Models and Algorithms for Understanding and Supporting Learning Goals in Information Retrieval

by

Rohail Mustafa Syed

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2020

Doctoral Committee:

Associate Professor Kevyn Collins-Thompson, Chair
Dr. Paul N. Bennett, Microsoft Research AI
Professor Qiaozhu Mei
Professor Rada Mihalcea

Rohail Mustafa Syed

rmsyed@umich.edu

ORCID iD: 0000-0003-3504-2975

Acknowledgements

I am very grateful to my advisor, Kevyn Collins-Thompson, for being a great mentor throughout my time at the University, providing invaluable guidance and support along the way. I would additionally like to thank my dissertation committee members - Paul N. Bennett, Qiaozhu Mei, and Rada Mihalcea for serving on my committee, providing interesting insights and helpful critiques that have improved this dissertation.

My experience at the School of Information opened me up to fresh perspectives and insights, and substantially changed my way of thinking about the world, about research, and about problem-solving. I am very thankful that I got to meet, have great discussions with, and learn from my colleagues especially Sungjin Nam, Ryan Burton, and Heeryung Choi over the years.

I extend my deepest appreciation to my parents and family. They have not only been a great source of support and guidance for me during my time at the University but have long before inculcated in me a deep sense of intellectual curiosity, a love for continuous learning, and the creativity to think in new and unexpected ways.

Finally, I'd like to thank the institutions that awarded funding to make this research possible. This work was supported in part by Dept. of Education grant R305A140647 to

the University of Michigan. This work was also supported in part by the Michigan Institute for Data Science. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsors.

Table of Contents

List of Tables

List of Figures

Abstract

While search technology is widely used for learning-oriented information needs, the results provided by popular services such as Web search engines are optimized primarily for generic relevance, not effective learning outcomes. As a result, the typical information trail that a user must follow while searching to achieve a learning goal may be an inefficient one, possibly involving unnecessarily difficult content, or material that is irrelevant to actual learning progress relative to a user's existing knowledge. My work addresses these problems through multiple studies where various models and frameworks are developed and tested to support particular dimensions of search as learning. Empirical analysis of these studies through user studies demonstrate promising results and provide a solid foundation for further work.

The earliest work we focused on centered on developing a framework and algorithms to support vocabulary learning objectives in a Web document context. The proposed framework incorporates user information, topic information and effort constraints to provide a desirable combination of personalized and efficient (by word length) learning experience. Our user studies demonstrate the effectiveness of our framework against a strong commercial baseline's (Google search) results in both short- and long-term assessment.

While topic-specific content features (such as frequency of subtopic occurrences) naturally play a role in influencing learning outcomes, stylistic and structural features of the documents themselves may also play a role. Using such features we construct robust regression models that show strong predictive strength for multiple measures of learning outcomes. We also show early evidence that regression models trained on one dataset of search as learning can

show strong test-set predictions on an independent dataset of search as learning, suggesting a certain degree of generalizability of stylistic content features. The models developed in my work are designed to be as generalizable, scalable and efficient as possible to make it easier for practitioners in the field to improve how people use search engines for learning. Finally, we investigate how gaze-tracking and automatic question generation could be used to scale a form of active learning to arbitrary text material. Our results show promising potential for incorporating interactive learning experiences in arbitrary text documents on the Web. A major theme in these studies centers on understanding and improving how people learn when using Web search engines. We also put specific emphasis on long-term learning outcomes and demonstrate that our models and frameworks actually yield sustainable knowledge gains, both for passive and interactive learning. Taken together, these research studies provide a solid foundation for multiple promising directions in exploring search as learning.

Chapter 1

Introduction

As more people use the internet for learning purposes (De Rosa, 2006; Griffiths and Brophy, 2005; NetDay, 2004; Ng and Gunstone, 2002; Rainie and Hitlin, 2009; Syed, Collins-Thompson, Bennett, Teng, Williams, Tay, and Iqbal, 2020), there is a need to develop intelligent systems that can optimize the educational experience for such users. While there has been significant progress in developing effective Intelligent Tutoring Systems (ITS) (Koedinger, Anderson, Hadley, and Mark, 1997) and Web search algorithms personalized for individual users (Collins-Thompson, Bennett, White, de la Chica, and Sontag, 2011; Tan, Gabrilovich, and Pang, 2012), there has been little work in combining the two concepts. Such a hybrid search system would have the potential to yield significant improvements in the search as learning process. The hybrid system would have the advantages of the scalability, familiarity and the ubiquity of general Web search as well as the advantages of a model that personalizes selection of resources through the lens of a cognitive model of expected learning outcomes. The principal focus of most studies I've completed in this dissertation center on developing and understanding how such a system should be defined and how effective it actually is in improving learning outcomes.

Recent work in the information retrieval space has focused on the application of traditional Web search for educational information seeking tasks (Collins-Thompson, Rieh, Haynes, and Syed, 2016). Many prior studies have shown that Web search is an increasingly

common starting point for users engaging in search tasks designed for learning or discovering more about particular topics (Abualsaud, 2017; Bailey, Chen, Grosenick, Jiang, Li, Reinholdtsen, Salada, Wang, and Wong, 2012; De Rosa, 2006). Given the large-scale nature of Web search engines, both in terms of content and users, there has been heightened attention towards determining what strategies people are using in Web search to learn, what types of search retrieval algorithms result in better learning outcomes (Collins-Thompson et al., 2016)(Collins-Thompson, Hansen, and Hauff, 2017) and what type of retrieval frameworks can be designed to accommodate personalized learning experiences (Collins-Thompson and Callan, 2004) at the scale of general Web search (Syed and Collins-Thompson, 2017b). The focus of this dissertation is on constructing and investigating Web search frameworks and algorithms that facilitate exploratory search intents of an educational nature.

Past studies have shown that when people use the Web for starting an exploratory information seeking task, they often start with search engines. An OCLC study found that 89% of college students and 84% of all people used Web search engines to initiate their "search for information on a particular topic" (De Rosa, 2006). If students fail to find appropriate or helpful documents at the earliest stage of searching, they may be discouraged or unmotivated to continue. Such a scenario could lead many students interested in learning to abandon search tasks due to mismatches between what the student was expecting and what the search system returned. It is therefore important to develