparro [256] hold with contemporary SERPs. We observed the following findings with respect to our research questions.

- RQ1** The SERP has a significant main effect on the number of clicks on web results and the number of hovers on videos on the video results page, but not on other user interactions.
- RQ2** Task complexity has a significant effect on user interactions. While participants interact more with web results as the task becomes more complex, we observe significantly more interactions with image results for **Remember** tasks compared to the more complex **Analyse** or **Understand** tasks.
- RQ3** The interplay between SERP types and task complexity does not have a significant effect on user interactions.
- RQ4** There is little difference in the evaluation of the four SERP types by participants.

Out of eight observations, we found evidence to confirm two (**O7, O8**), with partial evidence for a further four (**O2, O3, O4, O5**). These findings indicate that the user interactions over different interfaces for solving tasks of varying complexity have remained mostly similar over time. However, we employed different information needs—and recruited different participants—from the prior studies. Nevertheless, the evidence contrary to **O1** and **O6** has interesting implications—introducing SERPs that users are not familiar with might result in a decrease in interaction. Although the grid layout can present search results in a condensed format (displaying more items in a given screen space compared to the list layout), users might still end up exploring more in the familiar list layout. Additionally, interactions with vertical results are not only dependent on the complexity of the tasks, but also the type of information need. As we observed, certain simpler tasks might warrant more interaction with vertical results than more complex tasks [265].

Reproducing IIR studies Several variables exist that might affect the observations of an IIR study. An unexhaustive list includes the selection of users, interfaces, and task

types. Although both Arguello et al. [16] and Siu and Chaparro [256] described how their respective interfaces looked, they did not point to any resources which would help us replicate them. Moreover, we believe that the more users become familiar with a particular interface, the more important it is to present a similar interface to them during a study examining their behaviours. As mentioned in Section 2.3.1, we have created templates of SERPs that resemble [google.com](#) and [you.com](#), and released them for further use. We believe our templates will be useful for the community to eliminate confounding variables in IIR studies that might arise due to SERP presentation. Secondly, Arguello et al. [16] and Siu and Chaparro [256] did not mention the entire set of tasks used in their studies. As a result, we came up with our tasks of different complexity, as presented in Table 2.1. Two studies by Urgo et al. [285, 286] both list examples of tasks pertaining to different complexities which also offer useful resources for future IIR studies. Our tasks differ with respect to the fact that we designed tasks that specifically enticed participants to not only look at web results, but also to image and video search results as well. It is important to have a fixed set of tasks and similar interfaces to reproduce and enable reliable comparison of observations (e.g., the number of queries, documents opened, etc.) with prior IIR studies. Lastly, in most cases, it will not be possible to have the same participants while reproducing IIR studies. Crowdsourcing provides a solution for capturing user interactions as it has been shown that there is little difference in the quality between crowdsourced and lab-based studies [318]. Power analysis can be used to determine the number of participants required given the experimental conditions of a particular study. It also might be useful to release experimental logs from these studies, after careful ethical checks and considerations. This will permit future researchers to examine them closely, and use them to develop, for example, models of user interaction and search behaviour.

Recent Research Extensions and Future Work There are several areas with scope for future refinement. First, although we tried to select information needs that cover a broad range of topics, we cannot be certain that the results generalise to information needs with other characteristics. Gritz et al. [102] investigated the influence of two interface layouts, *columns* vs *tabs* for multi-modal retrieval results in an academic task on user search behavior, search efficiency, and usability. Although there were no differences between the two layouts with respect to the search behavior of the participants, the subjective usability evaluation revealed significant differences in terms of the supportiveness of the interface, which was rated significantly higher in the column layout. Similar studies need to be conducted in other domains with their own information seeking goals. Second, we did not provide querying functionality to users—and hence it will be worthwhile to explore if that has an overall effect on user interactions. Thirdly, the positions of vertical results on the main page of the SERP were fixed, and we know from previous work [249, 265] that user interactions with verticals is affected by where they are displayed on the SERP. Allen et al. [6] observed user task performance is affected by time constraint where the layout of results play a role on search outcomes and experiences under imposed time constraints. Future work can aim to investigate the effect of other constraints like screen size, lack of web searching experience etc. can have an effect on user interactions. Fourthly, the findings from this study can be further applied to designing and evaluating SERP presentations and the placement of other heterogeneous content. For example, Schultheiß [247] plan to explore how search engine marketing influences user knowledge gain as they search the

web and how task complexity might play a role. [23] recently explored how the presence of distractors like advertisements, sponsored links and clickbait affect user interactions. Taken together, all these studies pave the way for research directions which include designing search interface that help user in their search goals, helping them navigate a plethora of (often distracting) information. Finally, Balog and Zhai [31] mentions the importance of task complexity and SERP layout while building user simulation models for evaluating information access systems. Understanding and modelling user interactions will also help us work on methodologies for interface optimisation [295] and SERP evaluation, along the same veins of prior studies [24, 63, 238, 273, 315].

3

SEARCH AS LEARNING

*Active reading strategies, such as highlighting and note-taking, enhance learners' knowledge and understanding of a topic. Previous research, mostly based on observational studies with a single document, points to these benefits. However, with web search engines now serving as the primary source for learners to find relevant content, users often lack adequate tools for effective sense-making. In the IIR community, efforts have been made to address this by providing notepad-like interfaces to track users' search context during exploratory searches. Despite these endeavours, there is a gap in explicitly measuring the impact of such tools on knowledge and understanding in complex learning-oriented search tasks. In this chapter, we addressed this research gap by conducting a crowdsourced between-subjects study ($N = 115$), where participants were assigned to one of four conditions: (i) **CONTROL** (a standard web search interface); (ii) **HIGH** (highlighting enabled); (iii) **NOTE** (note-taking enabled); and (iv) **HIGH+NOTE** (both highlighting and note-taking enabled interface). We assessed participants' learning with a recall-oriented vocabulary learning task and a cognitively more taxing essay writing task. We found that (i) active reading tools do not aid in the vocabulary learning task. However, (ii) participants in **HIGH** covered 34% more subtopics, and participants in **NOTE** covered 34% more facts in their essays when compared to **CONTROL**. Furthermore, (iii) we observed that incorporating active learning tools significantly changed the search behaviour of participants across a number of measures. Lastly, (iv) out of five hypotheses derived from the education literature on the effect of different active reading strategies on learning outcomes, we could confirm three in the SAL context. Our findings have important design implications for learning-oriented search systems. Learners can benefit from search interfaces equipped with these tools, but some strategies employing these tools are more effective than others.*

This chapter is based on the following papers:

- ❑ Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, Claudia Hauff. 2021. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In CHIIR. 229-238 [[235](#)].
- ❑ Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, Claudia Hauff. 2021. How Do Active Reading Strategies Affect Learning Outcomes in Web Search? In ECIR. 368-375 [[234](#)].

3.1. INTRODUCTION

The process of what is now known as **Search as Learning** (SAL) [67] was first formally defined by Marchionini [174] as an iterative process where learners purposefully engage with a search engine by reading, scanning and processing a large number of documents, *with the ultimate goal of gaining knowledge about one specific learning objective*. Learning via exploring, finding, analysing and evaluating documents [10, 183] containing information relevant to the desired learning objective is a time-consuming and cognitively demanding process. While learners often equate searching for information with searching the web [42, 197], web search engines are not equipped with tools to help users during the complex searches that are necessary in the context of learning [18, 99].

Outside of the web search scenario, *active reading strategies* such as annotating content (highlighting, note-taking, etc.) have been shown to have multiple benefits when engaging in long and complex learning tasks [214, 314]. These tools enable learners to limit their working memory load, as well as articulate and reformulate their thoughts. In turn, this can lead to substantial improvements in the understanding and retention of knowledge [140, 175]. Some of these work also explore different strategies by which learners *use* these tools and their effects on learning outcomes [35, 118, 155, 314]. **Active reading tools** play a number of roles in the text comprehension process. *Highlighting* is used for text selection, and *note-taking* for organisation. Both have been shown to help with the learning process—especially in recall oriented tasks, like a *fill in the blanks test* [218, 314], or *multiple choice questions (MCQ)* [37, 297]. Despite the apparent benefits of active reading tools within a learning context, highlighting and note-taking tools are not found in contemporary web search engines. Efforts have however been made to develop *information organisational tools*. By providing a note-taking interface, they allowed users to keep track of their search context, collect relevant articles and improve sense-making during search [39, 80]. However, none of these work explicitly measured the effect these tools had on learning. A/B testing was not conducted either, meaning no comparison of benefits could be made against a control group.

This chapter addresses the aforementioned research gap. More specifically, in this chapter, we explore the impact that active reading tools—*integrated into the search interface*—have on learning-oriented search tasks, with respect to behavioural and learning outcomes. Furthermore, we investigate how different highlighting and note-taking strategies (shown to be beneficial in learning outside of a SAL setup) affect learning outcomes during a complex, learning-oriented search task. We implement two active reading widgets—a highlighting tool and a note-taking tool—within an experimental search system. We conduct a between-subjects study ($N = 115$) where participants were assigned to one of four conditions, where the search interface contained (or lacked) the aforementioned tools: **CONTROL**, our control interface; **HIGH**, with text highlighting; **NOTE**, with note-taking; and **HIGH+NOTE**, including both tools. Participants were assigned to one of two search topics, with their learning assessed over two tasks: a recall-oriented vocabulary learning (*receptive*) task [189, 233]; and a cognitively demanding essay writing (*critical*) task [162, 260]. As such, this user study aims to address the following research questions.

Table 3.1: The five hypotheses and rationalisations used for this exploratory study.

Hypothesis	Rationale
H1 Learners who consider highlighting to be an important active reading strategy benefit less from it than learners who do not.	According to [314], learners who are <i>less</i> accustomed to highlighting put more effort into the act of highlighting and ultimately a better learning outcome is recorded for them.
H2 Learners directly copying considerable portions of their notes from documents they have viewed benefit less than participants who rephrase content in their own way.	Copying large portions of text reduces the attention of learners to critical details [35]. Rephrasing text while note-taking leads to a deeper processing and understanding of the said text while writing summaries [105].
H3 The number or amount of highlights by learners is <i>not</i> an indicator of learning outcomes.	Prior studies [150, 199, 314] have shown that the amount of highlights is not an indicator of learning outcomes.
H4 Learners who take wordier notes cover more facts in their essays.	Prior works [118, 201] depict conflicting observations regarding wordy notes. For this study, we assume that wordier notes contain more facts [201].
H5 Trained highlighters and note-takers learn significantly more than their untrained counterparts.	[155] and [46] trained learners on effective highlighting and note-taking strategies respectively. They observed that the trained group of learners had significantly greater learning outcomes compared to control groups.

3

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

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

RQ3 *What are the effects of different highlighting and note-taking strategies used by learners on their learning outcomes?*

Key findings. (i) The integration of active learning tools within the search interface does not aid in the receptive tasks. (ii) **HIGH** participants covered 34% more subtopics and **NOTE** participants covered 34% more facts in their essays compared to **CONTROL**. Providing both tools does not improve critical learning. (iii) The type of active learning tools has a significant impact on search behaviour. We find that participants with access to the tools queried less and viewed fewer documents. At the same time, **HIGH** and **HIGH+NOTE** participants spent more time reading documents; their **NOTE** counterparts spent considerable time writing notes. (iv) Out of five hypotheses (summarised in Table 3.1), inspired from the education literature on the effects of different highlighting and note-taking strategies on learning outcomes, we are able to confirm three in the SAL context.

3.2. RELATED WORK

The learning literature suggests that effective utilisation of active reading strategies (such as text highlighting, writing out keywords, note-taking and reflecting) helps to improve metacognitive monitoring of the learning process [83, 87, 198, 220, 297]. In turn, active reading strategies help to improve comprehension. Here, we outline prior work that have examined active learning strategies (pertaining specifically to text highlighting and note-taking), along with a wider discussion of recent, associated work in the SAL domain.

3

TEXT HIGHLIGHTING

Important concepts, ideas and information within a passage of text are often explicitly marked (or *highlighted*) by a learner. This is one of the most common ways to self-regulate learning from text [124, 156, 314]. However, prior work have limitations. They typically examine text highlighting or other active learning tools on *printed text*, or a *single* digital document. Leutner et al. [155] found that teaching learners to use active reading strategies like highlighting—together with lessons on self-regulation—was beneficial for learning. In contrast, Ponce and Mayer [218] found that providing highlighting functionality over a single document did improve the memorisation of highlighted terms, but *did not* lead to improved essay writing skills for their participants (when compared to a control condition, where no highlighting tool was present). Yue et al. [314] demonstrated that the highlighting of printed text improved the recall of keywords for a *fill in the blanks* task. The participants were able to answer more questions correctly from texts that they had highlighted when compared against texts without highlighting.

Ben-Yehudah and Eshet-Alkalai [37] compared text highlighting in both printed and on-screen text, and compared participants' learning here against a control setup (with no highlighting) for both mediums. They observed that highlighting helped in text comprehension (evaluated through a MCQ test), but only for printed text. The authors reasoned that under their setup, highlighting on the on-screen platform was not as *convenient* or *natural* when compared to highlighting on printed text. As a result, participants had to expend greater cognitive loads in the act of highlighting alone. The increase in cognitive load was therefore likely to harm the comprehension of the text. The authors also hypothesised that if highlighting for on-screen text were to become more convenient and natural for the learner, greater cognitive capabilities would be available for a deeper understanding and processing of the text. Liu et al. [161] observed that when used alone, text highlighting *may not be beneficial*. Externalising thoughts together with highlighting can however be effective. We draw on inspiration from these prior work, and examine the benefits (if any) that text highlighting provides to learners in a learning-oriented web search task.

NOTE-TAKING

The externalisation of thoughts can be achieved through careful note-taking as learners read and comprehend information presented to them. Through qualitative interviews, Capra et al. [54] found that users in exploratory search tasks reported note-taking as one of the most common activities during the search session. However, the effects of note-taking on knowledge gain or learning was not explored. Liu et al. [161] observed that for video learning, users showed higher learning gains (compared to a control group) using their active reading tool over video transcripts—which offered text highlighting, note-taking and

questioning functionality. Camporro and Marquardt [53] conducted a study to understand user preferences between paper and on-screen note-taking, where on-screen notes were written on a tablet device. A majority of participants were found to prefer on-screen note-taking, so long as it did not increase their cognitive load by distracting them from listening to presentations. In contrast to the studies that have considered note-taking in the context of a *single document or video lecture*, we explore in this chapter the benefits of note-taking within learning-oriented search tasks that spans *multiple webpages*.

SEARCH AS LEARNING

Previous research within the SAL domain has focused on: (i) understanding user behaviours when undertaking a learning-oriented search task [47, 84, 127, 160, 163, 189]; (ii) exploring different types of users and their behaviours (e.g., novices vs. experts) [40, 84, 202, 233]; and (iii) the optimisation of retrieval functions for learning [267–269].

Liu and Song [160] observed that learners who adapted their source selection strategies (e.g., reading encyclopedia documents from *Wikipedia* for vocabulary learning tasks; or reading Q&A documents from platforms such as *Stack Overflow* for *critical learning tasks*, such as analysing an issue or solving a problem) showed better learning outcomes when compared to learners who did not adapt these strategies. Kalyani and Gadiraju [127] also explored the effects of cognitive complexities for learning tasks (such as *remembering* vs. *applying* knowledge) on search behaviours, and observed that more cognitively taxing tasks led to a higher number of interactions with the search interface.

Characteristics of users have also been shown to influence the amount of learning that takes place during a search session. Gadiraju et al. [95] observed that participants with little prior knowledge achieved higher learning gains than learners with at least some knowledge *a priori*. In contrast, O'Brien et al. [202] found no difference in learning outcomes (measured by essay quality) between domain experts and non-experts. Liu et al. [162] reported that participants in their study underwent knowledge changes during different stages of a search session. However, the changes did not depend on their prior knowledge about a topic. More recently, Roy et al. [233] examined *when* learning occurs during a search session. They observed a difference between participants with higher and lower level prior knowledge levels, with the former showing higher learning gains towards the end of the search session. In this chapter, we are interested in observing the benefits of active reading tools over two different learning tasks—a *low-level, receptive* vocabulary learning task, and a *high-level, critical* essay writing task. Since search and user characteristics have been shown to affect a user's behaviours and learning outcomes [95, 127, 160], we explore how *the inclusion of active reading tools* affect search and learning behaviours during a learning-oriented search task.

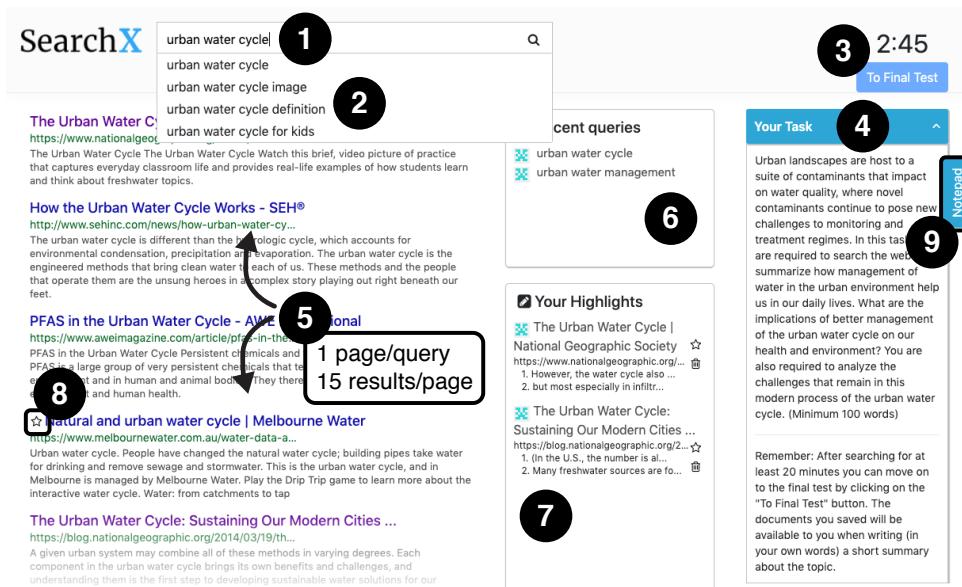


Figure 3.1: The SearchX interface as used for this study with annotations—refer to §3.3.1. This screenshot is an amalgamation of what would have been seen over all experimental conditions; refer to §3.4.1 for details.

3.3. HIGHLIGHTING AND NOTE-TAKING

To carry out our study, we used SearchX [221], a modular open-source retrieval framework that provides out-of-the-box support for crowdsourced interactive IR experiments. The standard interface provides a series of *widgets*, which, when taken together, comprise the look and feel of a contemporary web search engine’s SERP. Figure 3.1 shows the interface we used for our study. Figure 3.2 shows the two widgets we implemented: *highlighting* and *note-taking*.¹

3.3.1. SEARCHX INTERFACE

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

Starting with ①, users are able to enter and submit queries using a standard query box. In addition, we also provide *Query Autocompletion (QAC)* ② to assist users in formulating their queries. To the top right of the interface ③ is a clock indicating the elapsed time that has passed since the search task began, along with a button *To Final Test*. This button

¹All source code, tasks, and descriptions are available online at

<https://github.com/roynirmal/searchx-front-highlighting>

<https://github.com/roynirmal/searchx-back-highlighting>.

becomes enabled after 20 minutes of the search session and when clicked, moves the participant to the next stage in the experiment. The task description is provided at ④ to allow participants to re-familiarise themselves with the task at hand. Results are presented underneath the query box ⑤. Using the conventional and familiar *link/URL/snippet* layout, up to 15 results are presented. Clicking links (blue denoting unread; purple denoting previously examined) will open the document viewer widget, as shown in Figure 3.2a. In our experiment, pagination is *not* included as studies have shown that users often do not move to the second page of results or beyond [95, 189]. We also include a widget that lists queries that the participant previously issued during the session ⑥. The widget lists queries in chronological order, with the most recent query placed at the top. In addition, we also provide a widget that lists previously made highlights ⑦; it presents all documents that contain at least one highlight, as well as the corresponding highlights. Note that if highlighting is disabled for an experimental condition, this widget will simply list documents that participants decide to *save*. That is, participants will instead *save a list of documents* that are deemed to be useful to them in addressing the task. This is achieved by *starring* a document, as shown at ⑧. This is in contrast to when highlighting is present, where participants will curate a *list of highlights that are created over each document examined* (here, starring a document is unavailable). Lastly, we provide note-taking functionality with the **Notepad** button ⑨—see §3.3.4.

3.3.2. SEARCHX LOGGING

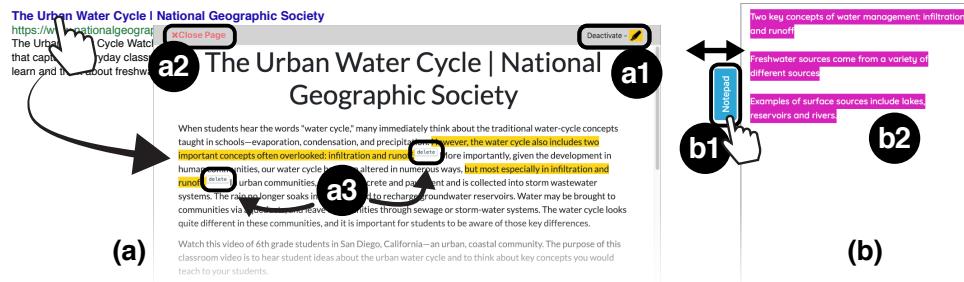
The SearchX system generates fine-grained search logs, allowing us to capture a number of key behavioural measures.² The system also provides a number of quality control features. As an example, participants who switched out of the search interface more than three times were automatically disqualified.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

3.3.3. TEXT HIGHLIGHTING

Encapsulated within the *document widget*, as shown in Figure 3.2a, is the highlighting tool. When presented with a SERP, a participant identifies a document that they wish to examine in more detail. By clicking the link associated with the document, the document widget then appears *on top of the SERP*, with the title and document content shown within the popup that appears. Participants may then begin to highlight portions of text within the document; the highlighter is enabled by default. The participant clicks and drags over the text they wish to highlight, and let go of the mouse or trackpad they are using when they have selected what they wish to highlight. Highlights are automatically saved by the system and made available in the ⑦ **Your Highlights** widget.

Highlights can also be deleted; this is demonstrated by the small **delete** button that appears at the end of the highlight in question, as shown by a3 in Figure 3.2a. The highlighting feature can also be disabled by clicking the button at a1 in Figure 3.2. The

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



3

Figure 3.2: Examples of the two new widgets introduced to SearchX for this study. (a) On the left is the document view, complete with text highlighting capabilities. (b) On the right is the note-taking widget, which is visible when **Notepad** is clicked. Note that these features were not available to all participants of the study; refer to §3.4.1 for more information.

document widget can be closed by clicking **Close Page** at a2, which will then return the participant to the SERP.

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

3.3.4. NOTE-TAKING

In addition to the text highlighting tool, we have also implemented a note-taking widget, stylised as **Notepad**. With experimental conditions that permit it, the note-taking widget is available initially as a non-intrusive ‘tab’-style button, as shown in Figure 3.2b at b1. When the participant clicks on this button, the note-taking widget appears to the right of the viewport, *floating above* all other elements of the SERP. This means that the widget is visible in any state, regardless of whether the document widget is present or not.

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

3.4. USER STUDY DESIGN

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

3.4.1. EXPERIMENTAL CONDITIONS

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

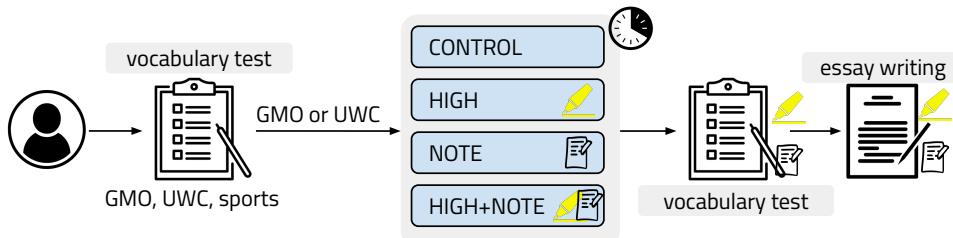


Figure 3.3: Overview of the study's workflow.

CONTROL The standard SearchX search interface is provided *without* highlighting or note-taking capabilities. As outlined in §3.3.1, users are able to save documents and 7 becomes the *Saved Documents* widget.

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

HIGH In this condition the highlighting widget is enabled (i.e., 7 as shown in Figure 3.1).

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

3.4.2. PROCEDURE

Our study flow is illustrated in Figure 3.3. It is inspired by recent studies in the SAL domain [95, 189, 233]. Independent of the experimental condition, participants were first asked to answer seven questions (designed to prime them towards learning-oriented searches, highlighting and note-taking). We asked them to reflect the last time they made such searches and their opinion regarding the benefits of active reading tools. Next, they completed two vocabulary knowledge tests, each one covering 10 vocabulary questions on a particular topic (we also include a third topic as a participant engagement check, outlined in §3.4.3). The topic they know the least about is then chosen as the topic to learn more about during the search session. We randomly assign each participant to one of our four experimental conditions. Participants have to stay in the search phase for at least twenty minutes (hence the timer at 3 in Figure 3.1), as this has been shown to be a reasonable time for people to accrue knowledge [84, 127, 189]. After the minimum search time has passed, participants can continue to the post-test, which consists of a vocabulary test on their topic (we ask *the same* 10 questions as in the pre-test) as well as an essay writing assignment (with a minimum required length of 100 words). During the post-test phase, the participants can access their saved documents, highlights and notes (though editing is now prohibited). Gauging participants' learning across receptive and critical learning tasks (§3.4.3) provides us with a more comprehensive understanding of how much a participant has *learned* about a particular topic than only a recall-oriented vocabulary learning task as conducted in [189, 233]. Lastly, we asked our participants seven reflective questions to gauge the perceived usefulness and ease of using our tools. These questions were restricted to participants that received one or both of those widgets.

3.4.3. TOPICS

In line with prior studies [162, 260], we construct two learning-oriented tasks in our experiment: one *receptive learning task* and one *critical learning task*. Receptive learning is defined as understanding, remembering and reproducing what is taught [151]. Concretely, we ask participants to provide definitions (if they can) of ten vocabulary terms relevant to the topic at hand. In contrast, critical learning includes criticising and evaluating ideas from multiple perspectives [151]. In our study, we ask our participants to analyse challenges and provide their own view of the topic. Overall, the two learning tasks encompass the lower level and higher level cognitive process dimensions of Anderson and Krathwohl's taxonomy [300].

The two topics we used for this study along with the ten vocabulary terms that participants were asked to define in the pre- and post-test, and sub-topics corresponding to the topics were chosen from [73]. They are presented below.³ We also included the topic of *sports*, as an engagement check for all participants in the pre-test: if participants exhibited the same or less prior knowledge on *sports* compared to the other two topics, they were rejected. This is in line with [189, 233] as we expect participants to have reasonably high knowledge regarding the vocabulary terms for the sports topic.

Urban Water Cycle (UWC): Urban landscapes are host to a suite of contaminants that impact water quality, where novel contaminants continue to pose new challenges to monitoring and treatment regimes. In this task, you are required to search the web and summarise how management of water in the urban environment can help us in our daily lives. What are the implications of better management of the urban water cycle on our health and environment? You are also required to analyse the challenges that remain in this modern process of the urban water cycle. (Minimum 100 words).

Vocabulary terms: Lesoto Highlands, eutrophication, Endocrine disrupting compounds, typhoid fever, coagulation, activated carbon filtration, membrane filtration, cholera, Legionella bacteria, recontamination.

3 subtopics: benefits of water management (WM) on health, benefits of WM on environment, remaining challenges

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

Genetically Modified Organisms (GMO): Genetically Modified Organisms (GMOs) have become controversial as their benefits for both food producers and consumers are accompanied by potential biomedical risks and environmental side effects. Imagine in a 'Biotechnology' course, you chose the topic of GMOs. You intend to introduce the benefits of GMOs on modern society to your class. At the same time, you analyse why GMOS can become a potential risk on health, the economy, and society at large – and finally give your conclusion on whether we should progress our research on GMOs and their commercial use. In order to complete this presentation, you need to search for relevant information and prepare an essay for yourself. (Minimum 100 words).

3

Vocabulary terms: transgenic, genomes, selective breeding, microinjection enzyme, chromosome, plasmid, myxoma, kanamycin, severe combined immunodeficiency, Leber's congenital amaurosis.

5 subtopics: benefits of GMOs, risk on health, risk on environment, risk on economy, own conclusion

Sports: Imagine you are taking an introductory course on Sports. For your term paper, you have decided to write about *Sports Development and Coaching*.

Vocabulary terms: olympics, weight lifting, karate, martial arts, aerobics, athletes, soccer, baseball, snowboarding, hockey.

3.4.4. SEARCHX SETUP

SEARCH RESULTS AND QAC SUGGESTIONS

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

DOCUMENT PREFETCHING

As shown in Figure 3.2a, the document widget presents web pages in a heavily altered format. Page-specific styling is removed to yield the content in black text on a white background, complete with images associated with the page (but excluding images within stylesheets, for example). This is done to: (i) make it easier for participants to highlight text (without complex page layouts); (ii) observe their highlights; and (iii) reduce the likelihood of distractions. As such, presenting the document in a timely manner presented a major technical challenge.

To parse content before being viewed, we *prefetched* the documents in the results list returned by the Bing Search API for each query issued. Web pages were accessed and crawled, and stored in a cache. As some pages may have been unavailable (through server downtime, for example), pre-warming the cache with results from previously issued queries was undertaken to minimise the risk of prolonged (5 seconds or more) delays in returning results to participants. Queries for the same topics were selected from the study by [73],

Table 3.2: The number of participants exploring each topic in our study, together with related statistics. Two-way ANOVA tests revealed no significant differences in average number of queries between topics ($F(1,107) = 1.83, p = 0.07$). \pm indicates the standard deviation.

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

3

with the top 50 results saved to the cache. By completing this step, 60,000 documents were prefetched and stored.

REMOVAL OF WIKIPEDIA(-LIKE) PAGES

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

3.4.5. PARTICIPANTS

Since insights from crowdsourced experiments are comparable to lab-based ones [94, 318], we recruited participants for our study using the crowdsourcing platform *Prolific*.⁶ The platform has been shown to be an effective choice for relatively complex and time-consuming interactive information retrieval experiments [308]. The study was undertaken over a two day period in the autumn of 2020. To ensure reliable and high-quality responses, we required our participants to have: (i) successfully completed 100 prior submissions on the *Prolific* platform; (ii) possess an approval rate of 90% or higher; and (iii) have native proficiency in English. Including the minimum search time of 20 minutes, the complete study took approximately forty five minutes to complete. For their time, participants were compensated at the rate of GBP£7.50 per hour.

We computed the required sample size in a power analysis for a *Between-Subjects* ANOVA using the software *G*Power* [89], resulting in the sample size of 120 participants. In all, 131 participants completed our study; 16 submissions were rejected based on our quality control criteria.⁷ This led to the headline figure of $N = 115$. Of the valid participants,

⁴https://en.wikipedia.org/wiki/Genetically_modified_organism

⁵This curated list is provided by Câmara et al. [73].

⁶<https://www.prolific.co/>

⁷Quality control criteria included counting browser blurring events (discussed in §3.3.2); participants should issue at least two queries, view two documents, and finish the post-test with a reasonable essay (as deemed through a manual evaluation).

64 identified as male, and 48 identified as female—with 3 withholding their gender identity. In terms of age, participants reported a median age of 33 (youngest 18; oldest 72). A total of 37 participants reported the highest formal education level as a *high school degree/diploma*. 48 reported a *Bachelor's degree*, with 11 possessing a *Master's degree*. The remaining 19 participants reported other education levels.

Table 3.2 reports the number of participants per topic, over each of the four conditions trialled. Of the 115 participants, 71 were assigned to the **GMO** topic, with the remaining 44 to the **UWC** topic. Remember that topics were assigned to participants based on their pre-task surveys (participants received the topic they had the least knowledge about), leading to a skewing towards the **GMO** topic. The table also contains basic statistics on the number of queries issued which is comparable to that reported in previous studies [127, 189] and shows that participants were fairly active on our platform; refer to §3.5 for more information.

3.4.6. MEASURING LEARNING

REALISED POTENTIAL LEARNING (RPL)

Our vocabulary learning task is evaluated via the *Vocabulary Knowledge Scale* [298] which the participants use to rate their knowledge in line with prior work [69, 189, 250, 267, 268].

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

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

$$ALG = \frac{1}{n} \sum_{i=1}^n \max(0, s^{post}(t_i) - s^{pre}(t_i)). \quad (3.1)$$

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

$$RPL = \frac{ALG}{\frac{1}{n} \sum_{i=1}^n 2 - s^{pre}(t_i)}. \quad (3.2)$$

3

Table 3.3: Example annotation of facts and subtopics and the computation of F-Fact and T-Depth. Note that sentences demonstrating knowledge of the topic are colour coded—each colour pertains to an individual subtopic (see the *T-Depth* column).

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

T-DEPTH, F-FACT AND READABILITY

For the critical learning task, we determine participants' knowledge expressed in their essays by following the work of Wilson and Wilson [301], who proposed and compared a number of measures for this very task. Concretely, we employ *F-Fact*, which counts the number of individual facts present in a summary, and *T-Depth*, which rates to what extent each subtopic is covered in the summary on a scale of 0–3 (from *not covered at all* to *covered with great focus*), as both of these measures were shown to be good indicators of learning. Both of these measures require a manual annotation effort. A concrete example of how we annotated facts and subtopic coverage in participants' summaries is provided in Table 3.3. Three annotators split the 115 essays among them. There were 18 essays which were analysed by all annotators; observing a Pearson correlation of 0.78 ($p = 0.002$) for *T-Depth* scores, and a correlation of 0.76 ($p = 0.002$) for *F-Fact* scores which indicates high inter-annotator agreement. Lastly, as neither of those measures is concerned with the readability of participants' essays, we also computed the Flesch-Kincaid readability scores.⁸ A high score indicates that the text is fairly easy to read, whereas a lower score indicates that the text is fairly complex and can be best understood by university graduates.

HYPOTHESES FOR RQ3

After obtaining the essay scores, we operationalised our five hypotheses (as detailed in Table 3.1) based on our collected data as follows:

- H1:** Learners were asked *Do you think highlighting is useful?* during the pre-questionnaire. This was an open question; we manually analysed their answers and divide them into *pro*, *unsure* and *anti* highlighters.⁹
- H2:** We calculated how many terms from the learners' notes are taken verbatim from the documents they read. The more terms that overlapped, the more we assumed text was directly taken from the examined documents.
- H3:** We divided (median-split) learners into *heavy* and *light* highlighters based on two separate conditions: (i) the total number of highlighting actions; and (ii) the total number of words highlighted.
- H4:** We divided (median-split) learners into *heavy* and *light* note-takers based on the total number of words written in their note-taking tool.
- H5:** We make two assumptions to distinguish between trained and untrained highlighters and note-takers: (i) learners who frequently engaged in highlighting and note-taking prior to the study are considered to be trained (learners were asked the open question: *How often do you highlight and take notes while learning?* during the pre-questionnaire),¹⁰ and (ii) based on their education level—learners having a bachelor's, master's or a doctorate degree are considered to be trained.

⁸We use `textstat` for computing the Flesch readability score.

⁹Pro - *A great extent*; Unsure -*It's a mild benefit to me*; Anti - *I don't think highlighting itself helps me all that much*.

¹⁰Trained - *Almost always if I see something very new to me*; Untrained - *Rarely*

3.5. RESULTS

First, we need to assess the reliability of the participants self assessment regarding their vocabulary knowledge. We randomly sampled 50 answers for knowledge levels (3) and (4); labelling them as *correct*, *partially correct*,¹¹ or *incorrect*.¹² We found that for knowledge level (3), 38% of the answers were correct, 56% were partially correct and remaining 6% were incorrect. Out of the answers self-assessed as (4) 72% were correct, 26% were partially correct and remaining 2% were incorrect. Based on these numbers we argue that the self-assessments of the participants are largely reliable. On average participants marked $2.2(\pm 1.8)$ answers as knowledge levels (3) or (4). This indicates that the participants still needed to learn fair bit of the topics for our tasks.

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

3.5.1. RQ1: HIGHLIGHTING, NOTE-TAKING AND LEARNING

Our first research question, **RQ1**, considers *how beneficial the highlighting and note-taking widgets are for learning-oriented search tasks, when compared to a standard web search interface*. Table 3.4 presents an overview of our learning measures (amongst other behavioural measures) across our four experimental conditions. We report the RPL (**III**), T-Depth essay scores (**IV**), F-Fact essay scores (**V**), and Flesch essay scores (**V**). We first examine the effects of highlighting and note-taking on vocabulary learning.

Our analysis shows that mean RPL scores varied between 0.11 (**CONTROL**) and 0.15 (**NOTE**), all with similar levels of variance. Indeed, our ANOVA analysis yielded no significant differences between the four conditions. The reported mean RPL figures showed that participants gained less than 20% of the knowledge that *could* have been acquired when considering the results of their receptive learning surveys. This finding shows that although highlighting tools have been shown to improve receptive knowledge while learning from a single document [37, 218, 297, 314], they do not aid receptive learning to a similar extent in complex search sessions. Further analysis showed a very small fraction of vocabulary terms that were present in the recorded text highlights (**XVII**) or notes (**XXI**).

T-Depth, F-Fact and Flesch essay scores, that pertain to evaluating critical learning ability, are presented on rows **IV**, **V** and **VI** respectively in Table 3.4. Looking first at the T-Depth essay scores, we see a general trend showing that for conditions where additional tools were available (**HIGH** at 1.64 ± 0.59 , **NOTE** at 1.40 ± 0.61 , and **HIGH+NOTE** at 1.48 ± 0.67), more subtopics were covered by participants in sufficient detail than when compared to those assigned to **CONTROL** at 1.22 ± 0.43 . Post-hoc analysis yielded a significant difference between the **CONTROL** and **HIGH** conditions ($F(3, 107) = 2.72, p = 0.04$). Significant differences were also found between conditions **CONTROL** (14.56 ± 10.36) and **NOTE** (19.59 ± 8.53) when looking at the F-Fact scores ($F(3, 107) = 2.68, p = 0.04$). We observed higher mean F-Fact scores corresponding to **HIGH**, **NOTE**, and **HIGH+NOTE** when compared to **CONTROL** (although this difference was not statistically significant for

¹¹An example of partially correct answer from UWC topic: *an illness for the vocabulary term typhoid fever*.

¹²An example of incorrect answer from GMO topic: *Relating to plasma for the vocabulary term plasmid*.

Table 3.4: Mean (\pm standard deviations) of RPL and search behaviour metrics across all participants in each condition. A dagger (†) denotes two-way ANOVA significance, while C , H , N , B , indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) over the four conditions CONTROL, HIGH, NOTE and HIGH+NOTE respectively.

Measure	CONTROL		HIGH		NOTE		HIGH+NOTE	
	I. Number of participants	32	29	29	29	25	25	25
II. Search session duration (minutes)	23m40s($\pm 10m26s$)	27m53s($\pm 9m50s$)	20m33s($\pm 6m46s$)	29m17s($\pm 15m15s$)				
III. RPL	0.11(± 0.10)	0.14(± 0.21)	0.15(± 0.15)	0.11(± 0.10)				
IV. T -Depth scores of essays [†]	1.22(± 0.43) ^H	1.64(± 0.59) ^C	1.40(± 0.61)	1.48(± 0.67)				
V. F -Fact scores of essays [†]	14.56(± 10.36) ^N	16.55(± 5.51)	19.59(± 8.53) ^C	15.92(± 8.06)				
VI. Flesch scores of essays [†]	32.19(± 39.78)	21.40(± 62.71)	15.86(± 61.54) ^B	46.43(± 16.75) ^N				
VII. Mean #. of essay terms	181.56(± 76.35)	200.83(± 85.61)	225.86(± 112.53)	193.00(± 87.94)				
VIII. Number of queries [†]	5.81(± 3.54) ^{H,B}	4.63(± 2.68) ^C	4.93(± 2.53)	4.28(± 1.74) ^C				
IX. Average time between queries (seconds)	281.95(± 271.50)	307.22s(± 265.15)	254.20s(± 148.67)	289.49s(± 206.90)				
X. Average time between documents (secs.)	96.60(± 49.19)	68.53(± 174.82)	121.30(± 65.72)	114.25(± 174.20)				
XI. Average document dwell time (secs.) [†]	339.50(± 34.74) ^{H,N,B}	522.13(± 38.86) ^{C,N}	166.03(± 26.16) ^{C,H,B}	470.64(± 50.80) ^{C,N}				
Docs.								
XII. Number of unique documents viewed [†]	12.16(± 5.95) ^{H,N,B}	8.07(± 3.95) ^C	8.24(± 3.57) ^C	8.60(± 3.50) ^C				
XIII. Number of unique document snippets viewed [†]	99.72(± 57.34) ^{H,B}	72.93(± 35.19) ^C	87.03(± 38.37)	71.00(± 25.25) ^C				
XIV. Number of highlight additions	—	53.53(± 38.64)	—	50.56(± 43.99)				
XV. Number of highlight deletions	—	7.20(± 17.56)	—	4.36(± 11.25)				
XVI. Number of words highlighted per highlight action	—	29.48(± 17.37)	—	30.30(± 7.87)				
XVII. Fraction of vocabulary terms present in highlights	—	0.04(± 0.06)	—	0.06(± 0.08)				
XVIII. Fraction of essay terms present in highlights	—	0.38(± 0.16)	—	0.44(± 0.21)				
XIX. Percentage of user who took notes ^s	—	—	86.2%	52%				
Notes								
XX. Number of words in note-pad	—	—	1014.17(± 2475.23)	379.04(± 907.39)				
XXI. Fraction of vocabulary terms present in notes	—	—	0.04(± 0.09)	0.00(± 0.02)				
XXII. Fraction of essay terms present in notes [†]	—	—	0.37(± 0.25) ^B	0.20(± 0.28) ^N				
Perception								
XXXIII. Ease of highlighting tool (1 (easy) - 5 (difficult))	—	1.48(± 0.91)	—	1.88(± 1.17)				
XXXIV. Usefulness of highlighting tool (1 (not useful) - 5 (useful))	—	3.93(± 1.33)	—	3.80(± 1.38)				
XXXV. Ease of notepad tool (1 (easy) - 5 (difficult))	—	—	2.07(± 1.28)	2.24(± 1.39)				
XXXVI. Usefulness of notepad tool (1 (not useful) - 5 (useful))	—	—	3.76(± 1.27)	3.08(± 1.55)				

HIGH and **HIGH+NOTE**). This suggests that participants in other conditions discussed a greater number of facts in their essays when compared to their counterparts in **CONTROL**.

Turning our attention to the readability of the participant’s essays, we observe that the Flesch readability scores (row VI, Table 3.3) also differ significantly between conditions. Essays written by participants subject to **CONTROL** on average were easier to read than **HIGH** and **NOTE**. Additionally, essays written with the **NOTE** condition were significantly more complex to read than those written by **HIGH+NOTE** ($F(3,107) = 2.64, p = 0.04$). We should also note that we observed negative Flesch scores for essays of 14 participants across all conditions. This typically happens when participants do not write complete sentences (e.g., bullet points) which renders the pieces of text to be more difficult to read.

From the above, we can see that highlighting and note-taking functionality aid different aspects of essay writing, with the former helping with subtopic coverage, and the latter with fact coverage. However, using both in tandem (**HIGH+NOTE**) does not lead to any significant learning outcome improvements compared to **CONTROL**. Our results contradict those found by Ponce and Mayer [218], who did not observe any significant differences in essay quality amongst participants with and without highlighting capabilities on the systems they used. However, it is important to note here that in the aforementioned study the participants had access only to a *single document*, and essays were evaluated using different measures (a presence of nine pre-defined items in the essays).

3.5.2. RQ2: HIGHLIGHTING, NOTE-TAKING AND SEARCH BEHAVIOUR OF USERS

RQ2 considered *how active reading tools altered the search behaviour of participants*. For this question, we observe that the participants having access to one or both active reading tools issued fewer queries than those in **CONTROL**, and significantly so for **HIGH** and **HIGH+NOTE**. Previous studies [95, 189, 233, 312] have shown that participants issuing more queries observe higher knowledge gains in the receptive vocabulary learning task. The observations in our study might explain the lack of significant difference in RPL scores. However, despite issuing fewer queries (4.63 ± 2.68 vs. 5.81 ± 3.54 , row VIII, Table 3.4), **HIGH** participants cover significantly more subtopics in their essays than their **CONTROL** counterparts ($F(3,107) = 2.68, p = 0.04$). Looking deeper, we observe that participants in **HIGH** spend significantly more time reading documents than those in **CONTROL** (row XI) ($F(3,107) = 5.63, p = 0.001$). This suggests that the highlighting tool facilitates user reflection more while reading a particular document, thereby internalising concepts more effectively than participants in **CONTROL**. The higher document dwell times for **HIGH** and **HIGH+NOTE** participants are in line with findings by Ben-Yehudah and Eshet-Alkalai [37], where highlighting was shown to increase reading time of documents. Comparing the highlighting behaviour of the two groups in Figure 3.4, we observe a similar trend. Most of the highlighting activities are performed at the beginning of the search session. Later, highlights decrease to below 5 on average. This is coupled with the fact that fewer participants are involved in highlighting activity.

Document dwell time is however significantly lower for participants in **NOTE** (166.03 ± 26.16 secs.) when compared to all other conditions (e.g., 522.13 ± 38.86 secs. for **HIGH**). Although not significant, participants on average in **NOTE** spent more time on the SERP between reading two consecutive documents. This together with the lower number of snippets viewed indicates that participants in **NOTE** take notes after reading a particular

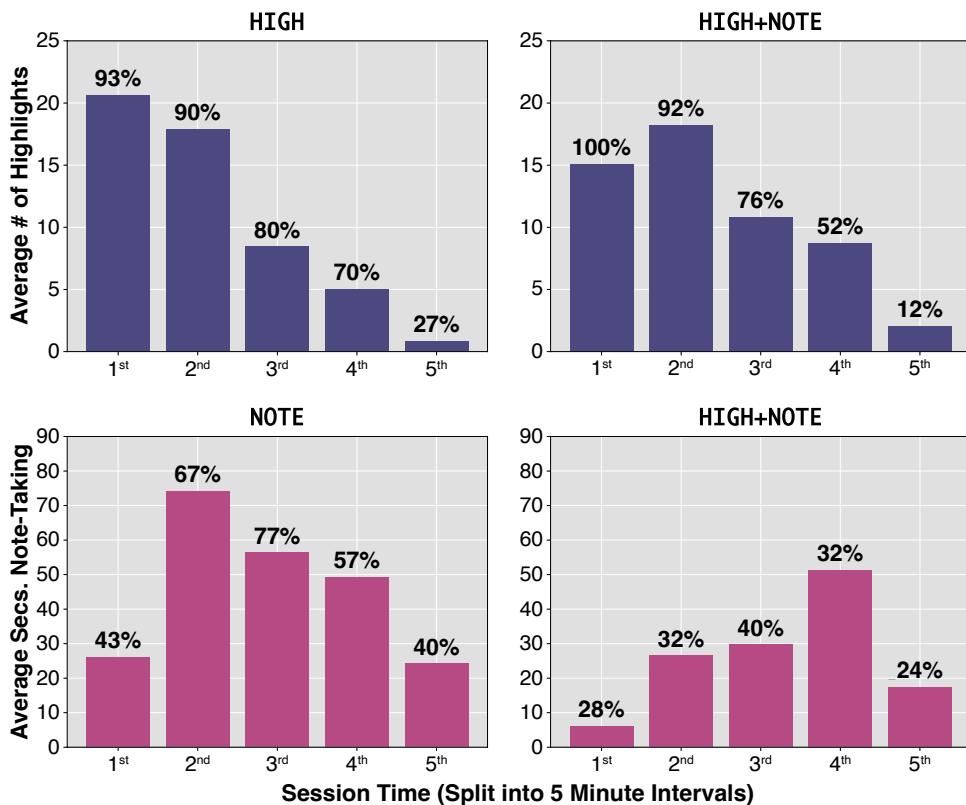


Figure 3.4: Average #highlights and average #seconds of note-taking activity in each five minute interval. The number on top of each bar shows the % of participants with 1+ highlights or 1+ seconds note-taking activity in that interval.

document. We also observe that a significantly large portion of essay terms come from the notes of **NOTE** participants compared with **HIGH+NOTE** participants. This can explain the significantly more complex essays (indicated by the Flesch scores) written by **NOTE** participants when compared to that of their **HIGH+NOTE** counterparts. **NOTE** participants also wrote the longest essays on average—albeit not significantly so. Moreover, when compared to **HIGH+NOTE**, **NOTE** participants took more notes (row XX, there was no significant difference due to the high variance). From Figure 3.4, we see that **NOTE** participants take more notes towards the beginning—with **HIGH+NOTE** towards the end. For the latter, this also coincides with the time period where they are highlighting less. Thus, spending time on taking notes can be a contributing factor for participants acquiring more knowledge—and consequently using this in their essays.

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

3.5.3. RQ3: ACTIVE READING STRATEGIES AND LEARNING

The basic learner statistics for each condition are shown in Table 3.4. We observe that **HIGH** learners cover significantly more subtopics in their essays (**T-Depth**, III), whereas **NOTE** learners write significantly more facts than their **CONTROL** counterparts (**F-Fact**, IV). Essays written by **NOTE** learners were also significantly more complex to read compared to **HIGH+NOTE** learners (**Flesch**, V). Incorporating both highlighting and note-taking tools does not lead to a significant improvement in learning outcomes.

H1: We did not observe a significant difference (Table 3.5) for Flesch scores (V) and F-Fact (III) between the three groups of highlighters belonging to **HIGH** and **HIGH+NOTE** when compared to the three groups of **CONTROL**. However, we observed significant differences for T-Depth ($F(2, 77) = 6.44, p = 0.002$). Post-hoc tests revealed that unsure highlighters belonging to both **HIGH** and **HIGH+NOTE** cover significantly more subtopics in their essays than their **CONTROL** counterparts. Anti-highlighters belonging to **HIGH** show better learning outcomes compared to anti-highlighters belonging to **CONTROL**, whereas pro-highlighters belonging to **HIGH** and **HIGH+NOTE** gain no benefits. This is in line with the findings of [314] and shows evidence for our hypothesis. This might be attributed to the fact that learners who are not sure about the benefits of highlighting put more effort in the act of highlighting itself. This also indicates that highlighting makes some learners process text in a way different from how they normally would, which eventually leads to a better understanding of the text.

H2: From Table 3.4, we find that notes of learners from both **NOTE** and **HIGH+NOTE** on average have 10% overlap with the documents they read (row X). Hence, when we combine all note-takers, we see that those who have more than 10% of their notes over-

Table 3.5: H1: Learners are divided into *pro-highlighters*, *unsure* or *anti-highlighters*. [†] indicates two-way ANOVA significance, while ^C, ^H, ^B, indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) with Holm-Bonferroni correction.

	CONTROL			HIGH			HIGH+NOTE		
	Pro	Unsure	Anti	Pro	Unsure	Anti	Pro	Unsure	Anti
I #users	9	13	10	13	11	5	11	7	7
II #words highli.	—	—	—	1529.8 (333.1)	1944.6 (1018.2)	1174.2 (126.6)	1703.0 (319.0)	1826.7 (790.1)	974.1 (490.3)
III F-Fact	13.1(1.9)	16.3(3.9)	13.6(3)	17.1(1.3)	14.6(1.5)	19.6(3.6)	16.2(2.4)	17.9(3.7)	13.6(2.5)
IV T-Depth [†]	1.2(0.2)	1.2(0.1) ^{H,B}	1.2(0.1) ^H	1.4(0.1)	1.6(0.1) ^C	2.3(0.2) ^C	1.2(0.2)	1.7(0.1) ^C	1.8(0.3)
V Flesch	35.7(7.7)	25.9(12.1)	37.3(15.4)	8.0(18.4)	27.3(21.7)	43.3(31)	48.9(5.2)	41.7(2.9)	47.2(8.7)

Table 3.6: **H3, H4:** Learners are divided into two groups (*heavy* and *light*) based on the median values for each active reading strategy. The learning metrics are computed separately for each group. The significant differences obtained from TukeyHSD pairwise test are highlighted in **bold**.

	F-Fact		T-Depth		Flesch Scores	
	Heavy	Light	Heavy	Light	Heavy	Light
I. #Highlight Actions	15.9	16.6	1.5	1.6	32.4	33.5
	±1.2	±1.4	±0.1	±0.1	±7.1	±11.3
II. #Highlighted Words	17.0	15.5	1.4	1.7	26.4	39.5
	±1.3	±1.3	±0.1	±0.1	±9.5	±9.2
III. #Words in Note-pad	20.0	15.7	1.4	1.5	11.6	48.4
	±1.8	±1.4	±0.1	±0.1	±12.0	±2.8

3

lapped with the viewed documents, covered significantly more facts (F-Facts) than whose notes overlapped less than 10% ($t(38) = 2.04, p = 0.04$), which shows evidence *against* our hypothesis. However, the former explored less subtopics and wrote more complex essays (although not significantly) than the latter. This shows that although copying considerable portions of text into notes might not be beneficial for certain aspects of essay writing like topical coverage, they can be useful when the essays require more factual information.

H3: From Table 3.4, we observe no significant difference between learners of **HIGH** and **HIGH+NOTE** when comparing learning metrics, the number of highlight actions (VII) and words highlighted (VIII). Following this, dividing learners into *heavy* and *light* highlighters, we see from Table 3.6 the amount of highlighting is not an indicator of learning since there is no significant difference between the two cohorts (I, II), thus providing evidence *for* our hypothesis. This indicates that the act of highlighting alone does not benefit learning—it has to be coupled with a deeper cognitive processing of the text.

H4: NOTE learners cover significantly more facts in their essays compared to their **CONTROL** counterparts (IV), cover significantly more essay terms in their notes (XI), and write more complex essays (V) than their **HIGH+NOTE** counterparts (Table 3.4). Furthermore, albeit not significantly, **NOTE** learners write wordier notes (XI) compared to **HIGH+NOTE** learners (Table 3.4). This shows evidence *for* our hypothesis that wordy notes benefit learners in our given task. Table 3.6 further corroborates our hypothesis where we see that learners who take wordier notes (*heavy* note-takers) cover significantly more facts in their essays, and write significantly more complex essays (III). This indicates that taking wordy notes and having access to them while writing their essays help learners to cover more factual information.

H5: When we divide learners based on their prior highlighting experience, we observe a significant difference for T-Depth (Table 3.7)—untrained highlighters cover more subtopics in their essays (I). Prior note-taking experience does not benefit learners. We also do not see any significant learning difference between trained and untrained highlighters/note-takers when we divide them based on their education level. These results show evidence *against* our hypothesis that being trained in highlighting and note-taking benefits learners. This

Table 3.7: **H5:** Participants are divided into two groups (*trained* and *non-trained*) based on their self reported highlighting and note-taking frequency and based on their education level. The learning metrics are computed separately for each group. The significant differences obtained from TukeyHSD pairwise tests are in **bold**.

	F-Fact		T-Depth		Flesch Score	
	Trained	Non trained	Trained	Non trained	Trained	Non trained
I. Prior highlighting frequency	16.8 ±1.3	15.8 ±1.3	1.4 ±0.1	1.7 ±0.1	28.8 ±12.2	36.6 ±6.5
II. Highlighter Education Level	16.4 ±1.2	15.7 ±1.5	1.7 ±0.1	1.5 ±0.1	36.6 ±8.8	27.4 ±10.8
III. Prior note-taking frequency	18.9 ±1.5	16.6 ±1.8	1.6 ±0.1	1.3 ±0.1	28.7 ±8.8	31.8 ±10.3
IV. Note-taker education level	19.5 ±1.6	15.7 ±1.7	1.5 ±0.1	1.4 ±0.1	23.8 ±10.8	36.9 ±6.4

indicates that if learners are prevented from learning using strategies they employ, the cost of prevention does not outweigh the benefits of using a highlighting or a note-taking tool. Although these results do not follow the observations from [46, 155], it needs to be considered that in those studies, the experimental groups of learners were trained specifically about efficient highlighting and note-taking strategies.

3.6. CONCLUSIONS

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

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

queries, and spending more time examining documents.¹³ On the other hand, note-taking leads to participants spending *less* time reading documents, and taking more notes.

In our work we also investigated the extent to which five findings (i.e., our hypotheses) from the education literature [37, 46, 155, 314], with regards to effect of highlighting and note-taking strategies on learning outcomes, hold up in a SAL context. We confirmed three of those hypotheses, and showed that while engaging in complex learning-oriented search tasks on the web, the acts of highlighting and note-taking themselves may not benefit learners. Rather, it is *how* these tools change the way the learners scan and processes text that is more important for learning while searching.

3

Recent Research Extension and Future Work Recent follow up studies have looked into how highlighting and note-taking, together with other strategies, can help in various user-sided aspects of the SAL process like self-regulated learning [282], goal setting [283, 284], etc. and system sided aspects like query suggestions [209], tools for sensemaking [264] etc. Further analysis of behavioural log data could provide insights into the document understanding process. For example, would recorded highlights and notes indicate more relevant/interesting sections of a given document, and if so, could retrieval algorithms be manipulated to promote documents that contain these ‘hotspots’? Findings could also eventually lead to the comparison of manual and automatic tools for active reading, and automatic thought externalisation. Extensions to this study could expand the work on examining how search behaviours can act as proxies for predictive measures of learning during search [40, 41, 84, 162, 233, 301]. Few work explored the impact of other widgets (e.g., entity cards [240], idea generation tool [59], cross session, cross device search assistance [100] etc.) for learning oriented search tasks. Advances in these directions could lead to the development of an adaptive search system. The observations from this work has design implications for search interfaces, where we must consider incorporating active reading tools within web search engines.

¹³These findings are reflected by *Search Economic Theory (SET)* [21] that indicates with similar time limits, as the number of queries issued drops, more documents will be examined (or longer will be spent on them).

4

MODELING USER INTERACTIONS AND WIDGET POSITIONING

Models developed to simulate user interactions with search interfaces typically do not consider the visual layout and presentation of a Search Engine Results Page (SERP). In particular, the position and size of interface widgets—such as entity cards and query suggestions—are usually considered a negligible constant. In contrast, in this work, we investigated the impact of widget positioning on user behaviour. To this end, we focussed on one specific widget: the Query History Widget (QHW). It allows users to see (and thus reflect) on their recently issued queries. We built a novel simulation model based on Search Economic Theory (SET) that considered how users behave when faced with such a widget by incorporating its positioning on the SERP. We derived five hypotheses from our model and experimentally validated them based on user interaction data gathered for an ad-hoc search task, ran across five different placements of the QHW on the SERP. We found partial support for three of the five hypotheses and observed that a widget’s location indeed has a significant impact on search behaviour.

This chapter is based on the following paper:

- ❑ Nirmal Roy, Arthur Câmara, David Maxwell, Claudia Hauff. 2022. Incorporating Widget Positioning in Interaction Models of Search Behaviour. In ICTIR. 53-62 [231].

4

4.1. INTRODUCTION

Economic theory, specifically *microeconomic theory*, assumes that an individual or *firm* will tend to maximise their profit—subject to budget or other constraints [288]. Microeconomic theory can also provide us with an intuitive means to model human-computer interactions [19]. Given a *demand* (that may arise from factors such as the nature of the context, the underlying task, or the system used), a *user* will exert *effort* to interact with the system by expending *internal resources* such as their working memory, attention, or energy. Users of a system will also incur a *cost* by expending *external resources* such as time, money, or physical effort (such as moving a mouse, or typing on a keyboard) [184]. In the context of IR, interactions between the user and system may lead to *benefits* in terms of information obtained, or resolved information needs [26, 28]. Rational users looking to maximise profit from their interactions can do so by either maximising their benefit or by minimising their expended cost and effort—and thus subscribe to the *Principle of Least Effort* [317].

Assuming that searchers behave in a rational way (a reasonable assumption to make [22]), we can model their interactions with a search engine to obtain insights into the different decisions made during the interaction process. In turn, these insights can help us provide explanations as to *why* users behave in a certain way. Importantly, such a model allows us to generate *testable hypotheses* as to how user behaviour will likely change when interface designs are modified based on a cost/benefit analysis of interface elements. For example, a study by Azzopardi et al. [22] found partial support for the hypothesis that, *as the cost and/or effort of issuing a query increases, users of a search system will issue fewer queries and examine more documents per query*.

Traditionally, **Information Seeking and Retrieval (ISR)** models [34, 36, 85, 120, 143, 302] provide post-hoc explanations as to *what* happens during episodes of information seeking. While these models are undoubtedly useful, they cannot predict: *we cannot employ them to learn what is likely to happen* in terms of user behaviour when changes are made to the retrieval system in question. This predictive power is necessary, for instance, in order to simulate the effects changes to the presentation of a SERP have on user behaviour, without having to run many costly user studies. Ultimately, the goal here is to only run user studies on interface designs that have shown promise from prior simulations.

In contrast to aforementioned models of ISR, our work follows a recent line of research that focuses on building mathematical models based on SET [19, 20] which is inspired by microeconomic theory—or *Information Foraging Theory (IFT)* [216, 217]. These models allow us to relate changing costs (e.g., the cost of querying, or the cost of examining a search result snippet) to changing search behaviours. Prior work in this area have focused on how users interact with a ranked list [57, 187], their stopping behaviours [180, 303], the trade-off between querying and assessing [19, 20, 22], and browsing costs [27, 133]. In these aforementioned studies, the SERP typically has a simple layout: the user can submit queries and assess documents. In addition, *interface components* (hereafter referred to as *wIDGETS*) such as *Related Searches* are typically considered to be placed at a fixed position, and their specific position is not part of the formal model definition. However, contemporary SERPs are complex, and widgets can appear at various positions on the SERP as shown anecdotally in Table 4.1: there is no consensus on positioning or size of the *Related Searches* widget across web search engines. In addition, contemporary SERPs contain direct answers (leading to *good abandonment* [166, 303]), advertisements, and information cards—as well

Table 4.1: The placement of (as well as the number of) text columns, and the number of entries in the *Related Searches* widget across ten different web search engines. Results retrieved on May 2nd, 2021 for the query chess. *Placement* corresponds to the widget’s position within the SERP.

Search Engine	Placement	#Columns	#Entries
bing.com	Bottom left	1	8
google.com	Bottom left	2	8
duckduckgo.com	Bottom left	1	8
yandex.com	-	-	0
ask.com	Upper right	1	12
yahoo.com	Bottom left	2	8
qwant.com	Upper right	1	8
baidu.com	Bottom left	3	9
ecosia.org	Bottom left	No columns	8
dogpile.com	Top left	1	8

4

as result lists that integrate content from a number of different search verticals.

In our work, we focus on an aspect of individual *widgets* on a SERP that—as already mentioned—has so far been neglected in mathematical representations of user interaction: the *positioning* of a given widget on a SERP. With this focus, we select one specific SERP widget to provide an initial exploration of how to incorporate widget positioning into a SET-based model. Concretely, we focus on the QHW , which is shown in Figure 2.1. It allows a user to view and thus reflect upon their recently issued queries during a search session. The widget is easy to understand for users, and involves only a small number of interactions—making it ideal as a first widget to employ for our exploration. Our main research question is therefore as follows.

RQ1 *How can we utilise the Search Economic Theory model of user interaction to refine the design hypothesis space for widget positioning?*

To answer this question, we first derive a SET-based model that considers a widget’s positioning as an input variable. Based on our formal model, we derive five hypotheses as to the search behaviour users are likely to exhibit as the widget’s positioning changes. Subsequently, in order to validate our model (and therefore also the inclusion of the positioning component in the model), we conduct a user study with $N = 120$ participants that each complete one ad-hoc retrieval task using a SERP with the QHW —in one of five different positions.¹ We observe empirical evidence that provides partial support for three of our five hypotheses which shows that: (i) a widget’s location influences search behaviour; and (ii) we are able to successfully create a formal interaction model, incorporating positioning, and mostly find evidence for our derived hypotheses.

¹An overview of the different widget positions we employ for our experiments is shown in Figure 2.1.

4.2. MICROECONOMIC THEORY AND IIR

Many models of ISR have been defined in the past [19, 34, 36, 85, 92, 120, 143, 217, 302]. They can generally be categorised into two groups: *descriptive* models [34, 36, 85, 120, 143, 302] and *formal* (mathematical) models [19, 92, 217]. The former provide us with intuitions and a holistic view of a user's search behaviour (e.g., with the *Berrypicking model* [34], users *pick* through information patches—analogous to people collecting berries). While they provide us with explanations of why searchers behave in a particular manner, they do not allow us to *predict* how a user's search behaviour will change in response to changes to the SERP, the quality of the results, etc. For this step, formal models such as SET [19, 20], IFT [216] or the Interactive Probability Ranking Principle (iPRP) [92] are required. With the increasing complexity of SERPs and the increasing amount of decisions users have to take during search episodes (and thus the ever growing number of experimental variants one would have to explore when exploring new interface variants), being able to rely on formal models to explore promising areas of the user interface design space is vital for cost-effective and efficient iterations of novel search interfaces.

A key assumption of the listed formal models (which are related to each other as shown by Azzopardi and Zuccon [25]) is that users will modify their search behaviour to achieve the greatest possible net benefit from an interaction which is defined as the difference between the benefit of interaction and the cost of interaction. Thus, modelling the cost and benefit of interactions taking place on typical SERPs—and subsequently validating the designed models through user studies (or conversely finding that the proposed model is not sufficiently fine-grained enough to predict user behaviour well)—has been the focus of recent work in this area.

Specifically, Azzopardi and Zuccon [26] created user-oriented cost-benefit models to analyse a number of user decisions (including the length of the submitted query, the specificity of the query, the use of query suggestions vs. query reformulation, etc.) that are made during a search session—and at what point those decisions lead to maximum user benefit. The authors focus on model creation; the developed models are not empirically validated. In a similar vein, Azzopardi and Zuccon [27] developed a cost model to determine—for various screen sizes—the number of search result snippets that should be visible on the SERP,

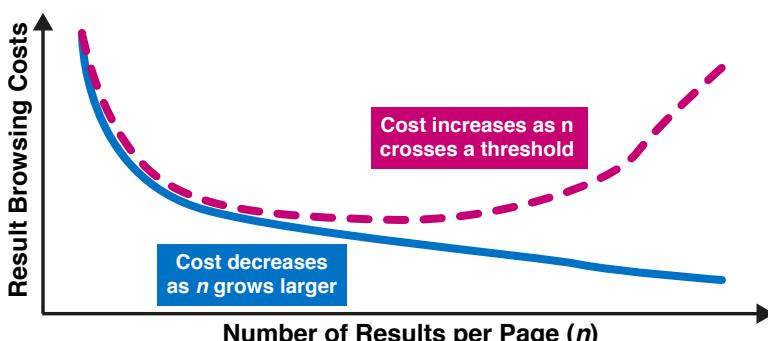


Figure 4.1: Example cost functions for two different SERP scenarios, adapted from Azzopardi and Zuccon [27]. Hypotheses can be derived from them, e.g., the optimal number of results (n per page) to show to maximise a user's benefit in the violet scenario is at the cost function's global minimum.

under the assumption that a user is looking for a document, and continues looking until that document is found. The developed formal model was initialised with hyperparameter values (such as the cost in seconds of typing out a character, or clicking a link) taken from the literature. Based on the developed cost functions (idealised examples of which are shown in Figure 4.1), several hypotheses were created—though their validation through a user study remained a point for future work. While this work already hinted at a distinction between desktop and mobile search (via the very different number of visible results in the viewport), Verma and Yilmaz [291] explicitly tackled this challenge and empirically determined (with 193 search sessions over $N = 25$ participants) to what extent existing user cost-benefit models are applicable (without change) to the mobile setting. The authors found that the parameters between desktop and mobile settings vary widely, and existing cost functions (with fixed hyperparameters, tuned to desktop search—and not adapted to the mobile setting) do not correlate very well with user satisfaction.

Using SET as their theoretical underpinning, Ong et al. [204] recently investigated the relationship between typing speed and search behaviour, both formally as well as empirically. While the authors did indeed observe a relationship between the two, they did find discrepancies between the observed user behaviours and those predicted by their model, conjecturing that their approximation of the model’s query cost (by typing speed) does not capture all important aspects of the query cost component. A similar methodology was used by Maxwell and Azzopardi [177]. Here, the authors derived five different hypotheses about how temporal delays (both query response delays and document download delays) affect search behaviour. These hypotheses were derived from SET- and IFT-based models, respectively. Empirically (with $N = 48$ participants), three of the five hypotheses on user behaviours held.

Prior work have successfully employed formal models to derive testable hypotheses of search behaviours. To the best of our knowledge, none of the prior work have however considered the *position* of a user interface widget as important enough to include in the derived model. In our work, we focus on this very issue: *how does the position of a search interface widget impact the search behaviour our model predicts, and to what extent do those predictions hold when examining interaction data derived from a user study?*

4.3. CONSIDERING WIDGET POSITIONING

In this section, we first discuss—at a high level—how to incorporate the positioning of a widget within an interface in a SET-based model. We then introduce our implementation of the QHW in more detail, and present the cost functions for our specific widget use case. We conclude this section with a number of hypotheses we derive from our mathematical model regarding the influence of the QHW position on a user’s search behaviour.

4.3.1. POSITIONING BASED ON FITTS’ LAW

One way to consider the positioning of widgets within an interface—in a microeconomic cost model of interaction—is to *estimate the time it will take for a user to find the widget on an interface/SERP* from a given *starting position*. One way to approximate this is by using *Fitts’ Law* [90]—an established, robust model of human psychomotor behaviour which has been frequently applied to computer and mobile interface design [132, 159, 266]. It states that

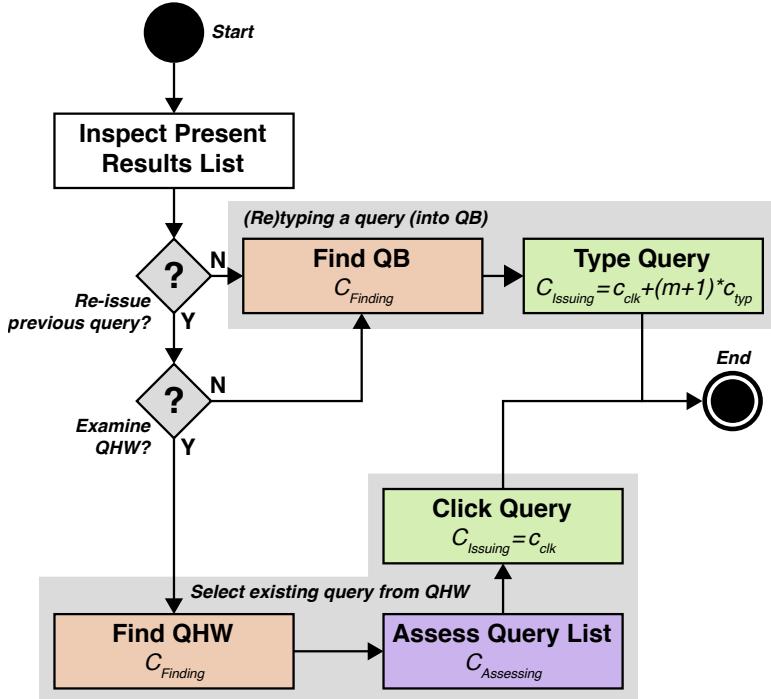


Figure 4.2: Flowchart of the modelled querying process. Before issuing each query, the user is presented with the choice of inspecting the QHW or typing in the query via the QB . Associated costs are outlined in §4.3.2.

the movement time for a user (moving their cursor on screen from a *source* to some *target*) is affected by the distance moved and the precision needed for such movement. The bigger and closer the target is, the easier it is to find and click. Shorter mouse movements are preferred, given that the object is large enough [266]. Therefore, given a search interface, the time taken to find a widget within a SERP is a function of its position and its size. In this work, Fitts' law is used as part of our SET-based user interaction model.

4.3.2. THE QUERY HISTORY WIDGET (QHW)

Let us now turn our attention to QHW , the interface widget that we developed the position-aware cost functions for. Shown in the callout in Figure 2.1, the QHW lists all previously issued queries in a search session. Our model considers the following scenario. A user, after inspecting a retrieved list of documents presented on a SERP, decide to issue a different query. Do they: (a) reissue an old query (i.e., a query submitted earlier in the *same* search session, perhaps because they wish to find a document from earlier); or (b) issue a new query (i.e., a query not yet submitted in the *same* search session, potentially leading to a new set of documents)? If the user decides to issue a new query, they will head to the *query box* (QB), and type the new query. If the user decides to reissue an old query, they must then decide whether to: (a) re-type the query in the QB ; or (b) scan the QHW , find the old query, and click it. A flowchart of the process described is shown in Figure 4.2.

In this chapter, we focus exclusively on the scenario where the user has *decided to reissue an old query*. We assume that the user knows they have issued this particular query in the past (a reasonable assumption, given that we only consider queries from a single search session), and expects to find it in the QHW . We develop a formal model in order to predict and understand the scenario where they will choose to re-type this old query in the QB , or when they will select it from the QHW instead—all *conditional on the position of the QHW on the SERP*. Note that this work does not focus on the reasoning behind re-issuing a query from earlier. We leave this for future work. Rather, we aim here to integrate positional information within a SET-based interaction model.

SPECIFYING COSTS

The total cost $C_{Reissuing}$ (in seconds) of re-issuing an old query (that consists of m characters, and is listed at position q inside the QHW) can be represented as three constituent components, as shown in the following equation.

$$C_{Reissuing} = C_{Finding} + C_{Assessing} + C_{Issuing} \quad (4.1)$$

$C_{Finding}$ is the cost of finding either the QHW or the QB . We approximate the cost in terms of time taken (in seconds). According to Fitts' Law [90], the movement time of the mouse cursor from some starting position on a display to some target (in this case, either the QHW or the QB) is equal to $a + b \log_2(\frac{D}{H} + 0.5)$ [55, 136], where D is the distance to the centre of the widget from the starting position of the cursor, H is the height of the widget (in 2D interfaces, the smallest value from the target's height or width is considered [170]), and a, b are constants that are empirically determined. Intuitively, the further the widget is from the starting position, the more time it will take for the users to find the widget.

$C_{Assessing}$ is the cost of assessing a widget. For QB , this cost is zero as users do not have to check a list of options. For QHW , it involves two actions: *scrolling and checking*. For example, consider that a user wants to find the q^{th} query (our *target query*) in the QHW . We associate a constant cost c_{scr} with scrolling over one query. Similarly, we associate a constant cost c_{chk} with checking whether a query is the *target query* or not. Given that the QHW displays t queries *above the fold* (e.g., in our experimental interface, as illustrated in Figure 2.1, we fixed $t = 4$), if $q \leq t$, then users do not incur any scrolling cost—and only the cost of checking to see if the query matches what they are seeking, or $q \times c_{chk}$. However, if $q > t$, users then have to scroll until the desired query is visible. This cost can be estimated by $(q - t) \times c_{scr} + q \times c_{chk}$, in line with [27].

$C_{Issuing}$ is the cost associated with entering the query. For QB , it is the cost of typing the query of length m ; this cost is $c_{clk} + (m + 1) \times c_{typ}$, where c_{clk} is the cost of clicking on the QB , c_{typ} is the cost of typing one character and $+1$ is included to account for the pressing of . For QHW , it is the cost c_{clk} of clicking on the desired query link.²

WHEN TO USE QHW

Based on the previous section, we can now write the cost functions $C_{Reissuing}^{QB}$ and $C_{Reissuing}^{QHW}$: re-issuing an old query by typing into QB , and by selecting a query from the QHW .

²While we have not yet described our implemented QHW widget in detail, we note that each old query is represented as a hyperlink; clicking a hyperlink reissues the query and displays the results for it on the SERP.

respectively. Based on our assumption of a rational user, we argue that a user will chose QB if the cost of using QB is less than the cost of using QHW . For completeness, we present both cost functions in Equations 4.2 and 4.3 below, as well as a short definition of the corresponding symbols. For simplicity and neatness, we suppress the subscript from $C_{Reissuing}^{QB}$ and $C_{Reissuing}^{QHW}$ for now on, referring these costs simply as C^{QB} and C^{QHW} , respectively. Rational users should choose QB over QHW if $C^{QB} < C^{QHW}$.

$$C^{QB} = a + b \log_2 \left(\frac{D_{QB}}{H_{QB}} + 0.5 \right) + c_{clk} + (m + 1) \cdot c_{typ} \quad (4.2)$$

D_{QB} = Distance of QB from starting position

H_{QB} = Height of query box (in pixels)

m = Query length (in characters)

c_{typ} = Cost of typing one character

$$C^{QHW} = \begin{cases} a + b \log_2 \left(\frac{D_{QHW}}{H_{QHW}} + 0.5 \right) + q \cdot c_{chk} + c_{clk}, & \text{if } q \leq t \\ a + b \log_2 \left(\frac{D_{QHW}}{H_{QHW}} + 0.5 \right) + (q - t)c_{scr} + q \cdot c_{chk} + c_{clk}, & \text{if } q > t \end{cases} \quad (4.3)$$

D_{QHW} = Distance of QHW from starting position

H_{QHW} = Height of QHW (in pixels)

q = Position of the target query in QHW

c_{chk} = Cost of checking a query in QHW

c_{scr} = Cost of scrolling over a query in QHW

c_{clk} = Cost of clicking a hyperlink in QHW

CONSTANTS

In our model, the above inequality depends not only on the positioning of QHW , but also on the value of a few constants. These are: the cost of clicking (c_{clk}); scrolling (c_{scr}); typing (c_{typ}); checking (c_{chk}) queries; the sizes of both QB (H_{QB}) and QHW (H_{QHW}); the number of queries above the fold (t); the distance from the bottom of the screen to QB (D_{QB}); and the considered starting point of the cursor. In order to derive meaningful hypotheses from our inequality and use the model to predict actual user behaviour, we need to provide meaningful estimates of these constants. We can either estimate them directly from the interaction logs we collect in our user study, or fix their values based on studies reported in the literature in line with [25, 27]. For example, the typical values of pertaining are shown in Table 4.2 where we take c_{typ} , c_{scr} , c_{clk} and the hyperparameter b from the literature. We note that a —defined in Equations 4.2 and 4.3—is cancelled in our comparison, and thus is ignored. In order to make use of the model, we also need to define certain other constants like distance of QB , or the height of QHW , etc.—which we also report in Table 4.2. We need these precise values to predict real world behavior by calculating the exact cost of each decision. We leave this as future work. In this chapter, we focus on using the general intuition behind the model equations to derive hypotheses of user interaction.

Table 4.2: Overview of the model's constants and values used.

Constant	Value
<i>Taken from the literature</i>	
c_{typ}	0.28 [22, 55]
c_{scr}	0.1 [28]
c_{clk}	0.2 [22, 55]
b	0.1 [55, 136]
<i>Defined for our experiments</i>	
D_{QB}	1000px
H_{QB}	50px
H_{QHW}	130px
t	4
c_{chk}	0.25
Cursor starting position	End of search result list at bottom of screen, ⑥ from Figure 2.1

4

4.4. HYPOTHESES

Having defined our model in Equations 4.2 and 4.3—along with all associated constants, we now derive five hypotheses pertaining to the query issuing behaviour that the model describes, and how position can influence search behaviours.

H1 *As the length of query q to be reissued increases, a user will be more likely to reissue the query via $QH\mathcal{W}$.*

This first hypothesis follows from Equation 4.2. As m (the length of query q in characters) increases, C^{QB} increases. At the same time, m does not influence $C^{QH\mathcal{W}}$.

H2 *If the number of queries to check in $QH\mathcal{W}$ increases, a user's likelihood of using $QH\mathcal{W}$ increases if its distance to the starting point decreases.*

In Equation 4.3 we see that, if q increases, D_{QHW} has to decrease to keep the overall cost of using $QH\mathcal{W}$ lower than that of QB .

H3 *The lower the distance of the $QH\mathcal{W}$ to the starting point, the more likely users will use it.*

This follows from Equation 4.3 where everything else being constant, the cost of reissuing a query is lowest when $D_{QHW} = 0$.

H4 *Users who type slower are more likely to use the $QH\mathcal{W}$ irrespective of where it is located.*

In §4.3.2, we provided fixed estimates for various constants in our model. One of those estimates is the cost of typing a character. Since the typing speed of users might vary to a considerable degree [204], the typing cost should be subject to further scrutiny. A user with slower typing will have a higher cost of typing queries, which will likely affect what widget they will use to reissue a query. For slow typing, c_{typ} is high, and C^{QB} becomes

higher than C^{QHW} for all reasonable values of D_{QHW} and q . Hence, with slow typing, the positioning of QHW is less crucial—or how many queries are present in it, as a user is more likely to use QHW anyway.

H5 A user's attention follows a F-shaped gaze pattern.

This pattern has been observed on heterogeneous SERPs in the past [82] and should be reflected in the amount of attention users pay to the QHW in different positions. Specifically, the interface with QHW in the top right corner of the screen is likely to receive more attention than QHW positioned at the bottom right corner. Similarly, QHW in the bottom left corner is likely to receive less attention than QHW in the top left part of the screen.

4

4.5. USER STUDY DESIGN

In order to examine whether there is support for our hypotheses, we conducted a between-subjects user study. Participants were presented with a SERP that was complemented with the QHW in different positions, depending on the condition they were assigned.

4.5.1. SYSTEM, CORPUS, TOPIC AND TASK

For our user study we employed SearchX [222], a modular, open-source search framework which provides quality control features for crowdsourcing experiments. We integrated the **LogUI** framework [182] into SearchX to allow us to accurately capture all keyboard events and mouse events (including hovers and clicks) over QB , QHW , and results.

SearchX was configured to use the *TREC AQUAINT* corpus. The corpus consists of over one million newspaper articles from the period 1996 – 2000. Articles were gathered from three newswires: the *Associated Press (AP)*, the *New York Times (NYT)*, and *Xinhua*. Using a traditional test collection provided us with the ability to easily evaluate the performance of participants where required. We index the collection using *Indri*, and use its own snippet generator for the summaries presented to participants. We employed Indri's Dirichlet prior smoothing model (with $\mu = 2500$).

We used the *wildlife extinction* topic (topic number 347) from the *TREC 2005 Robust Track* [292]. A total of 165 relevant documents were identified by TREC assessors for this topic within AQUAINT. This topic was selected as it has been successfully employed in prior user studies [22, 177]; the topic remains relevant to this day, and is likely to be of some interest to our participants.

We instructed our participants to identify documents that they perceived to be relevant to the TREC topic description that we provided to them. We primed our participants by asking them to imagine that they were to write an essay on the topic, and would use the identified documents as potential references at a later time.

4.5.2. INTERFACE AND INCENTIVES

Our search interface is presented in Figure 4.3. It contains: the standard query box QB (without autocomplete features) ①; a task timer and a bookmarked-documents counter ②; six search *results per page (RPP)* ③; functionality to mark documents in the form of a toggle icon ④; and QHW ⑤.

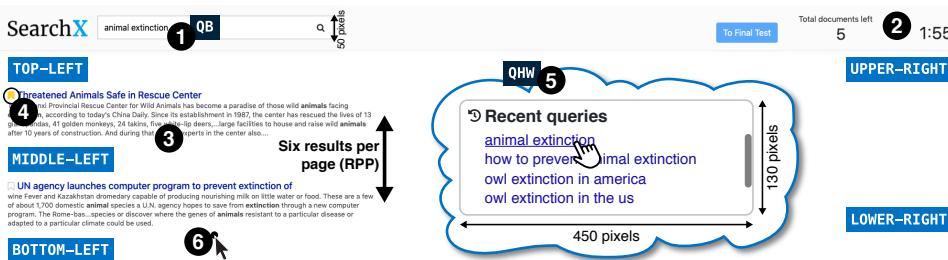


Figure 4.3: The SearchX search interface used for this study. Note the inclusion of the *QHW* in the callout—this was positioned in one of the areas as shown with blue boxes. Refer to §4.5.2 for information on the circled interface components.

4

As we were looking to incentivise participants to reissue existing queries, special considerations needed to be made to this effect—along with considering that the search interface used should be kept simple to avoid any undue attention given to components that were not considered by our model defined in §4.3.2. We evaluated our incentives in a small pilot study before deploying them to our study participants. Results from the pilot study are not included in our final analysis.

Participants were instructed that they could mark no more than *six* documents at a time. The marked documents counter helped participants to keep track of their number of marked documents. The idea behind this was that a strict limit on how many documents could be marked would incentivise participants when issuing queries later on in their search session (either via *QB* or *QHW*) to unmark previously marked documents (by toggling the icon)—and mark new ones that they perceived to be more promising. Participants were incentivised further by the potential for a bonus being awarded to the top six participants who achieved a high accuracy.

Before the study commenced, the participant's screen resolution was checked—a resolution check ensured that the resolution of the browser was 1920×1080 or greater. This resolution was found to show (with a high degree of certainty) that the entire search interface could fit on the participant's screen without the need for scrolling, meaning all six RPP were displayed, with none hidden *below the fold*. It also helped us to estimate the value of D_{QB} as presented in Table 4.2. There is also no pagination enabled on the SERP. These are due to the fact that our model does not include *page* scrolling or pagination, factors that could alter user behaviour.³ To this end, we also removed any hyperlinks to documents. To compensate, we increased the number of lines for each summary snippet from the established two to four. While longer snippets have been shown to increase confidence in decisions of relevance at the expense of accuracy [181], it was decided that additional surrogate text in this instance would help participants in judging documents without access to the full text.

³It does however consider scrolling costs *within QHW*.

4.5.3. OPERATIONALISING THE QHW

We operationalised the QHW as shown in the callout in Figure 4.3. The widget measures 450×130 pixels. At the top of the widget is the **Recent queries** title. Each query issued by the participant during the study is then prepended to the list shown in the lower portion of the widget. Queries are listed in reverse chronological order, with the most recently issued query appearing at the top.

As the QHW has a fixed width and height on the SERP, the widget could display at most four queries at a time, matching $t = 4$ as outlined in §4.3.2. Participants who wished to see more queries could scroll using their trackpad or mouse wheel to reveal older queries. All queries listed in the QHW were displayed as the standard blue hyperlink text—which underlines when hovered over—to provide a *proximal cue* [61] that they were hyperlinks that could be clicked. A click on the listed query then submits the listed query to the search engine, and displays the top six ranked documents.

4

In terms of positioning within the SERP, we trialled five different positions which are demonstrated in Figure 4.3 with blue boxes. Anecdotal evidence as presented in Table 4.1 suggests that there is no clearly defined position for widgets on a SERP (beyond the search results and entity cards), and thus we evaluated the major positions. Each of our five QHW positions (three on the left rail and two on the right) are represented in our user study as a unique condition.

TOP-LEFT Positioned at the top left, before the first result.

MIDDLE-LEFT Positioned on the left rail, below the third result.

BOTTOM-LEFT Positioned at the bottom left, immediately after the sixth and final result.

UPPER-RIGHT On the right rail, this condition positioned QHW underneath the clock; it is top-aligned with the first result. This position would be analogous to where an entity card sits on a contemporary web search engine’s SERP.

LOWER-RIGHT On the right rail, this condition positioned QHW under the clock; it is aligned at the bottom with the last result.

4.5.4. POST-TASK SURVEY

Inspired by the *User Experience Questionnaire (UEQ-S)* [114, 149], we asked participants five questions after the completion of the search task which. Questions explored the usage experience of the QHW . All questions were answered using a 7-point Likert scale, considering negative to positive responses. For example, to understand to what extent a widget positioning was unexpected for the participant, we ask “*What did you think about the position of the query history widget?*”, with the scale ranging from *unexpected* (1) to *expected* (7). Additionally, we ask about the support, ease of use, efficiency & clarity of the widget. Participants also received an open question for general comments and feedback about the interface.

4.5.5. CROWDSOURCED PARTICIPANTS

Participants for our study were recruited from *Prolific*, a crowdsourcing platform which has been shown to be an effective choice for complex and time-consuming *Interactive*

Information Retrieval (IIR) experiments [308]. In order to obtain high-quality and reliable data, we imposed the following constraints: (i) participants needed to have at least 100 prior Prolific submissions; (ii) have an approval rate of 95% or higher; and (iii) have native proficiency in English. The complete study took approximately fifteen minutes, which included the minimum search time of 10 minutes. For their time, participants were compensated at the rate of GBP£8.00 per hour.

Overall, a total of 125 participants took part in our study. From this total, we had to reject five as they did not comply with our quality checks.⁴ Our final cohort of 120 participants included 40 female and 80 males ones, with a reported average age of 35 years (youngest 18; oldest 77).

4.6. RESULTS

We now discuss the empirical validation of each of our five hypotheses which were introduced in §4.4. Recall that our research question asks whether widget positioning information can be meaningfully incorporated in a SET-based model.

A comparison of the main search behaviour indicators across conditions is shown in Table 4.3. On average, participants issued 12 queries (28 characters long)—and marked six documents, hitting the imposed limit). 114 participants reissued 5 queries on average, while six did not reissue any queries (either via QB or QHW). On average our participants spent 12 minutes on the search task. We collected, on average, 2148 log events per participant.

Additionally, we also measured how the participants behaved regarding marking documents. On average, participants marked 2.60 relevant documents during their session, and 5.10 non-relevant documents. As expected, participants also unmarked documents over their session, indicating that they were actually reflecting on what they had marked. On average, participants unmarked 1.70 documents, where 1.15 of these were non-relevant.

The results of our post-task survey indicate that our interface was easy to use (Table 4.3, row XV: on average a score above 5 on a 7-point Likert scale), and the purpose of the QHW was clear (row XVII: on average a score above 5). Apart from **MIDDLE-LEFT**, which received a comparably low and significantly worse *expected position* score than almost all other variants (the only exception being **LOWER-RIGHT**), the QHW variants were positioned at somewhat expected locations (Table 4.3, row XIII: on average a score above 4 on the 7-point Likert scale).

These numbers indicate that our task design (which encouraged the reissuing of queries) was successful. Finally, we point the reader to Figure 4.3 for examples of actual queries our participants submitted (as visible in the QHW callout). We also considered ⑥ from Figure 4.3 as the expected starting point where the cursor is positioned after they have scanned the search results. From this location, they move the cursor to QHW or QB to (re)issue queries. We argue that it is reasonable to expect individuals to examine all six results on the SERP before moving on. Coupled with the known correlation of eye gaze and cursor positioning on the screen [60], this assumption allows us to make estimations of D_{QB} and D_{QHW} in Equations 4.2 and 4.3.

4

⁴Our quality checks required that participants did not change the browser tab more than three times during the study, issued at least two queries, and marked at least two documents during their search session.

Table 4.3: Mean (\pm standard deviation) of search behaviour metrics across all participants in each variation of QHW . A dagger (\dagger) denotes one-way ANOVA significance, \S denotes χ^2 significance, while U,L , B,M , T indicate post-hoc significance ($p < 0.05$ with Bonferroni correction) over conditions **UPPER-RIGHT**, **LOWER-RIGHT**, **BOTTOM-LEFT**, **MIDDLE-LEFT** and **TOP-LEFT** respectively.

Measure		UPPER-RIGHT	LOWER-RIGHT	BOTTOM-LEFT	MIDDLE-LEFT	TOP-LEFT
I	Number of participants	24	24	25	24	23
II	Rank in terms of distance from ⑥ (D_{QHW} , pixels)	5(980)	3(610)	1(190)	2(515)	4(780)
<i>Search Log Statistics</i>						
III	Number of queries via QB^\dagger	$10.29(\pm 5.19)^B$	$11.67(\pm 5.04)$	$13.92(\pm 6.05)^U$	$12.29(\pm 5.18)$	$12.61(\pm 5.26)$
IV	Number of queries re-issued via QHW	$3.50(\pm 4.58)$	$3.79(\pm 5.16)$	$5.28(\pm 7.26)$	$5.21(\pm 6.67)$	$3.78(\pm 4.21)$
V	Number of queries re-issued via QB	$0.62(\pm 1.17)$	$0.67(\pm 1.01)$	$0.60(\pm 1.50)$	$0.54(\pm 0.78)$	$0.65(\pm 1.03)$
VI	Number of unique queries via QB	$9.67(\pm 4.72)^B$	$11.00(\pm 5.06)$	$13.32(\pm 5.84)^U$	$11.75(\pm 4.79)$	$11.96(\pm 5.43)$
VII	Number of hovers in QHW (500ms threshold) [†]	$5.50(\pm 4.61)^{M,T}$	$5.25(\pm 6.07)^T$	$6.72(\pm 7.42)$	$9.50(\pm 8.45)^U$	$11.74(\pm 6.09)^{U,L}$
VIII	Number of scroll events on QHW^\dagger	$2.75(\pm 4.11)^{B,M}$	$3.83(\pm 6.15)$	$7.52(\pm 10.03)^U$	$7.42(\pm 9.22)^U$	$6.13(\pm 7.82)$
IX	Frac. of queries re-issued via QHW , slow typing (57 users)	$0.95(\pm 0.12)$	$0.96(\pm 0.07)$	$0.93(\pm 0.12)$	$0.97(\pm 0.03)$	$0.90(\pm 0.18)$
X	Frac. of queries re-issued via QHW , fast typing (57 users)	$0.78(\pm 0.37)$	$0.88(\pm 0.15)$	$0.90(\pm 0.21)$	$0.80(\pm 0.31)$	$0.75(\pm 0.38)$
XI	Frac. queries re-issued below QHW fold [§]	$0.846^{B,M}$	0.911	$0.979^{U,L}$	0.957^U	0.892
XII	Frac. queries re-issued above QHW fold	0.849	0.758	0.882	0.852	0.835
<i>Post-Task Questionnaire</i>						
XIII	Expected Position, I : <i>unexpected</i> , 7 : <i>expected</i> [†]	$4.62(\pm 1.24)^M$	$4.22(\pm 1.88)$	$4.64(\pm 1.58)^M$	$3.17(\pm 2.12)^{U,B,T}$	$4.96(\pm 1.33)^M$
XIV	Task Support, I : <i>obstructive</i> , 7 : <i>supportive</i>	$5.25(\pm 1.39)$	$5.30(\pm 1.46)$	$5.40(\pm 1.26)$	$4.33(\pm 2.01)$	$4.91(\pm 1.76)$
XV	Ease of use, I : <i>complicated</i> , 7 : <i>easy</i>	$6.33(\pm 1.20)$	$6.17(\pm 1.23)$	$5.88(\pm 0.93)$	$5.42(\pm 1.61)$	$6.04(\pm 1.30)$
XVI	Help in task goal, I : <i>inefficient</i> , 7 : <i>efficient</i> [†]	$5.58(\pm 1.50)^M$	$5.04(\pm 1.55)^M$	$4.88(\pm 1.59)$	$3.83(\pm 2.33)^{U,L,T}$	$5.48(\pm 1.41)^M$
XVII	Purpose of widget, I : <i>confusing</i> , 7 : <i>clear</i>	$5.21(\pm 2.28)$	$6.17(\pm 1.50)$	$6.20(\pm 1.32)$	$5.33(\pm 1.74)$	$5.26(\pm 1.91)$

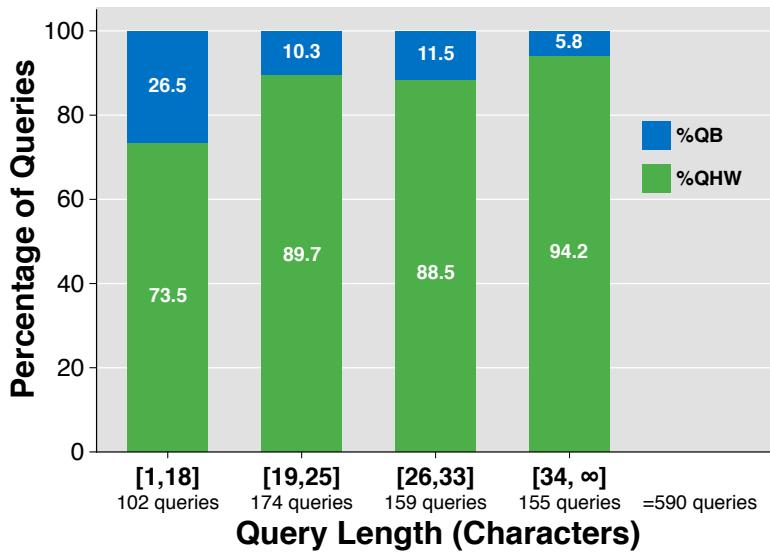


Figure 4.4: Overview of $\mathcal{QH}\mathcal{W}$ vs. \mathcal{QB} usage when reissuing queries of varying lengths. Shown here are the results over 590 reissued queries across all participants/conditions.

4

4.6.1. H1: QUERY LENGTH

Hypothesis **H1** states that as the length of some query to reissue increases, the likelihood of reissuing the query via $\mathcal{QH}\mathcal{W}$ —independent of the widget’s position—increases. To investigate **H1**, we consider all 590 reissued queries across all participants and conditions. Queries were partitioned into four groups (with boundaries at the 25th/50th/75th percentiles), according to their length m in characters—[1, 18], [19, 25], [26, 33] and [34, ∞]—and determined the fraction of queries reissued via \mathcal{QB} and $\mathcal{QH}\mathcal{W}$. Results are shown in Figure 4.4. We find that for the shortest reissued queries (with $m < 19$), 74% of queries are reissued via $\mathcal{QH}\mathcal{W}$, while this percentage rises to 94% for the longest ($m > 33$) queries. This trend provides support for **H1**: participants prefer to use $\mathcal{QH}\mathcal{W}$ for reissuing queries, and do so with a greater likelihood as query length increases. In order to observe if the trend is significant, we sampled one random reissued query from each participant to make the observations in the four query groups independent. A Chi-square test revealed that there is significant difference across the four query groups ($\chi^2(3, N = 114) = 10.58, p = 0.01$). Post-hoc tests showed that there were significant differences in the number of queries issued via $\mathcal{QH}\mathcal{W}$ between queries belonging to the 3 larger size groups when compared to the group representing smaller queries. However, there was no significant difference among the three groups representing larger queries. We therefore find partial support to our hypothesis that people are more likely to use the $\mathcal{QH}\mathcal{W}$ to reissue queries as query length increases.

4.6.2. H2: QUERY POSITIONING IN THE QHW

H2 centres around the number of queries in $\mathcal{QH}\mathcal{W}$. The hypothesis states that as the number of issued queries increases, the likelihood of a participant using $\mathcal{QH}\mathcal{W}$ increases

as the widget's distance to the starting point (⑥ in Figure 4.3) decreases. We rank the five QHW positions (conditions) we explore empirically according to their distance from the point on the screen where we expect the participant's cursor to be after they have scanned all six results. This information is shown in row **II** of Table 4.3. The positions are ranked as follows: (1) **BOTTOM-LEFT**; (2) **MIDDLE-LEFT**; (3) **LOWER-RIGHT**; (4) **TOP-LEFT**; and (5) **UPPER-RIGHT**. From our interaction logs, we also calculated the position in the QHW for a reissued query. As discussed in §4.5.3, queries are displayed in reverse chronological order, which means if users have to scan further down the list, they are looking for an older query to reissue. We collected the ranks of all reissued queries (590 queries in total), and divided them into two groups based on how many queries are displayed *above the fold* in QHW —reissued queries with a position of four or less (340 queries) and those with a rank greater than four (250 queries). Table 4.3, row **XI** shows that on average, participants are more likely to reissue a query when it is present *below the fold* from the conditions where QHW was positioned closer to the starting position (as observed in row **II**). Moreover, when participants want to reissue a recent query (displayed above the fold in QHW), they are on average less likely to reissue it from the respective QHW conditions compared to when participants want to reissue an older query (displayed below the fold), as shown in Table 4.3, rows **XI** and **XII**. The only exception is the farthest variant **UPPER-RIGHT**, where the likelihood of reissuing is similar for both recent and older queries. To observe if this trend showing evidence for our hypothesis is significant, we conducted a Chi-squared test following a similar approach to **H1**. We sample two reissued queries (one above and one below the fold of QHW) from each participant, and observe that for queries lower down the list (below the fold) in QHW (row **XI**), there is a significant difference for the fraction of time it was reissued via QHW across conditions ($\chi^2(4, N = 89) = 9.14, p = 0.02$). Post-hoc tests revealed significant differences between **BOTTOM-LEFT** and **MIDDLE-LEFT** in comparison to the **UPPER-RIGHT** condition. There was no significant difference when the query was present above the fold in QHW ($\chi^2(4, N = 114) = 3.86, p = 0.42$). Since we do not find a significant difference across all variants for queries reissued via QHW below the fold, we can only claim our hypothesis has been partially supported.

4.6.3. H3: DISTANCE OF THE QHW

H3 states that with decreasing distance of the QHW to the starting point, the higher its usage. As mentioned in **H2**, row **II** of Table 4.3 shows the distance of each condition of QHW from where we expect a participant's mouse cursor to be after scanning results (⑥ from Figure 4.3). We observe from row **IV** of Table 4.3 that there is no significant difference ($F(4, 115) = 0.544, p = 0.7$) in the number of reissued queries via QHW between the conditions. We do however observe a trend: participants in the two conditions closest to the starting point (**BOTTOM-LEFT** and **MIDDLE-LEFT**) issued on average more than five queries via QHW ; in the other three conditions, participants issued on average fewer than four queries via QHW . Finally, our QHW widget has the intended effect: reissued queries are far more likely to come via QHW than QB whose usage for reissued queries is shown in row **V** of Table 4.3. On average, fewer than one reissued query per participant is submitted via QB . Based on these results we cannot argue in favour of **H3**, even though the data trend is aligned with our hypothesis, the differences are not significant.

4.6.4. H4: SLOW TYPING

H4 states that participants who type slower incur a higher typing cost, and are likely to prefer to use QHW irrespective of its position on screen. From our interaction logs, we computed the mean typing speed of our participants (we considered 114 users who reissued at least one previous query) by averaging the total time they took to type a query by the query length. We then performed a median split (0.323 seconds per character) of our participants based on their mean typing speed, and categorised them as *Slow* and *Fast*.

In rows **IX** and **X** of Table 4.3, we report the fraction of times reissued queries were issued via QHW by *Slow* and *Fast* participants, respectively. Across all conditions, we find, on average, *Slow* participants relied on QHW more often than those in *Fast*. For example, in the **UPPER-RIGHT** condition, 95% of reissued queries on average were submitted via QHW over *Slow*; the value was 78% for *Fast*. The smallest difference in behaviour is observed for the **BOTTOM-LEFT** condition: here, *Slow* reissue on average 93% of queries via QHW , while *Fast* reissue 90% via QHW . In addition, we find *Fast* to exhibit more diverse behaviour than *Slow* as indicated by the standard deviations reported in rows **IX** and **X** of Table 4.3. This shows that participants in *Slow* rely on QHW more consistently than *Fast*. There is no significant difference for the fraction of time a query was reissued via QHW by *Slow* (row **IX**). Although this finding is in line with our hypothesis, we do not see any significant difference over *Fast* (**X**). We thus cannot claim to have support for **H4**.

4

4.6.5. H5: F-SHAPED GAZE PATTERN

H5 is not derived from our formal model, but instead based on prior work that have found users to pay attention to SERPs in a particular manner: the top-left part of the SERP receives the most attention, and attention decreases as one goes down the SERP on the left rail, and to the right rail. We hypothesise here that we can find a similar attention pattern for the different positions of QHW .

Contrasting to Dumais et al. [82] where gaze patterns were recorded via eye trackers, we did not perform webcam-based eye tracking and thus have to rely on other interaction logs to estimate attention. As found by Rodden et al. [227], eye gaze and mouse movements are correlated. We thus approximate how much attention participants were paying to QHW variants via the mean number of hover and scrolling events over QHW based on our interaction data.⁵ We only consider hover events that spanned at least 500ms. We make this choice as variants **MIDDLE-LEFT** and **TOP-LEFT**, due to their location, would fall ‘in the way’ of participants performing other tasks, like marking a document, or moving to QB —considering all hover events would have skewed the interaction data. We observe significant differences across QHW variants for these two hover-based ($F(4, 115) = 41.4, p = 0.003$) and scroll events ($F(4, 115) = 39.6, p = 0.01$) which are reported in Table 4.3 (rows **VII** and **VIII**). Post-hoc tests revealed that the **UPPER-RIGHT** condition receives significantly fewer hovers or scrolls (and thus less attention) than **MIDDLE-LEFT**, **TOP-LEFT** and **BOTTOM-LEFT**. Thus, attention decreases as we move to the right. In contrast, we do not confirm our hypothesis that attention decreases as users move down the screen: the **TOP-LEFT** and **BOTTOM-LEFT** conditions do not significantly differ in terms of our hover/scroll measures.

⁵We are aware of more advanced approaches to estimate gaze patterns from mouse movements, e.g., [103]—but leave this exploration for future work.

Overall, we have partial evidence for hypothesis **H5**: participants pay less attention to the right side of the screen as approximated by our hover/scroll measures, but this attention decrease is not observed as we move towards the bottom of the screen. Of course, it should be noted that we designed the interface in such a way that participants were able to see the entire SERP at once without the need to scroll and see below the fold.

4.7. CONCLUSIONS

In this work, we set out to answer the question of *how to incorporate interface positioning information in a SET-based model*. To this end, we derived a position-aware interaction model of search behaviour. We focused on the QHW , and formulated a model that can predict search behaviour related to the *reissuing* of queries from the same search session. We used Fitts' Law [90] to approximate the cost of finding the widget based on its five different positions on the screen. Based on our model and prior work, we developed five testable hypotheses. We conducted a between-subjects user study with $N = 120$ participants. We evaluated the impact of the *position* of QHW on search behaviour.

4

Of our five hypotheses, we found partial support for three.

H1 As the length of the to-be-reissued query increases, a user will be more likely to reissue the query via the QHW .

H2 If the number of queries to check in the QHW increases, the likelihood of users using the QHW increases as its distance to the starting point decreases.

H5 A user's attention span follows a F-shaped gaze pattern.

For the remaining two hypotheses—considering the relationship of the distance of QHW to the starting point and the widget's usage (**H3**), as well as the impact of typing speed on QHW 's usage (**H4**)—we observed trends aligned with our hypotheses. However, those trends were not statistically significant. Our empirical study therefore did *not* provide support for them.

Overall we argue that we successfully developed a position-aware interaction model of search behaviour. We did find that widget positioning plays a role and changes a user's search behaviour, and thus *position matters*—and should be incorporated into formal interaction models. Our model is purposefully simple and does not capture every possible facet of user interaction with a SERP and its widgets. Several additions and modifications can be made.

Generalisation We focused on a simplified use case of a single widget (the QHW). However, a modern search interface often contains multiple complex widgets simultaneously. Therefore, we aim to extend our work by creating user interaction models for more complex decisions pertaining to other widgets.

Cognitive effort Currently, our model ignores the cognitive cost of typing a query or looking for a query from a list present in the QHW . Modelling cognitive costs is not trivial and depends on, amongst other factors, the search phase the user is currently in, a user's prior knowledge—and task difficulty.

Layout and Graphics We have assumed that *only* the location of the QHW and QB impact the cost of finding these widgets. However, it has been shown that during web navigation, there is a difference between ease of finding a graphical widget (e.g., a shopping basket in an e-commerce website or a search box) versus one that is textual (e.g., various text hyperlinks in navigation menus) [115, 130]. The graphical properties of these widgets, like size, shape, colour, and highlights, can also impact the efficiency in finding links and widgets [115, 131, 212, 270]. They likely provide certain *cues* [61] that users latch onto.

Usability and Aesthetics Based on our current model, a widget that takes 90% of the SERP would be straightforward for a user to find. However, it would also make the whole user experience unpleasant at best. Therefore, modelling user interactions with multiple widgets could help us strike a balance to optimise the complete user experience. Prior work focusing on developing aesthetic measures (i.e., based on colour and symmetry) for widgets can also help develop a more nuanced model [195, 316].

Input Devices Our model assumes that the user interacts with a search engine in a standard browser, using a mouse and keyboard. However, this is not always the case. Extending our model to other types of interfaces like mobile and voice search (building on existing work [139]) is another interesting research venue to explore in future work.

5

VOICE MODALITY AND RELEVANCE JUDGMENT

*The creation of relevance assessments by human assessors (often nowadays crowdworkers) is a vital step when building IR test collections. Prior work has investigated assessor quality and behaviour, as well as tooling to support assessors in their tasks. We have a few insights, though, into the impact of a document’s presentation modality on assessor efficiency and effectiveness. Given the rise of voice-based interfaces, we investigated whether it is feasible for assessors to judge the relevance of text documents via a voice-based interface. We ran a user study ($n = 49$) on a crowdsourcing platform where participants judged the relevance of short and long documents—sampled from the TREC Deep Learning corpus—presented to them either in the *text* or *voice* modality. We found that: (i) participants were equally accurate in their judgements across both the *text* and *voice* modality; (ii) with increased document length, it took participants significantly longer (for documents of length > 120 words, it took almost twice as much time) to make relevance judgements in the *voice* condition; and (iii) the ability of assessors to ignore stimuli that are not relevant (i.e., inhibition) impacted the assessment quality in the *voice* modality—assessors with higher inhibition were significantly more accurate than those with lower inhibition. Our results indicate that we can reliably leverage the *voice* modality as a means to effectively collect relevance labels from crowdworkers.*

This chapter is based on the following paper:

- ✉ Nirmal Roy, Agathe Balayn, David Maxwell, Claudia Hauff. 2023. Hear Me Out: A Study on the Use of the Voice Modality for Crowdsourced Relevance Assessments. In SIGIR. 718-728 [230].

5.1. INTRODUCTION

Document relevance assessments by human assessors—with respect to a given set of *information needs*—is a vital step in the building of an IR test collection [125, 254]. Depending on the corpus, documents are represented in a variety of forms—including text (the most common form at TREC), images [75, 190], or videos [98, 167]. Prior work have investigated assessor quality, their behaviour, and tooling to support assessors—most often in the context of text documents [11, 142, 228, 245, 246]. Given the prevalent nature of text corpora, we continue in this vein and focus on an aspect that has received little attention so far: the *presentation modality* of the text documents during the judging process.

Thanks to the development of voice-based conversational search systems, people have become accustomed to being presented search results that are read out to them, an approach that is very different from the presentation of text on-screen. We posit that by utilising such audio-based devices, we can increase the scope for collecting relevance judgements for text documents in a number of ways. For example, assessors can contribute by judging documents on their smartphones [7, 287], if they have visual impairments [224, 290, 319], or if they come from a low-resource background [9, 224].

5

Two important aspects of collecting relevance judgements are: (i) the quality of assessments [245]; and (ii) the time taken by assessors to make their judgements [257]. Since relevance judgements are used to train and evaluate **Learning to Rank (Ltr)** systems, the quality of judgements impacts the effectiveness of such systems [66, 307]. The time taken by assessors to judge relevance may not only affect the quality of judgements, but also contribute to the cost of building (and maintaining) test collections. NIST assessors [71, 72] and crowdworkers [8, 144] are often paid by their time spent on a task (e.g., as on *Prolific*). The longer it takes assessors to judge, the costlier it becomes. There are a number of factors—not limited to topic difficulty [74, 245], document familiarity [246], or relevance judgement session length [246]—that have been shown to affect the quality of (and the time taken for) judging relevance.

In our work, we focus on two such factors in our pursuit to examine the feasibility of using the voice modality for text-document relevance assessments: *document length* [91, 241, 246, 255] and an assessor’s *cognitive abilities* [242, 245] expressed in terms of working memory and inhibition. Our selection of factors is motivated by a range of prior work. The serial [146] and temporal [243] nature of the voice medium makes it more difficult for listeners to “skim” back and forth over a piece of information as compared to reading it on-screen [191, 306, 311]. Voice interfaces also demand greater cognitive load when compared to text interfaces for processing information [146, 253, 281]. These are exacerbated as the amount of information to be conveyed increases in size [200, 251]. Understanding how these factors affect the relevance judgement process can help us design tasks for assessors with a wide range of abilities and for different document presentation modalities. While there exists various measures for cognitive abilities, we selected two—*working memory* (someone’s ability to hold information in short-term memory) [78] and *inhibition* (someone’s ability to ignore or inhibit attention to stimuli that are not relevant) [78]—which have been shown to play an important role in speech understanding [101, 236, 261]. We posit that they will also be crucial in the relevance judgement process, especially when documents are presented in the voice modality. Taken together, we investigate the following research questions.

RQ1 How does the modality of document presentation (text vs. voice) affect an assessor’s relevance judgement in terms of accuracy, time taken, and perceived workload?

RQ2 How does the length of documents affect assessors’ ability to judge relevance? Specifically, we look into the main effect of document length and the effect of its interplay with presentation modality.

RQ3 How do the cognitive abilities of an assessor (with respect to their working memory and inhibition) affect their ability to judge relevance? Specifically, we look into the main effect of the cognitive abilities of assessors and the effect of its interplay with the presentation modality.

To answer these questions, we conduct a quantitative user study ($n = 49$) on the crowdsourcing platform Prolific. Participants judged the relevance of 40 short and long documents sampled from the passage retrieval task data of the 2019 & 2020 TREC Deep Learning (DL) track [71, 72]. Our findings are summarised as follows.

- Participants judging documents presented in the voice modality are *equally* accurate as those judging them in the text modality.
- As documents got longer, participants judging documents in voice modality takes significantly longer than those in text modality. For documents of length greater than 120 words, the former took twice as much time with less reliable judgements.
- We also find that inhibition—or a participant’s individual ability to ignore or inhibit attention to stimuli that are not relevant—impacts relevance judgements in voice modality. Indeed, those with higher inhibition are significantly more accurate than their lower inhibition counterparts.

Overall, our results indicate that we *can* leverage the voice modality to effectively collect relevance labels from crowdworkers.

5

5.2. RELATED WORK

5.2.1. RELEVANCE JUDGEMENT COLLECTION

The general approach for gathering relevance assessments for large document corpora (large enough that a full judgement of all corpus documents is not possible) was established by TREC in the early 1990s [107]. Given a set of information needs, a pooled set of documents based on the top- k results of (ideally) a wide range of retrieval runs are assessed by topic experts. This method is typically costly and does not scale up [8] once the number of information needs or k increases. In the last decade, creating test collections using crowdsourcing via platforms like Prolific or *Amazon Mechanical Turk* (AMT) have been shown to be a less costly yet reliable alternative [8, 144, 246, 318]. While the potential of crowdsourcing for more efficient relevance assessment has been acknowledged, concerns have been raised regarding its quality—as workers might be too inexperienced, lack the necessary topical expertise, or be paid an insufficient salary. In turn, these issues may lead them to completing the tasks to a low standard [134, 172, 213]. Aggregation methods (e.g., majority voting) can be used as effective countermeasures to improve the reliability of judgements [117, 126].

There are a number of factors that have been shown to affect the relevance judgement process. Scholer et al. [245] observed that participants exposed to non-relevant documents at the start of a judgement session assigned higher overall relevance scores to documents than when compared to those exposed to relevant documents. Damessie et al. [74] found that for easier topics, assessors processed documents more quickly, and spent less time overall. Document length was also shown to be an important factor for judgement reliability. Hagerty [106] found that the precision and recall of abstracts judged increased as the abstract lengths increased (30, 60, and 300 words). In a similar vein, Singhal et al. [255] observed that the likelihood of a document being judged relevant by an assessor increased with the document length. Chandar et al. [58] found that shorter documents that are easier to understand provoked higher disagreement, and that there was a weak relationship between document length and disagreement between the assessor. In terms of time spent for relevance judgement, Konstan et al. [141] and Shinoda [252] asserted that there is no significant correlation between time and document length. On the other hand, Smucker et al. [258] found participants took more time to read, as document length increased (from ~10s for 100 words, to ~25s for 1000 words).

5

5.2.2. VOICE MODALITY

Voice-based crowdsourcing has been shown to be more accessible for people with visual impairments [290, 319], or those from low resource backgrounds [224]. It can also provide greater flexibility to crowdworkers by allowing them to work in brief sessions, enabling multitasking, reducing effort required to initiate tasks, and being reliable [113, 289]. However, information processing via voice is inherently different compared to when it is presented as text. The use of voice has been often shown to lead to a higher cognitive load [192, 293]. Individuals also exhibit different preferences. For example, Trippas et al. [279] observed that participants preferred longer summaries for text presentation. For voice however, shortened summaries were preferred when the queries were single-faceted. Although their study did not measure the accuracy of judgements against a ground truth, what participants considered the most relevant was similar across both conditions (text vs. voice presentation). Furthermore, the voice modality can leverage its own unique characteristics for information presentation. For instance, Chuklin et al. [64] varied the prosody features (pauses, speech rate, pitch) of sentences containing answers to factoid questions. They found that emphasising the answer phrase with a lower speaking rate and higher pitch increased the perceived level of information conveyed.

Concerning the collection of relevance assessments, Tombros and Crestani [276] found in their lab study that participants were more accurate and faster in judging relevance when the list of documents (with respect to a query) were presented as text on screen as compared to when they were read out to the participants—either in person, or via telephone. It should however be noted that this work was conducted more than two decades ago—barely ten years after the invention of the Web, when the now common voice assistants and voice-enabled devices were long to be developed.

The work closest to ours is the study by Vtyurina et al. [293], who presented crowdworkers with five results of different ranks from *Google*—either in text or voice modality. They asked their participants to select the two most useful results and the least useful one. The relevance judgements of participants in the text condition were observed to be

significantly more consistent with the true ranking of the results than those who were presented with five audio snippets. The ability to identify the most relevant result was however *not* different between the two cohorts. This study did not consider the effect of document length or cognitive abilities of participants on their relevance judgement performance, which is what we explore.

5.2.3. COGNITIVE ABILITIES

Prior work have explored how the cognitive abilities of assessors impact relevance judgements. Davidson [76] observed that openness to information—measured by a number of cognitive style variables such as open-mindedness, rigidity, and locus of control—accounted for approximately 30% of the variance in relevance assessments. Scholer et al. [245] found that assessors with a higher need for cognition (i.e., a predisposition to enjoy cognitively demanding activities) had higher agreement with *expert* assessors, and took longer to judge compared to their lower need for cognition counterparts. Our work focuses on *working memory* and *inhibition*.

Working Memory (WM) refers to *an individual's capacity for keeping information in short-term memory even when it is no longer perceptually present* [78]. This ability plays a role in higher-level tasks, such as reading comprehension [168] and problem solving [299]. MacFarlane et al. [169] observed that participants with dyslexia—a learning disorder characterised by low working memory—judged fewer text documents as non-relevant when compared to participants without the learning disorder. They posited that it might be cognitively more demanding to identify text documents as non-relevant for the cohort with dyslexia. With regards to processing speech, High WM has also been shown to be helpful in adapting to distortion of speech signals caused by background noise [101]. Rudner et al. [236] and Stenbäck [261] observed high WM individuals perceived less effort while recognising speech from noise.

Inhibition (IN) refers to the capacity to regulate attention, behaviour, thoughts, and/or emotions by overriding internal impulses or external '*lure*'—and maintaining focus on what is appropriate or needed [78]. To our knowledge, prior studies have not investigated the effect of IN on the relevance assessment process. High IN has been shown to help in speech recognition, especially in adverse conditions like the presence of background noise [261, 262].

A significant number of prior work have explored various aspects related to the process of relevance assessment. This work however considers the novel effect of document length and the cognitive abilities of assessors to explore the utility of the voice modality with regards to judging relevance.

5.3. METHODOLOGY

To address our three research questions outlined in §5.1, we conducted a crowdsourced user study. The study participants were asked to judge the relevance of *Query/Passage* (*Q/P*) pairings, where passages were presented either in the form of *text* (i.e., a piece of text) or *voice* (i.e., an audio clip). In our study, passage **presentation modality** is a *between-subjects* variable. We also controlled the **length of passages**; this is a *within-subjects* variable to ensure that participants judged passages of varying lengths. The *independent variables*

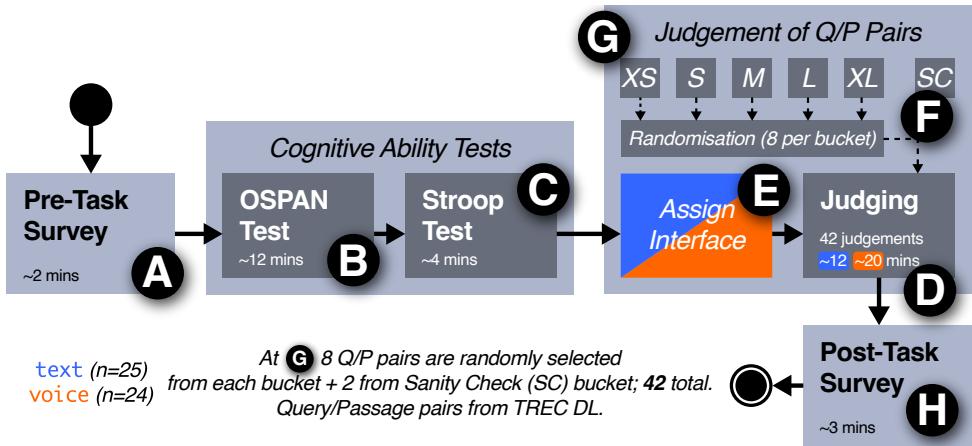


Figure 5.1: An overview of the user study protocol, including approximate times for participants to complete each component. Refer to §5.3.1 for mappings to the letters highlighting key aspects of the study procedure.

5

working memory and **inhibition** allow us to estimate the impact of the cognitive abilities of the participants on the accuracy of their judgements, time taken and perceived workload.

5.3.1. STUDY OVERVIEW

Figure 5.1 presents a high-level overview of the user study design.¹ The diagram highlights the main tasks that study participants undertook. Lasting approximately 32 minutes for **text** and 40 minutes for **voice**, the study consisted of four main parts: (i) the *pre-task survey* (§5.3.6); (ii) the *cognitive ability tests* (§5.3.3); (iii) the *judgements* (§5.3.4); and (iv) the *post-task survey* (§5.3.5).

After agreeing to the terms of the study, participants completed a pre-task survey ①. This survey included demographics questions, including questions about their familiarity with voice assistants—as reported in §5.3.6. Participants would then move onto two *psychometric tests*; as outlined in §5.3.3, these tests measured their cognitive abilities with respect to working memory ② and inhibition ③. Participants undertook a short practice task to help them familiarise themselves with the interface for each test.

After the psychometric tests, participants moved to the main part of the study: judging Q/P pairings ④. The experimental system first assigned the participants to either **text** or **voice** randomly ⑤ (§5.3.4). Based on the assigned condition, participants then judged a total of 42 Q/P pairings presented to them in a random order to mitigate the effect of topic ordering [245, 246] (§5.3.2)—40 were selected from the 2019 and 2020 TREC Deep Learning (DL) track, and the remaining two acted as a *sanity check* (SC) ⑥. The 40 passages belonged to different *answer length buckets* §5.3.2 ⑦. Finally, the participants would be taken to the post-task survey ⑧.

¹Note that circles refer to superimposed labels on the illustration in Figure 5.1.

5.3.2. QUERY/PASSAGE PAIRINGS

As mentioned, we obtained the Q/P pairings from the 2019 and 2020 TREC DL track—specifically the passage retrieval task [71, 72, 171]. The test partition of the datasets contain 43 and 54 natural language queries with passages that are judged by *NIST* assessors. Using a graded relevance scale, passages for each query were judged by assessors as: (i) *perfectly relevant* when the passage is dedicated to the query, containing an exact answer; (ii) *related* when the passage appears somewhat related to the query, but does not answer it completely; or (iii) *non-relevant*, when the passage has nothing to do with the provided query [71, 72]. We note that an additional relevance category exists (*highly relevant*). However, we ignore judgements of this category in our work (similar to [144]) in order to have a clear distinction between the different categories.

Sampling Procedure From the available test queries, we sampled 40 (due to budget constraints). As RQ2 states we are interested in how passage length affects assessments, we next determined five different buckets of passage length: from *very short* to *very long* (more details follow below). We randomly assigned the 40 queries to these five buckets, leading to eight queries per passage length bucket. For each query, we sampled three passages from the QRELS, with the additional condition that the sampled passages must fall into the query’s passage length bucket: one *perfectly relevant*, one *related*, and one *non-relevant* passage. And thus, each bucket contains 24 passages pertaining to eight queries. Table 5.1 demonstrates three Q/P examples, each coming from a different length bucket.

Sanity Check (SC) We also created two additional Q/P pairings to act as a sanity checks in order to perform quality control of the relevance judgements by our participants, as suggested by Scholer et al. [246].² *We did not consider the SC Q/P pairs in our data analysis.*

Judgements per Participant We presented all our participants with the *same* set of 40 queries + 2 SC queries in order to mitigate effects arising due to differences in queries [74]. Each participant judged one randomly sampled passage—out of the three available ones—for each of the 40 queries (ignoring the SC queries). We thus collected relevance judgements on a total of $40 \times 3 = 120$ Q/P pairs.³ Each participant judged 13 passages per QREL.

Passage Length Buckets To add more detail to our passage length bucketing procedure, we chose five types of length buckets: **Xs** (*Very Short*); **S** (*Small*); **M** (*Medium*); **L** (*Long*); and **XL** (*Very Long*). They corresponded to the 0–5, 5–50, 50–75, 75–99 and 99+%-ile of the lengths of all judged passages of the 97 test queries in our TREC-DL datasets. We selected the percentiles to have a range of 20 to 30 words per passage length bucket. The concrete word ranges for each passage length bucket can be found in Table 5.2.

²The sanity check questions were: (i) *Who was the lead vocalist of Queen?*, with the answer passage being perfectly relevant; and (ii) *What is the difference between powerlifting and weightlifting?*, with the answer passage being non-relevant.

³The list of collected Q/P pairs are available [here](#).

Table 5.1: Examples of *Query/Passage (Q/P)* pairs for different passage length categories. The (Q id) is taken from the TREC datasets. We also provide links to [audio] clips of the respective passages.

Passage Length	Query (Q id)	Ground Truth Relevance & Passage
Very Short (XS)	<i>What metal are hip replacements made of?</i> (877809)	RELEVANT Some prostheses, like hip and knee joints made of cobalt chrome, contain some trace of nickel and for patients with allergies to this may have to go with Titanium joints. [Audio]
Short (S)	<i>Who has the highest career passer rating in the nfl?</i> (1056416)	SOMEWHAT-RELEVANT Wilson is the only quarterback in NFL history to post a 100-plus passer rating in each of his first two seasons, and he's already won a Super Bowl. Dan Marino is really the only quarterback you could argue was better out of the gate. [Audio]
Long (L)	<i>What is the appearance of granulation tissue?</i> (1133579)	NON-RELEVANT The protective outer layer of the plant. Everything needs skin, or at least some sort of a covering, for plants, it's a system of dermal tissue. Which covers the outside of a plant and it protects the plant in a variety of ways. Dermal tissue called epidermis is made up of live parenchyma cells in the non-woody parts of plants. Epidermal cells can secrete a wax-coated substance on leaves and stems, which becomes the cuticle. Dermal tissue that is made up of dead parenchyma cells is what makes up the outer bark in woody plants. [Audio]

5

From Text Passage to Audio Clip We processed the passages to remove any unwanted punctuation, leading and trailing whitespace, and corrected a few spelling errors. These cleaning steps were necessary as we did not want the participants to be distracted by unclean text, and to create legible audio clips for the *voice* interface. We used *Amazon Polly*—an open-source text to speech system with an array of options for language and voice types—to generate the audio clips for the voice results.⁴ Specifically, we chose *Matthew*, a male US English voice, with a speed of 95% as the authors unanimously agreed that this particular setting (among other evaluated voice options) had the clearest pronunciation, in particular of difficult words that might appear in the passages.⁵ Lastly, we ran a pilot study ($n = 5$) where participants were asked to rate the pace, accent, and length of our generated

⁴<https://aws.amazon.com/polly/>

⁵Difficult words in this context include words from languages other than English (e.g., “...and include Gruyère, Emmental, Tête De Moine, Sbrinz..”), words specific to a domain (e.g., “...the manubrium, sternbrae, and xiphoid cartilage.”), etc.

Table 5.2: Overview of passage length buckets. Averages are reported together with the standard deviation.

Passage Length	Min-max #words	Avg. #words	Min-max audio clip length (s)	Avg. audio clip length (s)
Very Short	12 – 32	24.67(± 5.3)	3 – 13	10.04(± 2.4)
Short	33 – 53	41.67(± 3.8)	14 – 19	17.04(± 1.4)
Medium	54 – 74	63.17(± 5.6)	20 – 30	25.17(± 3.4)
Long	90 – 120	99.79(± 6.9)	31 – 42	36.04(± 3.04)
Very Long	121 – 151	139.96(± 8.2)	48 – 70	54.58(± 4.9)

audio clips on a seven-point scale. They reported an average score of 6.3, confirming the high quality of the audio clips for our task. Table 5.2 shows the minimum, maximum, and average length of the audio clips in seconds for the passages belonging to the five length buckets.⁶

5

5.3.3. COGNITIVE ABILITY TESTS

In order to measure the cognitive abilities of our participants with relation to judging the presented Q/P pairings, we chose two established psychometric tests that examine both an individual’s working memory and their inhibition. Prior work [101, 236, 261] has shown that working memory and inhibition play an important role in speech understanding.

Working Memory To measure working memory capacity, we used the *Operation-word-SPAN (OSPAK)* test [280] that has also been used in prior *Interactive IR (IIR)* work [62]. The OSPAN test measures an individual’s ability to recall letters displayed in sequence, while concurrently completing simple secondary tasks. Participants completed eight trials of varying lengths. During each trial, participants were shown a sequence of 3 – 7 letters, and were then asked to recall the letters in their original order from a grid display. Additionally, during each trial, participants completed simple mathematical problems between each letter shown in sequence (e.g., “is $8+6=15$?”). The final score was equal to the sum of sequence lengths of all trials perfectly recalled. A higher score in the OSPAN test indicates a participant’s greater ability to hold information (the letter sequence in correct order) in short-term memory when it is no longer perceptually present.

Inhibition To measure inhibition, we used the *Stroop test* which was first introduced in 1935 [263]. As an example, the Stroop test has been used to measure inhibitory attention control in learning [96, 129] and speech processing [261]. We used a computerised version of the test that was also used in the IIR study undertaken by Arguello and Choi [14]. During the Stroop test, participants were shown a sequence of words indicating one of four colours: red, green, yellow, or blue. Some of the words displayed are congruent (e.g., the word “blue” displayed in blue font), and others are incongruent (e.g., the word “blue” displayed in red font). For each word, participants had to indicate the *font colour* of the word as quickly

⁶Audio clips for all the passages are released [here](#).

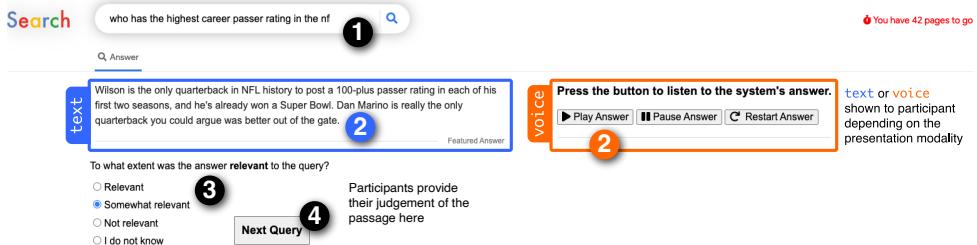


Figure 5.2: Composition screenshot of both the **text** and **voice** interfaces used by participants for judging query-passage pairs. Circled numbers correspond to the same in the narrative, found in §5.3.4.

5

as possible by clicking on the correct option presented as a list (the trial continued until the correct colour was chosen). Participants had to complete 48 correct trials (similar to the study by Arguello and Choi [14]), of which 24 are congruent and 24 are incongruent. The final score is equal to the participant's average response time (in milliseconds) for the incongruent trials, minus the average response time for the congruent trials. Response times are typically slower for the incongruent trials, an effect referred to as the *Stroop effect*. Lower scores are better for the Stroop test, with higher scores indicating a greater difficulty in focusing on the relevant stimulus (the colour of the word) and ignoring the non-relevant stimulus (the word itself).

5.3.4. ASSESSOR INTERFACE

Our study interface is shown in Figure 5.2, as a composition of both the **text** and **voice** interfaces. The **text**-specific components are highlighted in blue; **voice**-specific ones in orange. For each Q/P pairing they were required to judge, participants were presented with a static query box ① which could not be altered; it displayed the query for which the participant was to judge the passage for. Only one passage was shown ②; depending on the condition, this was either presented as text (for **text**), or a series of buttons to control the audio clip (for **voice**). In the case of **voice**, the participant had to press the **Play Answer** button to listen to the audio clip. They could also pause and restart the audio clip by pressing the **Pause Answer** and **Restart Answer** buttons respectively.

Once they had read or listened to the answer passage, participants then moved to the underlying form located at ③ to provide their judgement of the passage. Participants could choose between '*Relevant*', '*Somewhat relevant*', '*Non relevant*', and '*I do not know*'. We included the final option to ensure that participants were not forced to make a relevance decision in the case that they were not sure as it has been shown that assessors are not always certain of their judgements [2]. We did not provide the participants with the option to skip parts of the audio clip or adjust the speed. Certain checks were in place to ensure reliability of relevance judgements of participants, *in addition* to the two SC pairings as outlined in §5.3.2. For **text**, the form for marking relevance ③ appeared after five seconds. For **voice**, the form for marking relevance ③ appeared after 50% of the audio clip had been played. Participants could also proceed to judge the next query/passage pair by clicking the **Next Query** button ④ which was enabled only after a participant made their judgement. Once participants moved on to the next pairing, they could not go back to revise earlier judgements. No time limit was imposed on participants during the judging process.

5.3.5. OUTCOME MEASURES

In addition to the use of the two psychometric tests outlined in §5.3.3, we used interaction logging apparatus and additional surveys to capture both behavioural and experience data respectively.

Measuring Participant Behaviours We added the *JavaScript* library [LogUI](#) [182] into our web-based judgement interface; it allowed us to capture a variety of different behaviours and events such as: (i) when the page was loaded; (ii) clicks on the form to record the judgement made by a participant; and (iii) clicks on the [Play/Pause/Restart](#) buttons (for [voice](#)). From these events, we could compute the amount of time taken for an individual to make a judgement—that is, from when the page loaded (showing the query/passage pairing) to when the [Next Query](#) button was clicked ④ (Figure 5.2). In turn, this allowed us to compute the *time per relevance judgement*, as reported in our results.

Measuring Participant Experiences After completing the relevance judgements, participants completed the post-task survey. Participants were asked about their perceived workload based *only* on their perceived experiences of the relevance judgement tasks. To measure workload, we used five questions from the raw *NASA TLX* survey, as proposed by Hart and Staveland [108]. This instrument has been used (in slightly different forms) in several prior IIR studies (e.g., [14, 16, 232]). The five selected questions from the *NASA TLX* are designed to measure perceived: (i) mental demand; (ii) effort; (iii) temporal demand; (iv) frustration; and (iv) performance. We omitted the ‘*physical demand*’ question from the survey as it was not relevant to our task.⁷ Participants responded to the five *NASA TLX* questions using a seven-point scale (from “*poor*” to “*good*” for performance and from “*low*” to “*high*” for the remaining four).

5

Measuring Participant Performance We also computed the *accuracy* of our participants in the relevance judgement tasks. Accuracy was calculated in terms of how many Q/P pairs participants judged *correctly*—that is, their relevance judgement matching the ground truth from the QRELS. We also aggregated relevance judgements of participants on each Q/P pairing based on majority voting, as done by Kutlu et al. [144] to observe if collective judgements are more accurate. We used Krippendorff’s alpha (α) to measure inter-annotator agreement (as used by Damessie et al. [74]). Lastly, we calculated Cohen’s kappa (κ) [17, 29, 56] which measures the agreement of judgements with ground truths by considering chance.

5.3.6. PARTICIPANT DEMOGRAPHICS

We conducted an *a-priori* power analysis using *G-power* [88] to determine the minimum sample size required to test our **RQs**. The results indicated that the required sample size—to achieve 95% power for detecting an effect of 0.25, with two groups (modality) and five measurements (passage length)—is 46. As such, we recruited 50 participants from the Prolific platform. We disqualified one participant as they failed to correctly judge our sanity check Q/P pairs (§5.3.2). Our $n = 49$ (25 for [text](#), 24 for [voice](#)) participants were

⁷This was also done in prior studies, such as the study reported by Vtyurina et al. [293]

native English speakers, with a 98% approval rate on the platform—a minimum of 250 prior successful task submissions, and self-declared as having no issues in seeing colour. Participants were required to use a desktop/laptop device in order to control for variables that might affect results of the Stroop and OSPAN tests on other (smaller) devices. From our participants, 22 identified as female, 24 as male, with 3 declining to disclose this information. The mean age of our participants was 38 (*min.* 22, *max.* 69). With respect to the highest completed education level, 28 possessed a Bachelors (or equivalent), nine has a Masters (or equivalent), ten had a high school degree, and two had a PhD (or equivalent). We also asked participants how often they used a smart speaker to search for information, and listening to the provided answer—to which 13 reported daily usage, 20 said usage on a weekly basis, and 16 said never. Participants were paid GBP £11/hour.

5.4. RESULTS AND DISCUSSION

This section presents the results of our experiments pertaining to our three RQs. First, we provide details on the statistical tests we conducted, and how we utilised the cognitive ability tests to divide participants into *low-* and *high-ability* groups.

5

Statistical Tests For our analyses, we conducted a series of independent sample *t*-tests with Bonferroni correction ($\alpha = 0.05$) to observe if the modality of presentation has a significant effect on our dependent variables—accuracy of relevance judgements, the time taken to judge, and the perceived workload (RQ1).⁸ We also conducted a series of mixed factorial ANOVA tests (where modality of presentation is a *between-subjects* variable, and passage length is a *within subjects* variable) to observe if presentation modality, passage length, or the interaction between them have a significant effect on accuracy of relevance judgement and time taken (RQ2). Lastly, we conducted a series of three-way ANOVA tests to observe if the two user dispositions—working memory and inhibition—or their interaction with modality of presentation have a significant effect on the three dependent variables (RQ3). For RQ2 and RQ3, we followed up the ANOVA with pairwise Tukey tests with Bonferroni correction ($\alpha = 0.05$) to observe where significant differences lay. In the case where no significant difference was observed between the two conditions, we used equivalence testing between conditions through the *two one-sided t-tests (TOST)* procedure. The upper and lower bounds for the TOST was set at 7.5% ($-\Delta L = \Delta U = 7.5$) for accuracy, as Xu et al. [307] observed that LtR models were robust to errors of up to 10% in the dataset (we used 7.5% for *conservativeness*). For each scale of NASA-TLX, we set $-\Delta L = \Delta U = 2.04$, following Lee et al. [152], who used a bound of ± 18 on a 100-point NASA TLX. For our seven-point scale, it translates to ± 2.08 according to the formula of Hertzum [112].

Cognitive Ability Scores and High vs. Low Ability Groups To examine the effect of a participant's cognitive abilities on relevance judgement accuracy (RQ3), we performed a median split of the scores obtained by the participants in the OSPAN (*min.* 0, *max.* 50, *mean* = 25.4(± 12), *median* = 22) and Stroop test (*min.* = -300, *max.* = 650, *mean* = 171.25(± 184), *median* = 170) respectively. The mean scores of our participants for working memory and inhibition were within one standard deviation of the reference mean scores as reported

⁸All data and code pertaining to our analyses are [released](#).

Table 5.3: **RQ1:** Effect of modality of passage presentation on accuracy of relevance judgement, time taken per judgement in seconds and perceived workload (IV-VIII) per participant. We also report Krippendorff's α and Cohen's κ for accuracy. \dagger indicates significant difference in between the two conditions according to independent sample t-test. $*$ indicates the corresponding metric is equivalent for both conditions based on the TOST procedure.

Metrics	text	voice
I Accuracy *	68.40(± 9.15)%	65.94(± 8.56)%
α, κ	0.41, 0.61	0.37, 0.54
II Majority Voting Acc.	79.1%	75.8%
κ	0.76	0.71
III Time/Rel. Judge. (sec.) \dagger	17.56(± 9.08)	29.54(± 7.85)
IV Mental demand*	4.68(± 1.60)	4.83(± 1.37)
V Effort*	4.88(± 1.88)	4.00(± 1.50)
VI Temporal Demand*	4.04(± 1.86)	3.08(± 1.82)
VII Frustration \dagger	3.96(± 2.07)	1.83(± 0.82)
VIII Performance \dagger	4.16(± 1.93)	5.67(± 0.70)

5

in [14], validating our methodology. Participants were thus divided into a high- and low-ability group for each of working memory (based on OSPAN test scores) and inhibition (based on Stroop test scores). Note that for inhibition, a low test score indicates high ability. Prior studies have also analysed the effects of different cognitive abilities by dividing participants into low/high ability groups using a median split [3, 14, 62, 245].

5.4.1. RQ1: MODALITY OF PASSAGE PRESENTATION

Table 5.3 presents the main results for **RQ1**. There was no significant difference in judgement accuracy (row I, Table 5.3) between participants in **text** and those in **voice** ($t(47) = 0.97, p = 0.33$). TOST revealed that accuracy of judgements across both conditions were *equivalent* ($p = 0.02$). The inter-annotator agreement (α) was slightly higher in **text**. When using majority voting to aggregate relevance judgements (on average we had eight judgements per Q/P pair in each condition), we found that the accuracy increased from 68% and 66% to 79% and 76% respectively for **text** and **voice** (II, Table 5.3). This observation is in line with prior work [144], which shows that aggregating judgements from several assessors is more reliable than a single untrained assessor. Cohen's κ also increased with majority voting for both experimental conditions, indicating an increase in judgement reliability. Participants also showed similar trends of relevance judgement accuracy per relevance label category for both experimental conditions. As shown in Figure 5.3, participants in both conditions were most accurate in judging '*relevant*' passages (in line with findings by Alonso and Mizzaro [8]), followed by '*non-relevant*' passages. '*Somewhat relevant*' passages were most difficult to judge as participants in both conditions judged them correctly about half the time. With respect to the time taken to judge (III, Table 5.3), judgements in **text** were made significantly faster ($t(47) = -4.93, p < 0.001$) than in **voice**.

In terms of workload measured using NASA-TLX, there was no significant difference

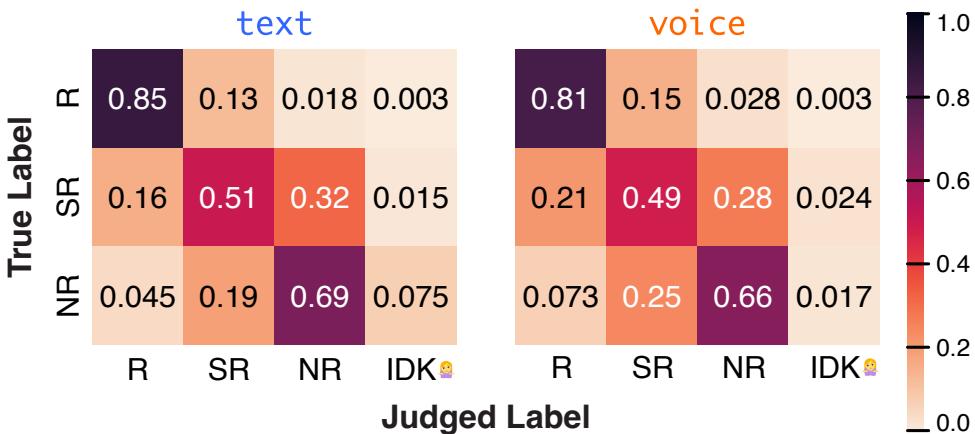


Figure 5.3: Accuracy of relevance judgements per label category for both **text** and **voice**. Diagonals represent percentage of time the true labels were *correctly* predicted by participants. Here, R = RELEVANT, SR = SOMEWHAT-RELEVANT, NR = NON-RELEVANT and IDK = I do not know.

5

in averages between the two cohorts in terms of perceived mental demand, effort, and temporal demand (**IV-VI**, Table 5.3). The TOST procedure revealed equivalent scores ($p < 0.05$) provided by participants for these three items of the NASA-TLX scale. For the other dimensions of NASA-TLX questionnaire, participants in **text** reported they felt significantly more frustrated (**VII**, Table 5.3) while performing the task than those in **voice** ($t(47) = 4.69, p < 0.001$). Participants in **voice** also reported significantly higher perceived performance (**VIII**, Table 5.3) when compared to the former ($t(47) = -3.60, p < 0.001$).

Overall, we found that participants listening to voice passages were equally accurate to their text counterparts. Vtyurina et al. [293] also observed that the probability of participants to identify the most relevant document was the same for both text and voice conditions. However, the authors implemented a different task design to ours. Their participants were presented with a list of results, and were significantly better at identifying the correct order of relevance when the summaries were presented in text modality. Insofar as to acknowledging the difference in task design, our observations with regards to the accuracy of participants with respect to relevance judgements across modalities are found to be partially in line with those of Vtyurina et al. [293]. We also observed that **voice** participants perceived a lower or equal workload when compared to those of **text**, in contrast to the other study's findings [293]. This can be attributed to their study setup. Contrary to ours, their presentation modality was a *within-subjects* variable. Our results indicate the proficiency of participants with both modalities for the given design of the task.

5.4.2. RQ2: PASSAGE LENGTH

Table 5.4 presents results related to **RQ2**. Like modality of presentation, passage length or its interaction with presentation modality did not have a significant effect on the relevance judgement accuracy (comparing rows **Ia** and **Ib**, Table 5.4). The TOST procedure revealed that for **XS** ($p = 0.01$) and **L** ($p = 0.001$) passages, judgement accuracy was *equivalent* across both conditions. Aggregating judgements via majority voting increased relevance judgement accuracy across all passage lengths for both text and voice conditions (comparing

Table 5.4: RQ2: Effects of passage length and presentation modality on accuracy of relevance judgements (with Krippendorff's α , Cohen's κ) and time taken. A bold number indicates that the metric for the corresponding presentation modality is significantly more than that for the other modality for the particular passage length. xs, s, m, xl indicates significant difference (within the same experimental condition) compared to XS, S, M, L, XL passage lengths. * indicates equivalence between the two conditions.

Metrics	Mode	Passage Length			XL
		XS	S	M	
I Accuracy (%)	text	66.7(± 19.0)*	74.5(± 14.7)	66.5(± 17.5)	61.0(± 18.0)*
	α, κ	0.37, 0.57	0.51, 0.67	0.43, 0.55	0.29, 0.50
II Maj. Voting Acc. (%)	voice	67.7(± 21.9)*	64.06(± 15.0)	72.4(± 19.4)	61.5(± 13.9)*
	α, κ	0.39, 0.56	0.44, 0.49	0.49, 0.63	0.27, 0.51
III Time Taken (sec.)	text	75	83	79	75
	κ	0.73	0.81	0.74	0.73
III Time Taken (sec.)	voice	79	79	79	67
	κ	0.78	0.78	0.73	0.62
III Time Taken (sec.)	text	14.11(± 6.3)	15.25(± 7.7)	15.15(± 6.8)	21.39(± 12.41)
	voice	17.3(± 5.0) $m_{l,xl}$	<u>25.47</u> (± 11.8) xl	<u>28.45</u> (± 14.8) xs,xl	<u>31.04</u> (± 6.15) xs,xl

rows **Ia-IIa** and **Ib-IIb**, Table 5.4). However, for **XL** passages (**IIa-IIb**, Table 5.4), the difference in accuracy after majority voting was more than 10% (with **text** being more accurate). We also observed a higher difference in Cohen's κ and Krippendorff's α for **XL** passages between the **text** and **voice** conditions. These results indicated a higher inter-annotator agreement and reliability of judgements for **text** compared to participants in **voice** with regards to **XL** passages.

With respect to the time taken for judging, we have already seen (Section 5.4.1) that presentation modality significantly affected the time to judge. Mixed factorial ANOVA showed that passage length had a significant main effect ($F = 21.6, p = 3.3e^{-15}$) on the time taken to assess. A post-hoc test revealed a significant difference in the time taken to judge of the following pairs of passage lengths (with the latter passage length category taking more time): **XS-M** ($p = 0.02$), **XS-L** ($p < 0.001$), **XS-XL** ($p < 0.001$), **S-XL** ($p < 0.001$) and **M-XL** ($p = 0.001$). There was also a significant interaction effect between passage length and presentation modality on the amount of time taken. Pairwise Tukey test revealed that except for **XS** passages, judging relevance in **voice** took significantly longer for participants as compared to doing the same in **text (bold numbers)**, row **III**, Table 5.5). In **voice (IIIb**, Table 5.5), it took participants significantly longer to judge relevance, as passages (audio clips) increased in length. Superscripts (in Table 5.4) indicate which pairs of passage length were significantly different in **voice** in terms of time taken per judgement.

5

In summary, we did not observe a significant difference in relevance judgement accuracy across different passage lengths in both conditions. We observed judging relevance of **XS** passages was *equivalent* in terms of accuracy and time taken across both **text** and **voice**. However, for **XL** passages, relevance judgements in **text** were more reliable (indicated by majority voting accuracy, α and κ when compared to that in **voice**). There was no clear trend between passage length and assessor agreement observed in contrast to findings from [58], possibly due to differences in the type of documents assessed. Although it took longer on average to judge a lengthier passage in **text**, there was no significant difference in terms of the time taken to judge relevance of different passage lengths (a similar trend as observed in [141, 252]). For longer passages, participants in **voice** took significantly longer to judge relevance than in **text**. For **XL** passages, we found that participants were taking twice as long in **voice** when compared to **text**.

Why does it take longer for participants to judge longer passages in the **voice condition?** In order to control for confounding variables, we did not let participants speed up the audio clips, nor did we provide them with a seeker bar to skip ahead. We found evidence that participants moved on to the next Q/P pairing as soon as they were satisfied with their assessment. Indeed, they did not wait for the audio clip to finish playing before moving on to the next Q/P pair for longer passages (Figure 5.4 (a)). We also let participants mark the relevance of a passage in **voice** only after 50% of the audio clip had been played (Section 5.3.1). However, as seen from Figure 5.4 (b), participants took longer to judge relevance (rather than right at the 50% mark). For **XL** passages, it was at the 66% of the audio clip on average. This suggests that it indeed took more time for participants in **voice** compared to **text** to assimilate the information and come to a judgement decision for longer passages.

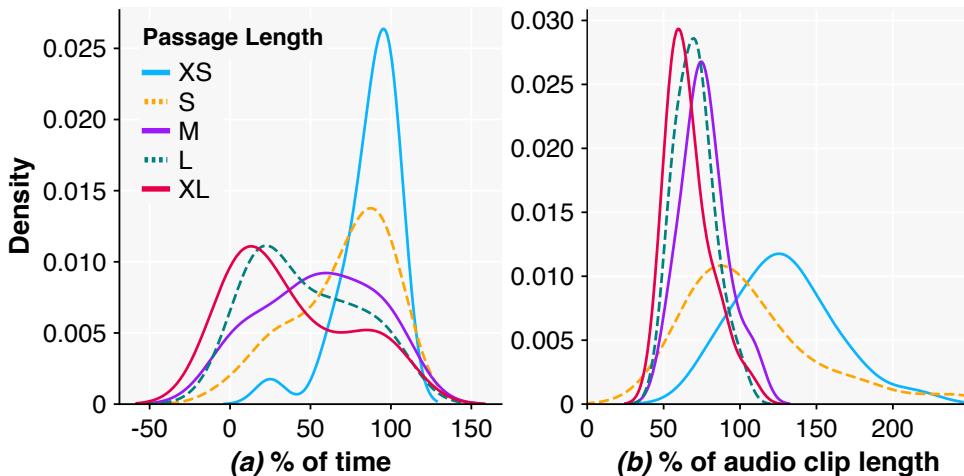


Figure 5.4: The trend of voice participants judging relevance w.r.t. time taken for passages of various length: (a) % of time participants listened to the entire audio clip; and (b) at what point was relevance judged (as a % of audio clip length).

5

5.4.3. RQ3: ASSESSOR COGNITIVE ABILITIES

Table 5.5 contains the results for our third research question. Here, \checkmark indicates a significant effect ($p < 0.05$) on the particular dependent variable, and \times indicates no significant effect.

None of the independent variables—modality of passage presentation (**PM**), working memory (**WM**), and inhibition (**IN**)—had a significant main effect on judgement accuracy. The interaction between the **IN** of participants and presentation modality (**IN** \times **PM**) had a significant effect on the accuracy ($F = 4.89, p = 0.03$). Pairwise Tukey test revealed that in **voice** participants with higher **IN** performed significantly better than those with lower **IN** ($70.5 \pm 7.2\%$ vs. $59.5 \pm 4.8\%$). The post-hoc test ($p = 0.01$) also revealed participants with low **IN** performed significantly better in **text** than those in **voice** ($70.0 \pm 9.5\%$ vs. $59.5 \pm 4.8\%$). We found significant main effects of **PM** on the time taken to judge relevance ($F = 22.17, p < 0.001$), reaffirming findings from Section 5.4.1 and Section 5.4.2.

With respect to the perceived workload, working memory had significant main effects on perceived temporal demand ($F = 7.88, p = 0.01$). A post-hoc test ($p < 0.001$) revealed that participants with high **WM** reported significantly less temporal demand as compared to those with low **WM** (2.5 ± 1.3 vs. 4.6 ± 1.7 respectively). **IN** also had significant main effects on perceived temporal demand ($F = 7.4, p = 0.01$). A post-hoc test ($p < 0.001$) revealed that participants with high **IN** reported significantly less temporal demand as compared to those with low **IN** (2.74 ± 1.4 vs. 4.59 ± 1.9 , respectively). Presentation modality had significant main effects on perceived frustration ($F = 8.36, p = 0.008$) and performance ($F = 5.83, p = 0.02$)—confirming observations from Section 5.4.1—with participants in **voice** reporting a lower workload. Lastly, the interaction between **WM** and presentation modality (**WM** \times **PM**) had a significant effect on perceived effort for the task ($F = 5.1, p = 0.03$). Post-hoc tests revealed that participants with high **WM** felt that judging using **text** required significantly more effort when compared to those in **voice** ($p = 0.001$).

In summary, we found that **IN** is a more important trait than **WM**, specifically for

Table 5.5: RQ3: Summary of main effects of *Presentation Modality (PM)*, *Working Memory (WM)*, *Inhibition (IN)*, and effects of the interaction of **WM** and **IN** with PM on accuracy of relevance judgement, time taken, and perceived workload. A ✓ indicates significant effect of a 3-way ANOVA test ($p < 0.05$) on the particular dependent variables and ✗ indicates no significant effect.

	PM	WM	IN	WMxPM	INxPM
I Accuracy	✗	✗	✗	✗	✓ $F = 4.89$ $p = 0.03$
II Time Taken (sec.)	✓ $F = 22.17$ $p < 0.001$	✗	✗	✗	✗
III Mental Demand	✗	✗	✗	✗	✗
IV Effort	✗	✗	✗	✓ $F = 5.1$ $p = 0.03$	✗
V Temporal Demand	✗	✓ $F = 7.88$ $p = 0.01$	✓ $F = 7.39$ $p = 0.01$	✗	✗
VI Frustration	✓ $F = 8.36$ $p = 0.008$	✗	✗	✗	✗
VII Performance	✓ $F = 5.83$ $p = 0.02$	✗	✗	✗	✗

relevance judgement accuracy in the **voice** modality. Low **IN** participants in the **voice** condition were less accurate—since we *did not control for the audio device of the participants*, and consequently not for the background noise they were subjected to, low **IN** participants in **voice** were less effective in focusing on the passages while judging relevance [261, 262]. We leave exploring the effect of background noise as future work. In our study, the interplay between cognitive abilities and modality of presentation on perceived workload had different effects. High **IN** and **WM** participants felt less temporal demand. High **WM** in **text** felt more perceived effort compared to those in **voice**. Our results imply that we should design tasks for collecting relevance assessments to match the preference and abilities of crowdworkers [9, 206].

5.5. CONCLUSIONS

We explored the feasibility of using **voice** as a modality to collect relevance judgements of query-passage pairs. We investigated the effect of passage length and the cognitive abilities of participants on judgement accuracy, the time taken, and perceived workload.

RQ1 On average, the relevance judgement accuracy was equivalent across both **text** and **voice**. Participants also perceived equal or less workload in **voice** when compared to **text**.

RQ2 For **XS** passages, the performance and time taken for relevance judgements was *equivalent* between both **voice** and **text**. As passages increased in length, it took participants significantly longer to make relevance judgements in the **voice** condition; for **XL** passages **voice**, participants took twice as much time and the judgements were less reliable compared to **text**.

RQ3 Inhibition impacted the relevance judgement accuracy in the **voice** condition—participants with higher inhibition were significantly more accurate than those with lower inhibition.

Our results from **RQ1** suggest that we can leverage the voice modality for this task. **RQ2** points to the possibility of designing hybrid tasks, where we can use the voice modality for judging shorter passages and text for longer passages. The results of **RQ3** showed that selecting the right participants for the relevance judgement task is important. We should be mindful to personalise the task to match the preference and abilities of crowdworkers [9, 206].

Future Work There are several open questions for future work. We did not provide participants with the option to speed-up voice passages—*does letting them speed-up or skip passage parts reduce time for longer passages without reducing accuracy?* We also did not test the limit of length—*how long can documents be for equal accuracy in the text and voice modality?* Future work should also explore mobile devices for playing voice passages—*can we collect relevance judgements by offering more flexibility to crowdworkers?* Lastly, since asking to provide rationales for judgements has been shown to improve relevance judgement accuracy of crowdworkers in the text modality [144], exploring the effects of rationale in voice-based relevance judgements should be a worthwhile endeavour.

6

CONCLUSIONS

In this thesis, we conducted four studies to deepen our understanding of user interactions with Information Retrieval (IR) systems on four different aspects of Interactive Information Retrieval (IIR)—(i) effect of Search Engine Results Page (SERP) layout and task complexity on user interactions; (ii) the influence of learning tools on user interaction and learning outcomes during a learning-oriented search process; (iii) modelling user interaction; and (iv) effect of document modality on the collection of relevance judgments. In this chapter, first, we summarise our main findings from the four chapters of our thesis, where we revisit our broad research questions and provide several potential future work directions that follow directly from those research questions. In the second part, we discuss the broader directions of future work in IIR.

6.1. (RE)-EXAMINING USER INTERACTIONS

In Chapter 2, we looked into the effect of SERP layout and task complexity on user interactions. As the SERP layout of search systems has evolved considerably over the last decade, in this chapter, we looked into to what extent user search behaviour has changed during this time. We reproduced the experimental setup of prior two studies— Arguello et al. [16] (published in 2012) as well as Siu and Chaparro [256] (published in 2014)—that both investigated how the presence or absence of heterogeneous content, the layout of the SERP (list vs. grid), and task complexity influence user interactions on the SERP. Specifically, we examined how many of the eight observations from the two studies still hold today (2022). To this end, we asked,

B-RQ1: How do the layout of the SERP and the complexity of the task at hand affect user interaction? To what extent has user interaction with web search engines changed in the last ten years?

To answer **B-RQ1**, we conducted a user study with 41 participants where the layout of the SERP and task complexity were the primary dependent variables. We experimented with four different SERP layouts and four different levels of task complexity. We found

that both SERP layout and task complexity significantly affected various aspects of user interaction with web search results. Firstly, we observed that participants had significantly more clicks on web results when they were present in the list layout as compared to the grid layout of the SERP. This observation was contrary to what [256] observed in their work. Secondly, participants interacted more with web results as the task became more complex, partially in line with observations from [16, 256]. Thirdly, we observed significantly more interactions with image results for the simpler **Remember** tasks compared to the more complex **Analyse** or **Understand** tasks contrary to what [16] observed. Lastly, we did not find significant differences in how users perceived the different SERP layouts regarding usability, aesthetics, etc. This observation was aligned with that from [16, 256]. Overall, we found evidence to confirm two, with partial evidence for a further four observations from the studies.

One of our goals in this chapter was to reproduce the setup of [16, 256] to the best of our ability to identify critical factors that can play an essential role during reproducing past IIR studies. We pointed out that interfaces, task descriptions and participant cohorts are crucial, among other things. Releasing resources/code to replicate the interface design will help eliminate confounding variables that are possible because of a different implementation. In this chapter, we created templates of SERPs that resemble google.com and you.com and released them for further use. IIR studies should also mention the entire list of search tasks/topics they employ together with study descriptions. It is essential to have a fixed set of tasks and similar interfaces to reproduce and enable reliable comparison of observations (e.g., the number of queries, documents opened, etc.) with prior IIR studies. Two studies by Urgo et al. [285, 286] list examples of tasks of different complexities, which also offer helpful resources for future IIR studies. Lastly, since it is challenging to have the same cohort of participants, crowdsourcing with power analysis (to determine the number of participants required given the experimental conditions of a particular study) can form a reliable alternative [318]. These will help us with the reproducibility of IIR studies and enable reliable comparison of results.

Future Work We limited the functionalities of the SERP offered to the participants—we did not let them query and provided them with static pages resembling SERP and clickable results. Do our observations still hold when users can issue their queries? Furthermore, we fixed the positions of vertical results on the SERP in our study. We know from previous work [249, 265] that user interactions with verticals are affected by where they are displayed on the SERP. Understanding how user behaviour changes with the position of SERP components is the motivation behind our research question in Chapter 4, and we highlight future work in this regard in Section 6.3. Understanding and modelling user interactions will also help research on methodologies for interface optimisation [295] and SERP evaluation. Prior work [111, 153, 248] have inferred search goals from user search behaviour like clicks on documents and querying behaviour. Do user interactions with all modern SERP components, widgets and vertical results help us better infer user search goals? Understanding search goals can also help us design adaptive SERP interfaces tailored to specific search goals. Can we modify the SERP layout based on the complexity of the task undertaken? Do such modifications affect user performance?

With the ever-increasing innovation in the development of modern devices (e.g., mobile

devices, voice assistants, augmented-reality glasses, etc.), there is a considerably large scope of future work in interactive information retrieval. Web searching is a significant activity on mobile devices [79]. Prior work has looked into mobile devices for web search in the context of SERP presentation [4, 139, 211], user behavior [5, 12] and user performance [77, 138]. How to incorporate results from different verticals and how that affects user performance for different levels of task complexities in the case of mobile devices is still an open question. Recently, there have also been new developments in visual and augmented reality (VR/AR) devices (e.g., Apple Vision Pro,¹ Meta Quest Pro,² etc.). Recent work [215, 244, 309] have also explored and developed prototypes on how augmented reality can be used for information retrieval. Users can interact with their physical environment to retrieve information, such as pointing their device at an object and receiving relevant details. However, many open questions remain in this field. How do we incorporate multi-modal and heterogeneous results in such a system? Should results layout in augmented reality interfaces depend on the complexity of the information need of the user? If so, how should they be displayed to enhance the search performance and experience of users?

6.2. SEARCH AS LEARNING

In Chapter 3, we explored to what extent two active reading tools for *highlighting* and *note-taking* affect user search behaviour and learning outcomes during a learning-oriented search process. These tools have been shown to help in *offline* or *classroom* learning scenarios. Previous work has looked into various aspects of the SAL domain, like understanding user behaviour, optimising retrieval functions for learning, etc. However, they do not explore the impact of active reading tools in the SAL process. Hence, in this chapter, we asked:

B-RQ2: How do active reading tools affect user interactions and learning outcomes during a learning-oriented search (SAL) process?

To this end, we conducted another user study with 115 participants observing the effect of two active reading tools (highlighting and note-taking) on their learning outcomes and search behaviour. We implemented a highlighting tool and a note-taking tool built into the search interface. We measured their learning over two tasks: a recall-oriented vocabulary learning task [189, 233] and a cognitively demanding essay writing task [162, 260]. We observed that neither the highlighting nor the note-taking tool helped participants in the vocabulary learning tasks. However, access to only the highlighting or note-taking tool helped them write better essays than the participants without the tools. We also explored how different highlighting and note-taking strategies help their learning outcomes. We showed that while engaging in complex learning-oriented search tasks on the web, merely highlighting and note-taking may not benefit learners. Instead, how these tools change how the learners scan and process text was more important for learning while searching.

Future Work This work explored user interaction with web search results and their highlighting and note-taking behaviour during the SAL process. As we observed that highlighting and note-taking strategies and various characteristics of learners affect their

¹<https://www.apple.com/apple-vision-pro/>

²<https://www.meta.com/nl/quest/quest-pro/>

learning outcomes, future work can examine how highlighting and note-taking behaviour can be used as proxies for their learning outcomes. Since, for an actual search system designed for learning, evaluating users' learning gain by conducting vocabulary or essay tests might not be a practical option, developing more implicit learning measures is also an open research direction. With reliable proxies of learning outcomes, we can also adapt retrieval functions to satisfy individual learners. Prior work have looked into optimising retrieval functions for learning [267–269] based on the documents clicked by a user. Analysis of highlighting and note-taking log data could provide a deeper understanding of the document understanding process. User behaviour while watching MOOC videos has been employed to evaluate the quality of video-based lectures and identify essential parts of such videos [148, 294]. Similarly, do highlights and notes indicate more relevant/interesting sections of a given document? Do they give a better idea of the topics the user learns? If so, could retrieval algorithms be manipulated based on their highlights or notes to promote documents? Possible research directions involving adaptation of the retrieval pipeline can look into (i) query expansion/suggestion based on highlighting/note-taking, (ii) learning to infer the knowledge state of users from their queries and highlighting/note-taking behaviour, (iii) fine-tuning transformers/LLMs for re-ranking of retrieved results based on the inferred knowledge state; (iv) learning to summarise retrieved answers (retrieval augmented generation [158, 173]) based on the inferred knowledge state.

6

Recently, social media applications like TikTok,³ Instagram,⁴ YouTube,⁵ etc. have become a significant source of accessing information on complex topics [93, 223, 259]. Video creators publish videos on various topics (from COVID-19 to improving snowboarding skills). These contents are searched and consumed by users to learn about those topics. Although there have been a few studies on how the presence of heterogeneous content (images, videos, etc.) affects user behavior [16, 186] and learning during web search [189], there are open research questions to understand the SAL process in these applications. To what extent user behaviour during searching and learning from such applications is different from that on a standard web search engine? How can we incorporate active reading tools to help learners in their sensemaking and knowledge-gaining process in such scenarios? Do those tools help them?

6.3. MODELING USER INTERACTIONS

In Chapter 4, we asked a more holistic question on the design of web search interfaces. An IIR practitioner/designer looking to incorporate a widget (e.g., note-taking tool) on a SERP has numerous potential options regarding the widget's location and other visual features (e.g., length/breadth, font size, etc.). However, they cannot run A/B tests on all possible options regarding the widget position due to budget constraints. Hence, in the thesis, we leveraged **Search Economic Theory** (SET), based on microeconomic theory, to build a model of user interaction that incorporates, for the first time, positional information of widgets. With such a user model, we can derive hypotheses of user behaviour based on the positioning of a widget. The ultimate goal is to select, based on the derived hypotheses, potential candidates for widget positions for running A/B tests. Previous work has utilised

³<https://www.tiktok.com/>

⁴<https://www.instagram.com/>

⁵<https://www.youtube.com/>

SET to develop models for predicting user interaction under various circumstances where widgets on the SERP are typically considered fixed, and their position is not part of the user model definition. Hence, our third research question in this thesis was,

B-RQ3: How can we utilise the Search Economic Theory model of user interaction to refine the design hypothesis space for widget positioning?

To this end, in Chapter 4, we derived a position-aware interaction model of search behaviour. We focused on the query history widget (QHW), as the widget is simple and easy to understand for users and involves only a small number of interactions—making it ideal as the first widget to employ for our exploration. We formulated a model to predict search behaviour related to reissuing queries from the same search session. We used Fitts' Law to approximate the cost of finding the widget based on its five different positions on the screen. Based on our model and prior work, we developed five testable hypotheses. We conducted a between-subjects user study with 120 participants. We evaluated the impact of the position of QHW on search behaviour. We did find that widget positioning plays a role and changes a user's search behaviour. We observed that users are likelier to use the QHW as query length increases. However, if the number of queries to check in the QHW increases, the likelihood of users using the QHW increases as its distance to the starting point decreases. Thus, position matters and should be incorporated into formal interaction models. Overall, we found partial support for three of the five hypotheses derived from our user model. This indicates that our model forms an acceptable starting point. However, we need to refine our model by incorporating new costs or having more realistic estimates of cost values from the study logs.

6

Future Work A natural extension of our work is to improve the realism of the user models by considering different search tasks of varying complexity, other widgets and other aspects of widgets (in addition to their position) like their shape, format, functionalities offered, etc. Graphical properties of widgets play an essential role in the amount of cognitive effort spent by the users and their perceived usability [115, 131, 270] consequently impacting their performance. Hence, modelling cognitive costs and user interactions with multiple widgets could help us strike a balance to optimise the complete user experience. Novel interface designs is an important topic, especially in the context of supporting learners on the web [52, 73]—it will also be worthwhile to research if incorporating various features of widgets like their position, size, functions, etc., in user models help select the optimal interface for learners.

Much work remains to determine how we can understand and subsequently model the cognitive processes and factors that influence how individuals behave when searching, especially in the presence of various widgets and other heterogeneous content on the SERP. The cognitive process and user behaviour also depend on (among other things) the device they use to search their prior topic knowledge. Thus, studies need to be conducted on building device-specific user models. What are the device-specific costs that are different from a desktop search? How do we estimate them? It also remains an open question about how realistic these user models can be or should be. It is important to note that the user models do not need to be perfect mirrors of human behaviour during a search scenario, but instead need to be *good enough*. By this, we mean that user models incorporating widget

position (or other) information should be able to help researchers derive helpful hypotheses of user behaviour that we can ultimately use to run A/B tests. In this regard, an important research direction is to evaluate how good a particular user model is before employing it for deriving such hypotheses. We need to ask, what are the essential characteristics of a good user model? Does a user model need to satisfy some properties to qualify to be a *good* one? What are those properties?

Developing user models can also enable us to run simulations of user interaction [52, 178, 179]. Such simulated data obtained from user models that incorporate information about SERP widgets can allow designing *adaptive user interfaces*—interfaces that can autonomously change the content, layout, or style depending on the user’s search context, capabilities and interests. These interfaces can provide significant user benefits by planning sequences of adaptations that gracefully lead them through gradual changes in their search process. An adaptive system must decide what to adapt and when or when not to make changes. They have been studied in the broader field of human-computer interaction, where methods like reinforcement learning have been employed to make these decisions on interface adaptations [147, 313]. In IIR, work has been done to adapt retrieval algorithms based on users’ prior and acquired knowledge during a SAL process [68]. It remains an open question on how to adapt SERP interface components to support users at various stages of their search journey. Estimating the utility or usefulness of an adaptation to the user is required for selecting an adaptation. Unfortunately, utility is complex to estimate accurately during a search process, both at design time and interactively from the kind of data these systems might have access to, such as clicks or document viewing duration. Moreover, in adaptive interaction, utility is also non-stationary. That is, the skills and interests of the user evolve. A change that would make sense in the beginning when the user is a novice to the design may be devastating for an experienced user. Open research questions in adaptive user interfaces include—what are the various cost and utility components while building user models for adaptive interfaces? How do we represent and estimate those components (that vary throughout the search journey of a user)? What tasks and search contexts can we apply and use such adaptive search interfaces? How can we evaluate the quality of the user model and, consequently, the effectiveness of the adaptations derived from those user models?

6.4. RELEVANCE JUDGEMENT COLLECTION

In the first three broad research questions, we primarily dealt with *text* documents and results. In Chapter 5, we focused on the *voice* modality in collecting qrels for test collection. As audio-only devices are getting popular, we explored to what extent we can leverage voice modality while building test collection using judgements from crowdworkers. While previous work have explored various factors (e.g., document length, topic difficulty, priming effects, etc.) affecting the relevance judgement effectiveness of crowdworkers, none of them has examined the impact of presentation modality of documents. Hence, in this thesis, we asked:

B-RQ4: How does the presentation modality of documents, text (reading on screen) vs. voice (listening to audio clips) affect the relevance judgement process of assessors?

How do the cognitive abilities of assessors and their interplay with presentation modality affect the effectiveness of relevance judgment?

To answer **B-RQ4**, we conducted our last user study (Chapter 5) with 49 crowdworkers where we measured their relevance judgements effectiveness in terms of accuracy and time taken. We also measured their perceived workload using the NASA-TLX survey. We explored the effect of the cognitive abilities of assessors on their judgement effectiveness. Each crowdworker had to judge the relevance of query-passage pairs either by reading the passages on-screen or listening to audio clips of the passages. We found that judgement accuracy was equivalent between crowdworkers reading the passages and those listening. However, as passages increased in length, it took participants significantly longer to make relevance judgements when listening to them than those reading them. Our results suggest that we can leverage the voice modality for this task and the possibility of designing hybrid tasks, where we might use the voice modality for judging shorter passages and text for longer ones.

Future Work There are multiple open avenues to extend the research conducted in this chapter. Asking to provide rationales for judgements has been shown to improve the relevance judgement accuracy of crowdworkers in the text modality [144], a natural extension to our work can be exploring the effect of the same in the voice modality. Future work should also explore whether the voice modality can be used for judging the relevance of long documents. In our study, we found evidence that crowdworkers often tend *not* to listen to the entirety of the presented audio clips of the passages. Does this decrease relevance judgement effectiveness in the case of long documents? Does the position of the relevant portion (beginning, middle, end) of the answer in the document affect relevance judgement effectiveness, especially in the voice modality of presentation? Another future research direction is considering domain-specific documents (e.g., medical, legal, etc.) with domain-specific vocabulary. Does the modality of presentation affect relevance judgement effectiveness regarding these documents? Lastly, we did not provide speed control for listening to audio clips or a seeker bar to skip segments of the clip. Future work can explore how to adapt the speed of audio clips depending on the abilities of the crowdworkers and if providing speed control lets them judge longer documents faster without hurting the accuracy. Can LLMs summarise longer documents and then let crowdworkers judge their relevance? If possible, it will address the problem of crowdworkers taking longer to judge longer documents in the voice modality of presentation.

6

6.5. COMMON NEXT STEPS

There are certain *next-steps* which are required to generalise the findings of our studies. In this section, we list a few areas of generalisation that are common to all the studies we conducted. Firstly, we recruited participants via a crowd-sourcing platform. We limited our participants to native English speakers and paid compensation according to the Prolific platform's suggested compensation rate. We acknowledge that the participants come from a limited geographical distribution. Nonetheless, reproducing our studies with different groups of users (e.g., users speaking other languages, users with hearing/visual disabilities, etc.) is a crucial step to generalise the results beyond the population of crowdworkers we

employed. Secondly, in our studies, we instructed the crowdworkers to search on a fixed number of topics. Observing if our results generalise to other topics is also an important step to consider for the future. For example, asking participants to learn/search about topics with multiple perspectives (*Should schools mandate uniforms for their students?*) might lead to different observations in our studies. Finally, we only allowed crowdworkers to participate in our studies using a desktop. Controlling the device was essential to eliminate the confounding effects of different devices. To what extent our results hold beyond the desktop interface is an open question. Crucial aspects of our studies, e.g., the design of note-taking and highlighting tools in Chapter 3 or the user model in Chapter 4, are specifically tailored towards web search on desktop. Hence, one important next step is to consider different devices in the context of our research questions. We list potential open research questions concerning search devices and interfaces beyond the desktop:

- **Chapter 2:** Are particular layouts more suitable for different devices (e.g., list layout for mobile devices or grid layout for devices like Echo Show, Nest Hub)? What is an effective way to incorporate results from different verticals on the SERP interfaces of different devices?
- **Chapter 3:** Prior work [104, 185, 296] have shown that user performance decreases when the search interface is complex as it increases their cognitive load. Thus, we can ask, how does incorporating highlighting or note-taking tools for web search in smaller devices like mobile phones affect their learning outcomes or search behaviour? What is an efficient way of providing users with the advantages of these tools on such devices? What do highlighting or note-taking tools look like for SAL using voice-based search tools like Alexa, Siri, etc.?
- **Chapter 4:** Do we need a different user model when considering user interaction in other devices like mobile? How different are the costs associated with such devices? To address these research directions, we have to answer questions like, is the cost of querying in a mobile interface different than that in a desktop one? Is the cost of finding a past query in a *QHW* designed for the mobile interface different than the desktop version? Answering these questions will let us develop device-specific hypotheses of user interaction by incorporating widget information.
- **Chapter 5:** Can crowdworkers judge relevance *on-the-go* using voice-based devices [278, 279] while performing a secondary task like driving or commuting? Such devices will play both the query and the passages in voice modality (contrary to our study, where the crowdworkers saw the query on their desktop). In such a scenario, what is the effect of answer length and their cognitive aspects on the relevance judgement accuracy? Can we help crowdworkers with visual impairments perform the relevance judgement task using voice-based devices?

6.6. BROADER RESEARCH DIRECTIONS

Generative AI is now mainstream. We can generate high-quality text, images and audio using generative models like **Large Language Models** (LLMs). How might this reshape

⁵<https://www.amazon.com/echo-show-10/>

⁵https://store.google.com/us/product/nest_hub_2nd_gen

the questions we have posed in this thesis? Based on the discussion of the findings of this thesis, we provide several areas for future work in the broader field of IIR.

6.6.1. LLMs AND USER INTERACTION

Recently, LLMs such as GPT-3 [49], ChatGPT (GPT3.5 and GPT-4) [205], LLaMA [277], with billions of parameters and pre-trained on a massive amount of data, have been released. They have demonstrated an excellent ability to generate semantically correct and coherent texts, images, audio and video. They can provide direct and concise answers to queries that users typically search on commercial web search engines like Google or Bing. Users have found it as a useful tool for learning and coding, summarising and scanning through reading materials while asking questions and even assisting tasks requiring creativity, brainstorming ideas and presenting information.⁶ While enthusiasts have claimed ChatGPT might replace traditional web search engines entirely because of the concise and direct nature of the answers, several real-life examples have also revealed that chatGPT is often incorrect (*hallucinates* answers), makes mistakes while sounding plausible and reasonable [121]. The pros and cons of applications like ChatGPT provide numerous scopes for future research in various aspects of IIR, including but not limited to SAL, user modelling, and collection of relevance judgment.

Recently, Microsoft introduced Bing Chat where LLM-powered chat assistance like ChatGPT is present with a traditional SERP.⁷ Using such LLM-based tools to aid their web search will result in different user interactions than in a more traditional setup of commercial search engines like Google or Bing. An example is the kind of queries users tend to pose—in the former, due to its conversational nature, users pose questions in the form of natural language sentences instead of keyword queries as is done in the latter. Users also expect answers instead of links to documents. Users can browse links while asking complex questions to find comprehensive answers and summarised information from the chat assistance. Thus, studies need to understand user behaviour and performance using these new web search tools. New user models also need to be developed that consider new dimensions (e.g., cost of querying or assessing an answer is different in a *conversational* interface where users search for information by asking questions to an LLMs) of search introduced by these tools.

6

6.6.2. SEARCH AS LEARNING

LLMs can be used for educational purposes, providing explanations and answers to questions and asking adjunct questions to learners to assess their current knowledge state. Thus, learners can use them for active learning, sensemaking, and goal-setting during the SAL process. Essential questions in this context are *how* and *when* to incorporate LLMs as a learning tool during the SAL process. Furthermore, with these tools, the learning process and search can be *personalised* based on the searcher's learning capabilities and prior knowledge. A learner can use the system prompt to provide their current knowledge on a topic, which can be used to tailor answers according to their need. However, this requires the learners to clearly articulate their current knowledge state and understand what they do not know—but this is often not the case [312]. Hence, in an ideal system, the knowledge

⁶ChatGPT reached one million users five days upon its release

⁷<https://www.microsoft.com/en-us/edge/features/bing-chat?>

state of a user must be inferred from the conversation. Systems like ChatGPT offer a crude solution to this—they keep track of the information provided during a conversation as context and use it to answer questions posed by the user. The question of how to effectively track the learning progress of a student through their online interaction with teaching materials has been addressed in the domain of **knowledge tracing** [1, 97, 210]. However, how to employ LLMs to explicitly trace the knowledge state of learners based on their interactions is still an open question. Liu et al. [164] used GPT-2 to ask programming questions to learners and estimated their knowledge state from the submitted code. Several open questions in this regard include: Can we use LLM-generated questions to estimate user knowledge state in other domains as well? How and when can those questions be generated? How do we represent the knowledge state of a user? How do we modify answers based on the inferred knowledge state?

While these tools have the advantage that they can free up the cognitive load of learners by simplifying complex topics during a SAL process, wrong information can also end up being counterproductive to the learning process. For example, LLMs can automatically summarise a document that a learner reads, rendering it unnecessary for them to take notes. However, it needs to be examined to what extent automation of processes, like note-taking, assists learners during the SAL process. It has been shown in education literature that active learning methods like note-taking are more beneficial as compared to providing the students with lecture slides [137, 154] or just listening to lectures [207, 239]. Hence, when provided with LLM-generated notes, can learners still comprehend texts and make sense of them in a similar/better way as they would have done if they were taking the notes (e.g.,) themselves? Furthermore, what happens if there are factual mistakes in the summaries? Users tend to overestimate the truthfulness of *direct* answers given by a system (such as those by ChatGPT) [219]. Are learners able to detect them? Users who are satisfied with the answers provided by such generative models can be tempted to skip significant parts of the actual reading process. How does that impact their learning outcomes? To what extent and when do learners end up learning/internalising the mistakes in factual information in those generated answers? Another critical question, especially when users try to search on topics with multiple perspectives (e.g., whether schools should mandate uniforms), is how to deploy these tools during the search process such that they foster critical thinking and prevent confirmation bias or filter bubbles.

6.6.3. RELEVANCE JUDGEMENT COLLECTION

Lastly, developing LLMs has opened up new possibilities (and concerns) for collecting relevance judgements [86] of text documents. A decade ago, employing crowdworkers to collect relevant judgements helped us scale up the costly annotation process of collecting them from experts. The scaling up led to the creation of more extensive test collections (e.g.,MSMARCO [196], DBpedia-Entity v2 [109], Hotpot QA[271, 310] etc.) with the compromise in the quality of judgements. With LLMs, history may repeat itself: they open up the possibility of annotating a massive amount of data. However, what about annotation quality? Recent work showed that LLMs could be effective, with accuracy as good as human labellers [275]. Additionally, collecting relevance judgements using LLMs is faster than collecting them using humans; they can judge the relevance of documents without being affected by documents they have seen before (contrary to crowdworkers [245] and with no

boredom or tiredness. Furthermore, they have much more information on a topic than a human and can process multiple languages, which will also help us create multilingual corpora. This raises the question: Can LLMs entirely replace humans in the judgment process? For what tasks should human assessors not be replaced by machines [70, 119]? However, we must also acknowledge how we must understand the potential limitations and negative externalities of using generative models for automated relevance label generation. Important open questions include—if and how biases in LLMs [70, 119] may also manifest in relevance labelling? For example, a biased model can underestimate the relevance of longer documents [116]. This bias might manifest more systemically when relevance labels are solicited from these models rather than crowdworkers. Do LLM-generated relevance labels show bias towards ranking models that themselves employ LLMs? Does optimising towards LLM-based labels make us susceptible to the risk of falling into the trap of overfitting to the peculiarities of the LLM rather than towards improving actual relevance? LLM-specific quality assurance methods will need to be developed, and research needs to be conducted on how to facilitate collaboration between humans and generative models about this task—how to employ LLMs, as well as other generative models, to aid human assessors in devising reliable judgments while enhancing the efficiency of the process?

6.7. FINAL REMARKS

The insights gained from these studies hold practical significance for IR system developers, user interface designers, and researchers in IIR. The findings can guide the development of innovative interfaces, interactive tools, and personalised user experiences, fostering a more effective information retrieval process. Additionally, our research opens avenues for further investigations, such as exploring novel cognitive-driven interactive approaches, integrating machine learning techniques into IIR, and evaluating IIR in emerging technologies like LLMs and user contexts.

BIBLIOGRAPHY

REFERENCES

- [1] G. Abdelrahman, Q. Wang, and B. Nunes. 2023. Knowledge tracing: A survey. *Comput. Surveys* 55, 11 (2023), 1–37.
- [2] A.L. Al-Harbi and M. Smucker. 2014. A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th information interaction in context symposium*. 195–204.
- [3] A. Al-Maskari and M. Sanderson. 2011. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management* 47, 5 (2011), 719–729.
- [4] A.O. Alanazi, M. Sanderson, Z. Bao, and J. Kim. 2020. The impact of ad quality and position on Mobile SERPs. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 318–322.
- [5] M. Aliannejadi, M. Harvey, L. Costa, M. Pointon, and F. Crestani. 2019. Understanding mobile search task relevance and user behaviour in context. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 143–151.
- [6] G. Allen, M. Beijen, D. Maxwell, and U. Gadiraju. 2023. In a Hurry: How Time Constraints and the Presentation of Web Search Results Affect User Behaviour and Experience. In *International Conference on Web Engineering*. Springer, 221–235.
- [7] M. Almeida, M. Bilal, A. Finamore, I. Leontiadis, Y. Grunenberger, M. Varvello, and J. Blackburn. 2018. Chimp: Crowdsourcing human inputs for mobile phones. In *Proceedings of the 2018 World Wide Web Conference*. 45–54.
- [8] O. Alonso and S. Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Vol. 15. 16.
- [9] M. Alsayasneh, S. Amer-Yahia, E. Gaussier, V. Leroy, J. Pilourdault, R.M. Borromeo, M. Toyama, and J.M. Renders. 2017. Personalized and diverse task composition in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* 30, 1 (2017), 128–141.
- [10] L.W. Anderson and D.R. Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

- [11] J. Anderton, M. Bashir, V. Pavlu, and J.A. Aslam. 2013. An analysis of crowd workers mistakes for specific and complex relevance assessment task. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1873–1876.
- [12] I. Arapakis, S. Park, and M. Pielot. 2021. Impact of response latency on user behaviour in mobile web search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 279–283.
- [13] J. Arguello and R. Capra. 2014. The effects of vertical rank and border on aggregated search coherence and search behavior. In *Proceedings of the 23rd acm international conference on conference on information and knowledge management*. 539–548.
- [14] J. Arguello and B. Choi. 2019. The effects of working memory, perceptual speed, and inhibition in aggregated search. *ACM Transactions on Information Systems (TOIS)* 37, 3 (2019), 1–34.
- [15] J. Arguello, F. Diaz, J. Callan, and B. Carterette. 2011. A methodology for evaluating aggregated search results. In *European Conference on Information Retrieval*. Springer, 141–152.
- [16] J. Arguello, W.C. Wu, D. Kelly, and A. Edwards. 2012. Task complexity, vertical display and user interaction in aggregated search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 435–444.
- [17] R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics* 34, 4 (2008), 555–596.
- [18] A.H. Awadallah, R.W. White, P. Pantel, S.T. Dumais, and Y.-M. Wang. 2014. Supporting complex search tasks. In *Proc. 23rd ACM CIKM*. 829–838.
- [19] L. Azzopardi. 2011. The economics in interactive information retrieval. In *Proc. 34th ACM SIGIR*. 15–24.
- [20] L. Azzopardi. 2014. Modelling interaction with economic models of search. In *Proc. 37th ACM SIGIR*. 3–12.
- [21] L. Azzopardi. 2014. Modelling Interaction with Economic Models of Search. In *Proc. 37th ACM SIGIR (SIGIR '14)*. 3–12.
- [22] L. Azzopardi, D. Kelly, and K. Brennan. 2013. How query cost affects search behavior. In *Proc. 36th ACM SIGIR*. 23–32.
- [23] L. Azzopardi, D. Maxwell, M. Halvey, and C. Hauff. 2023. Driven to Distraction: Examining the Influence of Distractors on Search Behaviours, Performance and Experience. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 83–94.
- [24] L. Azzopardi, P. Thomas, and N. Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st international acm sigir conference on research & development in information retrieval*. 605–614.

- [25] L. Azzopardi and G. Zuccon. 2015. An analysis of theories of search and search behavior. In *Proc. 1st ACM ICTIR*. 81–90.
- [26] L. Azzopardi and G. Zuccon. 2016. An analysis of the cost and benefit of search interactions. In *Proc 2nd ACM ICTIR*. 59–68.
- [27] L. Azzopardi and G. Zuccon. 2016. Two scrolls or one click: A cost model for browsing search results. In *Advances in Information Retrieval*. 696–702.
- [28] L. Azzopardi and G. Zuccon. 2018. Economics models of interaction: a tutorial on modeling interaction using economics. (2018).
- [29] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 667–674.
- [30] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A test collection with query variability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 725–728.
- [31] K. Balog and C. Zhai. 2023. User simulation for evaluating information access systems. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 302–305.
- [32] J. Banks. 2005. *Discrete event system simulation*. Pearson Education India.
- [33] F. Baskaya, H. Keskustalo, and K. Järvelin. 2012. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 105–114.
- [34] M.J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.
- [35] A. Bauer and K. Koedinger. 2006. Pasting and encoding: Note-taking in online courses. In *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*. IEEE, 789–793.
- [36] N. J. Belkin. 1990. The Cognitive Viewpoint in Information Science. *J. Inf. Sci.* 16, 1 (Jan. 1990), 11–15.
- [37] G. Ben-Yehudah and Y. Eshet-Alkalai. 2018. The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *J. Edu. Multimedia & Hypermedia* 27, 2 (2018), 153–178.
- [38] T. Berners-Lee, R. Cailliau, A. Luotonen, Henrik F. Nielsen, and A. Secret. 1994. The world-wide web. *Commun. ACM* 37, 8 (1994), 76–82.
- [39] K. Bharat. 2000. SearchPad: Explicit capture of search context to support web search. *Computer Networks* 33, 1-6 (2000), 493–501.

- [40] N. Bhattacharya and J. Gwizdka. 2018. Relating eye-tracking measures with changes in knowledge on search tasks. In *Proc. 10th ACM ETRA*. 1–5.
- [41] N. Bhattacharya and J. Gwizdka. 2019. Measuring learning during search: differences in interactions, eye-gaze, and semantic similarity to expert knowledge. In *Proc. 4th ACM CHIIR*. 63–71.
- [42] J.P. Biddix, C.J. Chung, and H.W. Park. 2011. Convenience or credibility? A study of college student online research behaviors. *Internet & Higher Education* 14, 3 (2011), 175–182.
- [43] P. Borlund. 2002. Evaluation of interactive information retrieval systems. (2002).
- [44] P. Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research* 8, 3 (2003), 8–3.
- [45] H. Bota, K. Zhou, and J.M. Jose. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 acm on conference on human information interaction and retrieval*. 131–140.
- [46] J.R. Boyle. 2011. Thinking strategically to record notes in content classes. *American Secondary Education* (2011), 51–66.
- [47] M. Bron, J. Van Gorp, F. Nack, L.B. Baltussen, and M. de Rijke. 2013. Aggregated search interface preferences in multi-session search tasks. In *Proc. 36th ACM SIGIR*. 123–132.
- [48] S.W. Brown. 2008. The Reference Interview: Theories and Practice. (2008).
- [49] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [50] G. Buscher, R.W. White, S. Dumais, and J. Huang. 2012. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 373–382.
- [51] J. Callan, J. Allan, C.L. Clarke, S. Dumais, D.A. Evans, M. Sanderson, and C. Zhai. 2007. Meeting of the MINDS: an information retrieval research agenda. In *ACM SIGIR Forum*, Vol. 41. ACM New York, NY, USA, 25–34.
- [52] A. Câmara, D. Maxwell, and C. Hauff. 2022. Searching, learning, and subtopic ordering: A simulation-based analysis. In *European Conference on Information Retrieval*. Springer, 142–156.
- [53] M.F. Camporro and N. Marquardt. 2020. Live Sketchnoting Across Platforms: Exploring the Potential and Limitations of Analogue and Digital Tools. In *Proc. 38th ACM CHI*. 1–12.

- [54] R. Capra, G. Marchionini, J. Velasco-Martin, and K. Muller. 2010. Tools-at-hand and learning in multi-session, collaborative search. In *Proc. 28th ACM CHI*. 951–960.
- [55] S. K. Card, T. P. Moran, and A. Newell. 1980. The keystroke-level model for user performance time with interactive systems. *Commun. ACM* 23, 7 (1980), 396–410.
- [56] J. Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22, 2 (1996), 249–254.
- [57] B. Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proc. 34th ACM SIGIR*. 903–912.
- [58] P. Chandar, W. Webber, and B. Carterette. 2013. Document features predicting assessor disagreement. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 745–748.
- [59] C. Chavula, Y. Choi, and S.Y. Rieh. 2023. SearchIdea: An Idea Generation Tool to Support Creativity in Academic Search. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 161–171.
- [60] M.C. Chen, J.R. Anderson, and M.H. Sohn. 2001. What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing. In *CHI'01 extended abstracts on Human factors in computing systems*. 281–282.
- [61] E.H. Chi, P. Pirolli, K. Chen, and J. Pitkow. 2001. Using Info. Scent to Model User info. Needs and Actions and the Web. In *Proc. 19th ACM CHI*. 490–497.
- [62] B. Choi, R. Capra, and J. Arguello. 2019. The effects of working memory during search tasks of varying complexity. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 261–265.
- [63] A. Chuklin and M. de Rijke. 2016. Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In *Proceedings of the 25th acm international on conference on information and knowledge management*. 175–184.
- [64] A. Chuklin, A. Severyn, J.R. Trippas, E. Alfonseca, H. Silen, and D. Spina. 2019. Using audio transformations to improve comprehension in voice question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 164–170.
- [65] C. Cleverdon, J. Mills, and M. Keen. 1966. Factors Determining the Performance of Indexing Systems Volume 1. Design. (1966).
- [66] P. Clough and M. Sanderson. 2013. Evaluating the performance of information retrieval systems using test collections. (2013).
- [67] K. Collins-Thompson, P. Hansen, and C. Hauff. 2017. Search as learning (dagstuhl seminar 17092). In *Dagstuhl reports*, Vol. 7.

- [68] K. Collins-Thompson, S.Y. Rieh, C.C. Haynes, and R. Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 163–172.
- [69] H.G. Colt, M. Davoudi, S. Murgu, and N.Z. Rohani. 2011. Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical endoscopy* 25, 1 (2011), 207–216.
- [70] C.B. Marta R. Costa-juss and N. Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *GeBNLP 2019* (2019), 33.
- [71] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos. 2021. Overview of the TREC 2020 deep learning track. *arXiv preprint arXiv:2102.07662* (2021).
- [72] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and E.M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).
- [73] A. Câmara, N. Roy, D. Maxwell, and C. Hauff. 2021. Searching to Learn with Instructional Scaffolding. In *Proc. 6th ACM CHIIR*.
- [74] T.T. Damessie, F. Scholer, and J.S. Culpepper. 2016. The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proceedings of the 21st Australasian Document Computing Symposium*. 41–48.
- [75] R. Datta, D. Joshi, J. Li, and J.Z. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)* 40, 2 (2008), 1–60.
- [76] D. Davidson. 1977. The effect of individual differences of cognitive style on judgments of document relevance. *Journal of the American Society for information Science* 28, 5 (1977), 273–284.
- [77] D. Demaree, H. Jarodzka, S. Brand-Gruwel, and Y. Kammerer. 2020. 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*. 373–377.
- [78] A. Diamond. 2013. Executive functions. *Annual review of psychology* 64 (2013), 135.
- [79] Jerry Dischler. 2015. Building for the next moment. *Inside Ad-Words* (2015).
- [80] D. Donato, F. Bonchi, T. Chi, and Y. Maarek. 2010. Do you want to take notes? Identifying research missions in Yahoo! Search Pad. In *Proc. 19th WWW*. 321–330.
- [81] S.T. Dumais, G. Buscher, and E. Cutrell. 2010. Individual differences in gaze patterns for web search. In *Proceedings of the third symposium on Information interaction in context*. 185–194.
- [82] S. T. Dumais, G. Buscher, and E. Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *Proc. 3rd IIiX*. 185–194.

- [83] J. Dunlosky, K.A. Rawson, E.J. Marsh, M.J. Nathan, and D.T. Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Sci. in the Public Interest* 14, 1 (2013), 4–58.
- [84] C. Eickhoff, J. Teevan, R. White, and S. Dumais. 2014. Lessons from the journey: a query log analysis of within-session learning. In *Proc. 7th ACM WSDM*. 223–232.
- [85] D. Ellis. 1993. Modeling the Information-Seeking Patterns of Academic Researchers: A Grounded Theory Approach. *The Library Quarterly: Information, Community, Policy* 63, 4 (1993), 469–486.
- [86] G. Faggioli, L/ Dietz, C. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*. 39–50.
- [87] W. Fass and G.M. Schumacher. 1978. Effects of motivation, subject activity, and readability on the retention of prose materials. *J. Educational Psychology* 70, 5 (1978), 803.
- [88] F. Faul, E. Erdfelder, A.G. Lang, and A. Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.
- [89] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (May 2007), 175–191.
- [90] P. M. Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *J. experimental psychology* 47, 6 (1954), 381.
- [91] X. Fu, E. Yilmaz, and A. Lipani. 2022. Evaluating the Cranfield Paradigm for Conversational Search Systems. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. 275–280.
- [92] N. Fuhr. 2008. A probability ranking principle for interactive information retrieval. *Information Retrieval* 11, 3 (2008), 251–265.
- [93] M. Fyfield, M. Henderson, and M. Phillips. 2021. Navigating four billion videos: teacher search strategies and the YouTube algorithm. *Learning, Media and Technology* 46, 1 (2021), 47–59.
- [94] U. Gadipati, S. Möller, M. Nöllenburg, D. Saupe, S. Egger-Lampl, D. Archambault, and B. Fisher. 2017. Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, 6–26.
- [95] U. Gadipati, R. Yu, S. Dietze, and P. Holtz. 2018. Analyzing knowledge gain of users in informational search sessions on the web. In *Proc. 3rd ACM CHIIR*. 2–11.

- [96] S.M. Gass, J.N. Behney, and B. Uzum. 2013. Inhibitory control, working memory and L2 interaction. In *Psycholinguistic and sociolinguistic perspectives on second language learning and teaching*. Springer, 91–114.
- [97] A. Ghosh, N. Heffernan, and A.S. Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2330–2339.
- [98] R. Gligorov, M. Hildebrand, J. Van Ossenbruggen, L. Aroyo, and G. Schreiber. 2013. An evaluation of labelling-game data for video retrieval. In *European Conference on Information Retrieval*. Springer, 50–61.
- [99] G. Golovchinsky, A. Diriye, and T. Dunnigan. 2012. The future is in the past: designing for exploratory search. In *Proc. 4th IIiX*. 52–61.
- [100] S. Gomes, M. Boon, and O. Hoeber. 2022. A study of cross-session cross-device search within an academic digital library. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 384–394.
- [101] S. Gordon-Salant and S.S. Cole. 2016. Effects of age and working memory capacity on speech recognition performance in noise among listeners with normal hearing. *Ear and hearing* 37, 5 (2016), 593–602.
- [102] W. Gritz, C. Otto, A. Hoppe, G. Pardi, Y. Kammerer, and R. Ewerth. 2023. Comparing Interface Layouts for the Presentation of Multimodal Search Results. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 321–327.
- [103] Q. Guo and E. Agichtein. 2010. Towards predicting web searcher gaze position from mouse movements. In *Proc. 28th ACM CHI*. 3601–3606.
- [104] J. Gwizdka. 2010. Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology* 61, 11 (2010), 2167–2187.
- [105] Å.M. Hagen, J. Braasch, and I. Bråten. 2014. Relationships between spontaneous note-taking, self-reported strategies and comprehension when reading multiple texts in different task conditions. *Journal of Research in Reading* 37, S1 (2014), S141–S157.
- [106] K. Hagerty. 1967. Abstracts as a Basis for Relevance Judgment. (1967).
- [107] D.K. Harman. 1993. Overview of the first TREC conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. 36–47.
- [108] S.G. Hart and L.E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- [109] F. Hasibi, F. Nikolaev, C. Xiong, K. Balog, S. E. Bratsberg, A. Kotov, and J. Callan. 2017. DBpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1265–1268.

- [110] M. Hearst. 2009. *Search user interfaces*. Cambridge University Press.
- [111] M.R. Herrera, E.S. de Moura, M. Cristo, T.P. Silva, and A.S. da Silva. 2010. Exploring features for the automatic identification of user goals in web search. *Information processing & management* 46, 2 (2010), 131–142.
- [112] M. Hertzum. 2021. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. *Ergonomics* 64, 7 (2021), 869–878.
- [113] D. Hettiachchi, Z. Sarsenbayeva, F. Allison, N. van Berkel, T. Dingler, G. Marini, V. Kostakos, and J. Goncalves. 2020. "Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [114] A. Hinderks, M. Schrepp, and J. Thomaschewski. 2018. A Benchmark for the Short Version of the User Experience Questionnaire.. In *WEBIST*. 373–377.
- [115] G. Hinesley, M. H. Blackmon, and M. J. Carnot. 2008. The Importance of Graphics: Implications for Educational Hypertext Material. In *EdMedia+ Innovate Learning*. 1412–1421.
- [116] S. Hofstätter, H. Zamani, B. Mitra, N. Craswell, and A. Hanbury. 2020. Local self-attention over long text for efficient document retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021–2024.
- [117] M. Hosseini, I.J. Cox, N. Milić-Frayling, G. Kazai, and V. Vinay. 2012. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings* 34. Springer, 182–194.
- [118] M.J. Howe. 1970. Using students' notes to examine the role of the individual learner in acquiring meaningful subject matter. *The Journal of Educational Research* 64, 2 (1970), 61–63.
- [119] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5491–5501.
- [120] P. Ingwersen and K. Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.
- [121] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.
- [122] J. Jiang, D. He, and J. Allan. 2014. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 607–616.

- [123] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, Vol. 51. Acm New York, NY, USA, 4–11.
- [124] D.D. Johnson and B. von Hoff Johnson. 1986. Highlighting vocabulary in inferential comprehension instruction. *J. Reading* 29, 7 (1986), 622–625.
- [125] K.S. Jones and C.J. van Rijsbergen. 1976. Information retrieval test collections. *Journal of documentation* (1976).
- [126] H.J. Jung and M. Lease. 2011. Improving consensus accuracy via z-score and weighted voting. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.
- [127] R. Kalyani and U. Gadiraju. 2019. Understanding User Search Behavior Across Varying Cognitive Levels. In *Proc. 30th ACM HT*. 123–132.
- [128] Y. Kammerer and P. Gerjets. 2014. The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *International Journal of Human-Computer Interaction* 30, 3 (2014), 177–191.
- [129] M.J. Kane and R.W. Engle. 2003. Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of experimental psychology: General* 132, 1 (2003), 47.
- [130] S. Karanam, H. van Oostendorp, M. C. P. Melguizo, and B. Indurkhya. 2013. Interaction of textual and graphical information in locating web page widgets. *Behaviour & Information Technology* 32, 5 (2013), 503–515.
- [131] S. Karanam, J. Viswanathan, A. Theertha, B. Indurkhya, and H. Van Oostendorp. 2010. Impact of placing icons next to hyperlinks on information-retrieval tasks on the web. In *Proc. 32nd CogSci*, Vol. 32.
- [132] N. Karousos, C. Katsanos, N. Tselios, and M. Xenos. 2013. Effortless tool-based evaluation of web form filling tasks using keystroke level model and fitts law. In *Proc. 31st ACM CHI*. 1851–1856.
- [133] A. Kashyap, V. Hristidis, and M. Petropoulos. 2010. Facetor: cost-driven exploration of faceted query results. In *Proc. 19th ACM CIKM*. 719–728.
- [134] G. Kazai, J. Kamps, and N. Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval* 16 (2013), 138–178.
- [135] D. Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.
- [136] D. Kieras et al. 2001. Using the keystroke-level model to estimate execution times. *University of Michigan* 555 (2001).

- [137] H. Kim. 2018. Impact of slide-based lectures on undergraduate students' learning: Mixed effects of accessibility to slides, differences in note-taking, and memory term. *Computers & Education* 123 (2018), 13–25.
- [138] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.J. Yoon. 2016. Understanding eye movements on mobile devices for better presentation of search results. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2607–2619.
- [139] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. 2017. What snippet size is needed in mobile web search?. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 97–106.
- [140] D. Kirsh. 2010. Thinking with external representations. *AI & society* 25, 4 (2010), 441–454.
- [141] J.A. Konstan, B. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl. 1997. GroupLens: Applying collaborative filtering to usenet news. *Commun. ACM* 40, 3 (1997), 77–87.
- [142] B. Koopman and G. Zuccon. 2014. Relevation! An open source system for information retrieval relevance assessment. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 1243–1244.
- [143] C. C. Kuhlthau. 1988. Developing a Model of the Library Search Process: Cognitive and Affective Aspects. *RQ* 28, 2 (1988), 232–242.
- [144] M. Kutlu, T. McDonnell, T. Elsayed, and M. Lease. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research* 69 (2020), 143–189.
- [145] D. Lagun, C.H. Hsieh, D. Webster, and V. Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 113–122.
- [146] J. Lai and N. Yankelovich. 2006. Speech interface design. (2006).
- [147] T. Langerak, S. Christen, M. Albaba, C. Gebhardt, and O. Hilliges. 2022. MARLUI: Multi-Agent Reinforcement Learning for Goal-Agnostic Adaptive UIs. *arXiv preprint arXiv:2209.12660* (2022).
- [148] K.V Lau, P. Farooque, G. Leydon, M.L. Schwartz, R.M. Sadler, and J.J. Moeller. 2018. Using learning analytics to evaluate a video-based lecture series. *Medical teacher* 40, 1 (2018), 91–98.
- [149] B. Laugwitz, T. Held, and M. Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symp. Austrian HCI & usability eng.* 63–76.

- [150] T. Lauterman and R. Ackerman. 2014. Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior* 35 (2014), 455–463.
- [151] H. Lee, J. Lee, K. Makara, B. J Fishman, and Y. Hong. 2015. Does higher education foster critical and creative learners? An exploration of two universities in South Korea and the USA. *Higher Education Research & Development* 34, 1 (2015), 131–146.
- [152] J. Lee, S.S. Rodriguez, R. Natarajan, J. Chen, H. Deep, and A. Kirlik. 2021. What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 512–520.
- [153] U. Lee, Z. Liu, and J. Cho. 2005. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*. 391–400.
- [154] S.P. León and I. García-Martínez. 2021. Impact of the provision of PowerPoint slides on learning. *Computers & Education* 173 (2021), 104283.
- [155] D. Leutner, C. Leopold, and V. den Elzen-Rump. 2007. Self-regulated learning with a text-highlighting strategy. *J. Psychology* 215, 3 (2007), 174–182.
- [156] D. Leutner, C. Leopold, and E. Sumfleth. 2009. Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior* 25, 2 (2009), 284–289.
- [157] O. Levi, I. Guy, F. Raiber, and O. Kurland. 2018. Selective cluster presentation on the search results page. *ACM Transactions on Information Systems (TOIS)* 36, 3 (2018), 1–42.
- [158] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [159] C.J. Lin and S. Ho. 2020. Prediction of the use of mobile device interfaces in the progressive aging process with the model of Fitts' law. *J. Biomedical Informatics* 107 (2020), 103457.
- [160] C. Liu and X. Song. 2018. How do Information Source Selection Strategies Influence Users' Learning Outcomes'. In *Proc. 3rd ACM CHIIR*. 257–260.
- [161] C. Liu, C.L. Yang, J.J. Williams, and H.-C. Wang. 2019. NoteStruct: Scaffolding Note-taking while Learning from Online Videos. In *Proc. 37th ACM CHI*. 1–6.
- [162] H. Liu, C. Liu, and N.J. Belkin. 2019. Investigation of users' knowledge change process in learning-related search tasks. *Proc. ASIS&T* 56, 1 (2019), 166–175.
- [163] J. Liu, N.J. Belkin, X. Zhang, and X. Yuan. 2013. Examining users' knowledge change in the task completion process. *IP&M* 49, 5 (2013), 1058–1074.

- [164] N. Liu, Z. Wang, R.G. Baraniuk, and A. Lan. 2022. GPT-based Open-Ended Knowledge Tracing. *arXiv preprint arXiv:2203.03716* (2022).
- [165] Z. Liu, Y. Liu, K. Zhou, M. Zhang, and S. Ma. 2015. Influence of vertical result in web search examination. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*. 193–202.
- [166] F. Loumakis, S. Stumpf, and D. Grayson. 2011. This Image Smells Good: Effects of Image Information Scent in Search Engine Results Pages. In *Proc. 20th ACM CIKM*. 475–484.
- [167] H. Luan, Y.T. Zheng, M. Wang, and T.S. Chua. 2011. VisionGo: towards video retrieval with joint exploration of human and computer. *Information Sciences* 181, 19 (2011), 4197–4213.
- [168] C. Lustig, C.P. May, and L. Hasher. 2001. Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General* 130, 2 (2001), 199.
- [169] A. MacFarlane, A. Albrair, C.R. Marshall, and G. Buchanan. 2012. Phonological working memory impacts on information searching: An investigation of dyslexia. In *Proceedings of the 4th Information Interaction in Context Symposium*. 27–34.
- [170] I.S. MacKenzie and W. Buxton. 1992. Extending Fitts' law to two-dimensional tasks. In *Proc. 10th ACM CHI*. 219–226.
- [171] I. Mackie, J. Dalton, and A. Yates. 2021. How deep is your learning: the DL-HARD annotated deep learning dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2335–2341.
- [172] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 1–32.
- [173] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen. 2021. Generation-Augmented Retrieval for Open-Domain Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4089–4100.
- [174] G. Marchionini. 2006. Exploratory search: from finding to understanding. *Comm. ACM* 49, 4 (2006), 41–46.
- [175] C.C. Marshall. 1997. Annotation: from paper books to the digital library. In *Proc. 2nd ACM Conf. Digital Libraries*. 131–140.
- [176] D. Maxwell. 2019. *Modelling search and stopping in interactive information retrieval*. Ph.D. Dissertation. University of Glasgow.

- [177] D. Maxwell and L. Azzopardi. 2014. Stuck in traffic: How temporal delays affect search behaviour. In *Proc. 5th IIiX*. 155–164.
- [178] D. Maxwell and L. Azzopardi. 2016. Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM international conference on information and knowledge management*. 731–740.
- [179] D. Maxwell and L. Azzopardi. 2016. 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*. 1141–1144.
- [180] D. Maxwell and L. Azzopardi. 2018. Information scent, searching and stopping. In *Advances in Information Retrieval*. 210–222.
- [181] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proceedings of the 40th ACM SIGIR*. 135–144.
- [182] D. Maxwell and C. Hauff. 2021. LogUI: Contemporary Logging Infrastructure for Web-Based Experiments. In *Advances in Information Retrieval*. 525–530.
- [183] R.E. Mayer, E. Griffith, I. Jurkowitz, and D. Rothman. 2008. Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. *J. Exp. Psychology: Applied* 14, 4 (2008), 329.
- [184] M. McGregor, L. Azzopardi, and M. Halvey. 2021. Untangling cost, effort, and load in information seeking and retrieval. In *Proceedings of the 2021 CHIR*. 151–161.
- [185] J. Mendel. 2010. *The effect of interface consistency and cognitive load on user performance in an information search task*. Ph.D. Dissertation. Clemson University.
- [186] A. Michalovich and A. Hershkovitz. 2020. Assessing YouTube science news' credibility: The impact of web-search on the role of video, source, and user attributes. *Public Understanding of Science* 29, 4 (2020), 376–391.
- [187] A. Moffat, F. Scholer, and P. Thomas. 2012. Models and metrics: IR evaluation as a user process. In *Proc. 17th ADCS*. 47–54.
- [188] A. Moffat, P. Thomas, and F. Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 659–668.
- [189] F. Moraes, S.R. Putra, and C. Hauff. 2018. Contrasting Search as a Learning Activity with Instructor-designed Learning. In *Proc. 27th ACM CIKM*. 167–176.
- [190] H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C.E. Kahn, and W. Hersh. 2010. Overview of the CLEF 2009 medical image retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 72–84.

- [191] C. Murad and C. Munteanu. 2020. Designing voice interfaces: Back to the (curriculum) basics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [192] C. Murad, C. Munteanu, L. Clark, and B.R. Cowan. 2018. Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 269–276.
- [193] A. Namoun. 2018. Three column website layout vs. grid website layout: An eye tracking study. In *International Conference of Design, User Experience, and Usability*. Springer, 271–284.
- [194] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. 2013. 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*. 953–964.
- [195] D.C. Ngo, L.S. Teo, and J.G. Byrne. 2003. Modelling interface aesthetics. *Information Sciences* 152 (2003), 25–46.
- [196] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.
- [197] D. Nicholas, I. Rowlands, D. Clark, and P. Williams. 2011. Google Generation II: web behaviour experiments with the BBC. *ASLIB proceedings* 63, 1 (2011), 28–45.
- [198] S.L. Nist and M.C. Hogrebe. 1987. The role of underlining and annotating in remembering textual information. *Literacy Research & Instruction* 27, 1 (1987), 12–25.
- [199] E. Norman and B. Furnes. 2016. The relationship between metacognitive experiences and learning: Is there a difference between digital and non-digital study media? *Computers in Human Behavior* 54 (2016), 301–309.
- [200] C. Nowacki, A. Gordeeva, and A.H. Lizé. 2020. Improving the usability of voice user interfaces: a new set of ergonomic criteria. In *International Conference on Human-Computer Interaction*. Springer, 117–133.
- [201] P.A. Nye, T.J. Crooks, M. Powley, and G. Tripp. 1984. Student note-taking related to university examination performance. *Higher Education* 13, 1 (1984), 85–97.
- [202] H.L. O'Brien, A. Kampen, A.W. Cole, and K. Brennan. 2020. The Role of Domain Knowledge in Search as Learning. In *Proc. 5th ACM CHIIR*. 313–317.
- [203] H.L. O'Brien and E.G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69.
- [204] K. Ong, K. Järvelin, M. Sanderson, and F. Scholer. 2018. Qwerty: The effects of typing on web search behavior. In *Proc. 3rd ACM CHIIR*. 281–284.

- [205] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [206] P. Organisciak, J. Teevan, S. Dumais, R. Miller, and A. Kalai. 2014. A crowd of your own: Crowdsourcing for on-demand personalization. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 2. 192–200.
- [207] H. Özçakmak. 2019. Impact of Note Taking during Reading and during Listening on Comprehension. *Educational Research and Reviews* 14, 16 (2019), 580–589.
- [208] H.L. O’Brien. 2010. The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with computers* 22, 5 (2010), 344–352.
- [209] S. Palani, Y. Zhou, S. Zhu, and S.P. Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [210] S. Pandey and G. Karypis. 2019. A Self-Attentive Model for Knowledge Tracing. *International Educational Data Mining Society* (2019).
- [211] M.H. Payandeh, M. Boon, D. Storie, V. Ramshaw, and O. Hoeber. 2023. Drag-and-Drop Query Refinement and Query History Visualization for Mobile Exploratory Search. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 432–437.
- [212] R. Pearson and P. Van Schaik. 2003. The effect of spatial layout of and link colour in web pages on performance in a visual search task and an interactive search task. *Intl. J. Human-Computer Studies* 59, 3 (2003), 327–353.
- [213] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.
- [214] D. Persico and K. Steffens. 2017. Self-regulated learning in technology enhanced learning environments. In *Tech. Enhanced Learning*. 115–126.
- [215] R. Piening, K. Pfeuffer, A. Esteves, T. Mittermeier, S. Prange, P. Schröder, and F. Alt. 2021. Looking for info: Evaluation of gaze based information retrieval in augmented reality. In *Human-Computer Interaction—INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part I* 18. Springer, 544–565.
- [216] P. Pirolli. 2007. *Information foraging theory: Adaptive interaction with information*. Oxford University Press.
- [217] P. Pirolli and S.K. Card. 1999. Information Foraging. *Psychological Review* 106 (1999), 643–675.
- [218] H.R. Ponce and R.E. Mayer. 2014. An eye movement analysis of highlighting and graphic organizer study aids for learning from expository text. *Computers in human behavior* 41 (2014), 21–32.

- [219] M. Potthast, M. Hagen, and B. Stein. 2021. The dilemma of the direct answer. In *Acm sigir forum*, Vol. 54. ACM New York, NY, USA, 1–12.
- [220] M. Pressley. 2000. What should comprehension instruction be the instruction of? *Handbook of reading research* 3 (2000), 545–561.
- [221] S.R. Putra, K. Grashoff, F. Moraes, and C. Hauff. 2018. On the Development of a Collaborative Search System.. In *DESIREs*. 76–82.
- [222] S.R. Putra, F. Moraes, and C. Hauff. 2018. SearchX: Empowering Collaborative Search Research. In *Proc. 41st ACM SIGIR*. 1265–1268.
- [223] Z. Qiyang and H. Jung. 2019. Learning and sharing creative skills with short videos: A case study of user behavior in tiktok and bilibili. In *Int. Assoc. Soc. Des. Res. Conf.* 25–50.
- [224] S.M. Randhawa, T. Ahmad, J. Chen, and A.A. Raza. 2021. Karamad: A Voice-based Crowdsourcing Platform for Underserved Populations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [225] S. Robertson. 2008. On the history of evaluation in IR. *Journal of Information Science* 34, 4 (2008), 439–456.
- [226] K. Rodden and X. Fu. 2007. Exploring how mouse movements relate to eye movements on web search results pages. *Web Information Seeking and Interaction* (2007), 29–32.
- [227] K. Rodden, X. Fu, A. Aula, and I. Spiro. 2008. Eye-mouse coordination patterns on web search results pages. In *CHI'08 extended abstracts on Human factors in computing systems*. 2997–3002.
- [228] K. Roitero, E. Maddalena, S. Mizzaro, and F. Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Information Processing & Management* 58, 6 (2021), 102688.
- [229] I. Rowlands, D. Nicholas, P. Williams, P. Huntington, M. Fieldhouse, B. Gunter, R. Withey, H.R. Jamali, T. Dobrowolski, and C. Tenopir. 2008. The Google generation: the information behaviour of the researcher of the future. In *Aslib proceedings*, Vol. 60. Emerald Group Publishing Limited, 290–310.
- [230] N. Roy, A. Balayn, D. Maxwell, and C. Hauff. 2023. Hear Me Out: A Study on the Use of the Voice Modality for Crowdsourced Relevance Assessments. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 718–728.
- [231] N. Roy, A. Câmara, D. Maxwell, and C. Hauff. 2021. Incorporating widget positioning in interaction models of search behaviour. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 53–62.
- [232] N. Roy, D. Maxwell, and C. Hauff. 2022. Users and Contemporary SERPs: A (Re-) Investigation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2765–2775.

- [233] N. Roy, F. Moraes, and C. Hauff. 2020. Exploring Users' Learning Gains within Search Sessions. In *Proc. 5th ACM CHIIR*. 432–436.
- [234] N. Roy, M.V. Torre, U. Gadiraju, D. Maxwell, and C. Hauff. 2021. How Do Active Reading Strategies Affect Learning Outcomes in Web Search?. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II* 43. Springer, 368–375.
- [235] N. Roy, M.V. Torre, U. Gadiraju, D. Maxwell, and C. Hauff. 2021. 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*. 229–238.
- [236] M. Rudner, T. Lunner, T. Behrens, E.S. Thorén, and J. Rönnberg. 2012. Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology* 23, 08 (2012), 577–589.
- [237] I. Ruthven. 2008. Interactive information retrieval. *Annual review of information science and technology* 42 (2008), 43–92.
- [238] T. Sakai and Z. Zeng. 2020. Retrieval evaluation measures that agree with users' SERP preferences: Traditional, preference-based, and diversity measures. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2020), 1–35.
- [239] I. Salame and A. Thompson. 2020. Students' Views on Strategic Note-Taking and Its Impact on Performance, Achievement, and Learning. *International Journal of Instruction* 13, 2 (2020), 1–16.
- [240] S. Salimzadeh, D. Maxwell, and C. Hauff. 2021. The Impact of Entity Cards on Learning-Oriented Search Tasks. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 63–72.
- [241] T. Saracevic. 1969. Comparative effects of titles, abstracts and full texts on relevance judgments. *Proceedings of the American Society for Information Science* 6, 1 (1969), 293–299.
- [242] T. Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American society for information science and technology* 58, 13 (2007), 1915–1933.
- [243] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. 2010. "Your word is my command": Google search by voice: A case study. In *Advances in speech recognition*. Springer, 61–90.
- [244] M. Schleußinger. 2021. Information retrieval interfaces in virtual reality—A scoping review focused on current generation technology. *Plos one* 16, 2 (2021), e0246398.
- [245] F. Scholer, D. Kelly, W.C. Wu, H.S. Lee, and W. Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 623–632.

- [246] F. Scholer, A. Turpin, and M. Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 1063–1072.
- [247] S. Schultheiß. 2023. 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*. 475–478.
- [248] B. Sguerra, M. Baranes, R. Hennequin, and M. Moussallam. 2022. Navigational, Informational or Punk-Rock? An Exploration of Search Intent in the Musical Domain. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 202–211.
- [249] Y. Shao, J. Mao, Y. Liu, M. Zhang, and S. Ma. 2022. From linear to non-linear: investigating the effects of right-rail results on complex SERPs. *Advances in Computational Intelligence* 2, 1 (2022), 1–16.
- [250] J.L. Shefelbine. 1990. Student factors related to variability in learning word meanings from context. *J. Reading Behavior* 22, 1 (1990), 71–97.
- [251] J. Sherwani, D. Yu, T. Paek, M. Czerwinski, Y.C. Ju, and A. Acero. 2007. Voicepedia: Towards speech-based access to unstructured information. In *Eighth Annual Conference of the International Speech Communication Association*.
- [252] M. Moritaand Y. Shinoda. 2012. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval*. 272–281.
- [253] B. Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (2000), 63–65.
- [254] A. Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.
- [255] A. Singhal, G. Salton, M. Mitra, and C. Buckley. 1996. Document length normalization. *Information Processing & Management* 32, 5 (1996), 619–633.
- [256] C. Siu and B.S. Chaparro. 2014. First look: Examining the horizontal grid layout using eye-tracking. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 1119–1123.
- [257] M.D. Smucker and C.L. Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 95–104.
- [258] M.D. Smucker, G. Kazai, and M. Lease. 2012. *Overview of the trec 2012 crowdsourcing track*. Technical Report. Sch. of Info., Uiv. Texas Austin.

- [259] S. Song, Y.C. Zhao, X. Yao, Z. Ba, and Q. Zhu. 2021. Short video apps as a health information source: an investigation of affordances, user experience and users' intention to continue the use of TikTok. *Internet Research* 31, 6 (2021), 2120–2142.
- [260] X. Song, C Liu, and H. Liu. 2018. Characterizing and exploring users' task completion process at different stages in learning related tasks. *Proceedings of the Association for Information Science and Technology* 55, 1 (2018), 460–469.
- [261] V. Stenbäck. 2016. *Speech masking speech in everyday communication: The role of inhibitory control and working memory capacity*. Vol. 1559. Linköping University Electronic Press.
- [262] V. Stenbäck, E. Marsja, M. Hällgren, B. Lyxell, and B. Larsby. 2021. The contribution of age, working memory capacity, and inhibitory control on speech recognition in noise in young and older adult listeners. *Journal of Speech, Language, and Hearing Research* 64, 11 (2021), 4513–4523.
- [263] J.R. Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology* 18, 6 (1935), 643.
- [264] S. Suh, B. Min, S. Palani, and H. Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.
- [265] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. 2010. Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 519–528.
- [266] C. Sutter and M. Ziefle. 2004. Psychomotor efficiency in users of notebook input devices: Confirmation and restrictions of Fitts' law as an evaluative tool for user-friendly design. In *Proc. 48th HFES*, Vol. 48. SAGE Publications Sage CA: Los Angeles, CA, 773–777.
- [267] R. Syed and K. Collins-Thompson. 2017. Retrieval algorithms optimized for human learning. In *Proc. 40th ACM SIGIR*. 555–564.
- [268] R. Syed and K. Collins-Thompson. 2018. Exploring Document Retrieval Features Associated with Improved Short-and Long-term Vocabulary Learning Outcomes. In *Proc. 3rd ACM CHIIR*. 191–200.
- [269] R. Syed, K. Collins-Thompson, P.N. Bennett, M. Teng, S. Williams, W.W. Tay, and S. Iqbal. 2020. Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation. In *Proc. 29th WWW*. 1693–1703.
- [270] F. P. Tamborello and M. D. Byrne. 2005. Information search: the intersection of visual and semantic space. In *Proc. 23rd ACM CHI*. 1821–1824.
- [271] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

- [272] P. Thomas, A. Moffat, P. Bailey, and F. Scholer. 2014. Modeling decision points in user search behavior. In *Proceedings of the 5th information interaction in context symposium*. 239–242.
- [273] P. Thomas, A. Moffat, P. Bailey, F. Scholer, and N. Craswell. 2018. Better effectiveness metrics for serps, cards, and rankings. In *Proceedings of the 23rd australasian document computing symposium*. 1–8.
- [274] P. Thomas, F. Scholer, and A. Moffat. 2013. What users do: The eyes have it. In *Asia information retrieval symposium*. Springer, 416–427.
- [275] P. Thomas, S. Spielman, N. Craswell, and B. Mitra. 2023. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621* (2023).
- [276] T. Tombros and F. Crestani. 1999. A study of users’ perception of relevance of spoken documents. *Rapport technique TR-99-013, Berkeley, CA* (1999).
- [277] H. Touvron, T. Lavigil, G. Izacard, X. Martinet, M.A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [278] J.R. Trippas. 2021. Spoken conversational search: audio-only interactive information retrieval. In *ACM SIGIR Forum*, Vol. 53. ACM New York, NY, USA, 106–107.
- [279] J.R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. 2015. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*. 991–994.
- [280] M.L. Turner and R.W. Engle. 1989. Is working memory capacity task dependent? *Journal of memory and language* 28, 2 (1989), 127–154.
- [281] M. Turunen, J. Hakulinen, N. Rajput, and A. Nanavati. 2012. Evaluation of mobile and pervasive speech applications. *Speech in Mobile and Pervasive Environments* (2012), 219–262.
- [282] K. Urgo and J. Arguello. 2022. Capturing Self-Regulated Learning During Search. In *Workshop on Investigating Learning During Web Search*.
- [283] K. Urgo and J. Arguello. 2022. Understanding the “pathway” towards a searcher’s learning objective. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2022), 1–43.
- [284] K. Urgo and J. Arguello. 2023. Goal-setting in support of learning during search: An exploration of learning outcomes and searcher perceptions. *Information Processing & Management* 60, 2 (2023), 103158.
- [285] K. Urgo, J. Arguello, and R. Capra. 2019. 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*. 117–124.

- [286] K. Urgo, J. Arguello, and R. Capra. 2020. The Effects of Learning Objectives on Searchers' Perceptions and Behaviors. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 77–84.
- [287] R. Vaish, K. Wyngarden, J. Chen, B. Cheung, and M.S. Bernstein. 2014. Twitch crowdsourcing: crowd contributions in short bursts of time. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3645–3654.
- [288] H. R. Varian. 1999. Economics and Search. *SIGIR Forum* 33, 1 (Sept. 1999), 1–5.
- [289] A. Vashistha, P. Sethi, and R. Anderson. 2017. Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 1855–1866.
- [290] A. Vashistha, P. Sethi, and R. Anderson. 2018. BSpeak: An accessible voice-based crowdsourcing marketplace for low-income blind people. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [291] M. Verma and E. Yilmaz. 2017. Search Costs vs. User Satisfaction on Mobile. In *Advances in Information Retrieval*. 698–704.
- [292] E.M. Voorhees. 2006. Overview of the TREC 2005 Robust Retrieval Track.. In *Proceedings of TREC-14*.
- [293] A. Vtyurina, C. Clarke, E. Law, J. R Trippas, and H. Bota. 2020. A mixed-method analysis of text and audio search interfaces with varying task complexity. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 61–68.
- [294] J. Wachtler, M. Khalil, B. Taraghi, and M. Ebner. 2016. On Using Learning Analytics to Track the Activity of Interactive MOOC Videos.. In *SE@ VBL@ LAK*. 8–17.
- [295] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. 2013. Incorporating vertical results into search click models. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 503–512.
- [296] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma. 2014. An eye-tracking study of website complexity from cognitive load perspective. *Decision support systems* 62 (2014), 1–10.
- [297] S. Wang, D.S. Unal, and E. Walker. 2019. MindDot: Supporting Effective Cognitive Behaviors in Concept Map-Based Learning Environments. In *Proc. 38th ACM CHI*. 1–14.
- [298] M. Wesche and T.S. Paribakht. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Lang. Review* 53, 1 (1996), 13–40.
- [299] J. Wiley and A.F. Jarosz. 2012. How working memory capacity affects problem solving. In *Psychology of learning and motivation*. Vol. 56. Elsevier, 185–227.
- [300] L.O. Wilson. 2016. Anderson and Krathwohl–Bloom's taxonomy revised. *Understanding the New Version of Bloom's Taxonomy* (2016).

- [301] M.J. Wilson and M.L. Wilson. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *JASIST* 64, 2 (2013), 291–306.
- [302] T. Wilson. 2000. Human Information Behavior. *Informing Science* 3 (01 2000), 49–55.
- [303] W. Wu, D. Kelly, and A. Sud. 2014. Using information scent and need for cognition to understand online search behavior. In *Proc. 37th ACM SIGIR*. 557–566.
- [304] Z. Wu, M. Sanderson, B.B. Cambazoglu, W.B. Croft, and F. Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1635–1644.
- [305] X. Xie, Y. Liu, X. Wang, M. Wang, Z. Wu, Y. Wu, M. Zhang, and S. Ma. 2017. Investigating examination behavior of image search users. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 275–284.
- [306] C. Xu, Z. Li, H. Zhang, A. Rathore, H. Li, C. Song, K. Wang, and W. Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 14–26.
- [307] J. Xu, C. Chen, G. Xu, H. Li, and E.R.T Abib. 2010. Improving quality of training data for learning to rank using click-through data. In *Proceedings of the third ACM international conference on Web search and data mining*. 171–180.
- [308] L. Xu, X. Zhou, and U. Gadiraju. 2020. How Does Team Composition Affect Knowledge Gain of Users in Collaborative Web Search?. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 91–100.
- [309] F. Yang, S. Kalloori, R. Chalumattu, and M. Gross. 2022. Personalized information retrieval for touristic attractions in augmented reality. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1613–1616.
- [310] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C.D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. <https://doi.org/10.18653/v1/D18-1259>
- [311] N. Yankelovich and J. Lai. 1998. Designing speech user interfaces. In *CHI 98 Conference Summary on Human Factors in Computing Systems*. 131–132.
- [312] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, and S. Dietze. 2018. Predicting user knowledge gain in informational search sessions. In *Proc. 41st ACM SIGIR*. 75–84.
- [313] A. Yuan and Y. Li. 2020. Modeling human visual search performance on realistic webpages using analytical and deep learning methods. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

- [314] C.L. Yue, B.C. Storm, N. Kornell, and E.L. Bjork. 2015. Highlighting and its relation to distributed study and students' metacognitive beliefs. *Edu. Psy. Review* 27, 1 (2015), 69–78.
- [315] F. Zhang, J. Mao, Y. Liu, X. Xie, W. Ma, M. Zhang, and S. Ma. 2020. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*. 379–388.
- [316] X. S. Zheng, I. Chakraborty, J. J.W. Lin, and R. Rauschenberger. 2009. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *Proc 26th ACM CHI*. 1–10.
- [317] G. K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.
- [318] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J.M. Jose, and L. Azzopardi. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval* 16, 2 (2013), 267–305.
- [319] K. Zyskowski, M.R Morris, J.P. Bigham, M.L. Gray, and S.K. Kane. 2015. Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1682–1693.

LIST OF FIGURES

1.1	Model of how users interact according to the Cranfield paradigm	2
1.2	A model of a simplified version of the IIR process inspired from the works of [176, 272]. Depending on the specific search system, available search widgets and the task at hand there will be other decision points or actions taken by the user.	3
1.3	Evolution of SERP over time (for the query snowboarding). Top-left: In early 2002, search engines presented results in the <i>ten blue links</i> format. Top-right: Earlier in the last decade, results from different verticals like images, videos, news etc. started to be interleaved with the web results. The SERP renderings of 2002 and 2012 are obtained from www.theoldweb.com . Bottom-left: Widgets like entity cards, direct answers, maps etc., were introduced towards the second half of the last decade. Bottom-right: An example from Bing-Chat, LLM-powered conversational agents are considered to be the future of web search. 2020 and 2023 SERPs are screenshots taken from Google and Bing chat in November 2023.	5
2.1	Examples of both the (a) list-based and (b) grid-based interfaces trialled. Note the inclusion of links for the separate <i>All</i> (as shown), <i>Images</i> , and <i>Videos</i> result pages. Heterogeneous content is displayed in red boxes, and is <i>not</i> present in the two homogeneous content interface conditions (SG and SL). Circled numbers correspond to the narrative of Section 2.3.1.	18
2.2	Interaction plots, showing effects of SERP types and task complexity over: (a) clicking on web results; (b) the mean session duration (in seconds); and (c) clicks on images presented on the SERP. ⁸	27
2.3	Distribution of ranks of the first clicked web results for participants over both grid-based interfaces SL and HL (a), and list-based interfaces SG and HG (b).	29
3.1	The SearchX interface as used for this study with annotations—refer to §3.3.1. This screenshot is an amalgamation of what would have been seen over all experimental conditions; refer to §3.4.1 for details.	38
3.2	Examples of the two new widgets introduced to SearchX for this study. (a) On the left is the document view, complete with text highlighting capabilities. (b) On the right is the note-taking widget, which is visible when Notepad is clicked. Note that these features were not available to all participants of the study; refer to §3.4.1 for more information.	40
3.3	Overview of the study’s workflow.	41

3.4	Average #highlights and average #seconds of note-taking activity in each five minute interval. The number on top of each bar shows the % of participants with 1+ highlights or 1+ seconds note-taking activity in that interval.	51
4.1	Example cost functions for two different SERP scenarios, adapted from Azopardi and Zuccon [27]. Hypotheses can be derived from them, e.g., <i>the optimal number of results (n per page) to show to maximise a user's benefit in the violet scenario is at the cost function's global minimum.</i>	60
4.2	Flowchart of the modelled querying process. Before issuing each query, the user is presented with the choice of inspecting the QHW or typing in	62
4.3	The screenshot shows the QHW interface used for this study. Note the inclusion of the QHW in the callout—this was positioned in one of the areas as shown with blue boxes. Refer to §4.5.2 for information on the circled interface components.	67
4.4	Overview of QHW vs. QB usage when reissuing queries of varying lengths. Shown here are the results over 590 reissued queries across all participants/- conditions.	71
5.1	An overview of the user study protocol, including approximate times for participants to complete each component. Refer to §5.3.1 for mappings to the letters highlighting key aspects of the study procedure.	82
5.2	Composition screenshot of both the text and voice interfaces used by participants for judging query-passage pairs. Circled numbers correspond to the same in the narrative, found in §5.3.4.	86
5.3	Accuracy of relevance judgements per label category for both text and voice . Diagonals represent percentage of time the true labels were <i>correctly</i> predicted by participants. Here, R = RELEVANT, SR = SOMEWHAT-RELEVANT, NR = NON-RELEVANT and IDK = <i>I do not know</i>	90
5.4	The trend of voice participants judging relevance w.r.t. time taken for passages of various length: (a) % of time participants listened to the entire audio clip; and (b) at what point was relevance judged (as a % of audio clip length).	93

LIST OF TABLES

2.1	Overview of information needs and their type. The rightmost column shows the most popular query obtained from our query selection pilot study, outlined in Section 2.3.3. Numbers in parentheses indicate how many crowdworkers ($n = 25$) submitted the most popular query variation.	21
2.2	Results of a factorial mixed ANOVA, where interface is between-subjects, and task is within-subjects variable. A ✓ indicates significant effect ($p < 0.05$) on the particular user interaction and ✗ indicates no significance.	24
2.3	User interactions for different interfaces across all tasks. † indicates that there is a significant main effect of SERP layout on that particular user interaction. HG , HL , SG , SL indicate significant difference with HG , HL , SG and SL respectively. Maximum values for each interaction is highlighted in bold . Rows VII-X indicate interactions on SERP. r.p. is short for results page.	26
2.4	User interactions for different task complexity across all search interfaces. † indicates that there is a significant main effect of task complexity on that particular user interaction. N , R , U , A indicate significant difference with navigational, remember, understand and analyse tasks. Maximum values for each interaction is highlighted in bold . Rows VII-X indicate interactions on SERP.	28
3.1	The five hypotheses and rationalisations used for this exploratory study.	35
3.2	The number of participants exploring each topic in our study, together with related statistics. Two-way ANOVA tests revealed no significant differences in average number of queries between topics ($F(1, 107) = 1.83, p = 0.07$). ± indicates the standard deviation.	44
3.3	Example annotation of facts and subtopics and the computation of F-Fact and T-Depth. Note that sentences demonstrating knowledge of the topic are colour coded—each colour pertains to an individual subtopic (see the <i>T-Depth</i> column).	46
3.4	Mean (\pm standard deviations) of RPL and search behaviour metrics across all participants in each condition. A dagger (†) denotes two-way ANOVA significance, while C , H , N , B indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) over the four conditions CONTROL , HIGH , NOTE and HIGH+NOTE respectively.	49
3.5	H1: Learners are divided into <i>pro-highlighters</i> , <i>unsure</i> or <i>anti-highlighters</i> . † indicates two-way ANOVA significance, while C , H , B indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) with Holm-Bonferroni correction.	53

3.6	H3, H4: Learners are divided into two groups (<i>heavy</i> and <i>light</i>) based on the median values for each active reading strategy. The learning metrics are computed separately for each group. The significant differences obtained from TukeyHSD pairwise test are highlighted in bold	54
3.7	H5: Participants are divided into two groups (<i>trained</i> and <i>non-trained</i>) based on their self reported highlighting and note-taking frequency and based on their education level. The learning metrics are computed separately for each group. The significant differences obtained from TukeyHSD pairwise tests are in bold	55
4.1	The placement of (as well as the number of) text columns, and the number of entries in the <i>Related Searches</i> widget across ten different web search engines. Results retrieved on May 2 nd , 2021 for the query chess . <i>Placement</i> corresponds to the widget's position within the SERP.	59
4.2	Overview of the model's constants and values used.	65
4.3	Mean (\pm standard deviation) of search behaviour metrics across all participants in each variation of <i>QHW</i> . A dagger (\dagger) denotes one-way ANOVA significance, \ddagger denotes χ^2 significance, while U, L, B, M, T indicate post-hoc significance ($p < 0.05$ with Bonferroni correction) over conditions UPPER-RIGHT, LOWER-RIGHT, BOTTOM-LEFT, MIDDLE-LEFT and TOP-LEFT respectively.	70
5.1	Examples of <i>Query/Passage (Q/P)</i> pairs for different passage length categories. The (Qid) is taken from the TREC datasets. We also provide links to [audio] clips of the respective passages.	84
5.2	Overview of passage length buckets. Averages are reported together with the standard deviation.	85
5.3	RQ1: Effect of modality of passage presentation on accuracy of relevance judgement, time taken per judgement in seconds and perceived workload (IV-VIII) per participant. We also report Krippendorff's α and Cohen's κ for accuracy. \dagger indicates significant difference in between the two conditions according to independent sample t-test. $*$ indicates the corresponding metric is equivalent for both conditions based on the TOST procedure.	89
5.4	RQ2: Effects of passage length and presentation modality on accuracy of relevance judgements (with Krippendorff's α , Cohen's κ) and time taken. A <u>bold</u> number indicates that the metric for the corresponding presentation modality is significantly more than that for the other modality for the particular passage length. xs,s,m,l,xl indicates significant difference (within the same experimental condition) compared to XS, S, M, L, XL passage lengths. $*$ indicates equivalence between the two conditions.	91
5.5	RQ3: Summary of main effects of <i>Presentation Modality (PM)</i> , <i>Working Memory (WM)</i> , <i>Inhibition (IN)</i> , and effects of the interaction of WM and IN with PM on accuracy of relevance judgement, time taken, and perceived workload. A \checkmark indicates significant effect of a 3-way ANOVA test ($p < 0.05$) on the particular dependent variables and \times indicates no significant effect.	94

CURRICULUM VITÆ

Experience and Education

2024-04–Present	NLP Scientist at OKRA AI.
2023-11–2024-02	Research internship at Amazon.
2022-7–10	Research internship at Amazon.
2021-6–10	Research internship at Amazon.
2017–2019	M.Sc. in Computer Science, TU Delft.
2016–2017	Technology Consultant, PwC India.
2012–2016	B.E. in Electrical Engineering, Jadavpur University, Kolkata.

Publications

ECIR 2024	Is Interpretable Machine Learning Effective at Feature Selection for Neural Learning-to-Rank? Lijun Lyu, Nirmal Roy , Harrie Oosterhuis, Avishek Anand
CHIIR 2024	On the Effects of Automatically Generated Adjunct Questions for Search as Learning Peide Zhu, Arthur Câmara, Nirmal Roy , David Maxwell, Claudia Hauff
SIGIR 2023	📄 Hear Me Out: A Study on the Use of the Voice Modality for Crowdsourced Relevance Assessments Nirmal Roy , Agathe Balayn, David Maxwell, Claudia Hauff
ECIR 2023	Viewpoint Diversity in Search Results Tim Draws, Nirmal Roy , Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, Nava Tintarev
SIGIR 2022	📄 Users and Contemporary SERPs: A (Re-) Investigation Nirmal Roy , David Maxwell, Claudia Hauff
ICTIR 2021	📄 Incorporating Widget Positioning in Interaction Models of Search Behaviour Nirmal Roy , Arthur Câmara, David Maxwell, Claudia Hauff

- ECIR 2021  How Do Active Reading Strategies Affect Learning Outcomes in Web Search?
Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, Claudia Hauff
- CHIIR 2021  Searching to learn with instructional scaffolding
Arthur Câmara, **Nirmal Roy**, David Maxwell, Claudia Hauff
- CHIIR 2021  Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment
Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, Claudia Hauff
- CHIIR 2020 Exploring Users' Learning Gains within Search Sessions
Nirmal Roy, Felipe Moraes, Claudia Hauff

 Included in this thesis.

 Won award.

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 (following are all the dissertations since 2016).

- 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 Knobbe (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 Céller (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 Bilin-gual 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 Joosse (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 Retrievability 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 Maaike 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 Boses (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 Smetsers (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 Slootmaker (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 Heinerman (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 Dumitrasche (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

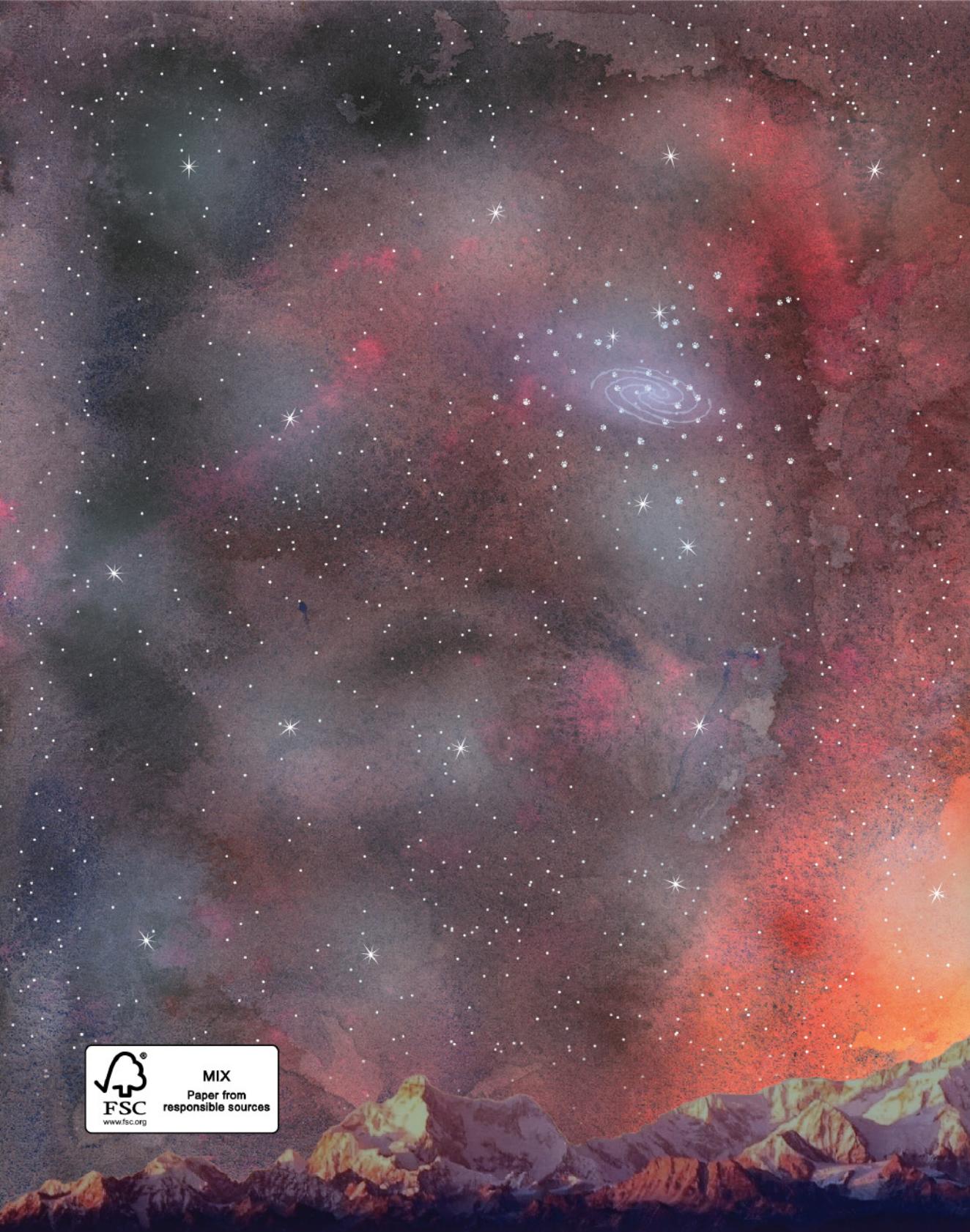
-
- 27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
- 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
- 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
- 30 Bob Zadok Blok (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 Mercuur (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 Gijsbers (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 Tragano (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 Čausević (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 Shojaifar (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 withdrawn
- 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
- 16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
- 17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
- 18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
- 19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
- 20 Ritsart Anne Plantenga (UL), Omgang met Regels
- 21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
- 22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
- 23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
- 24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour



Dm 2024 Babyn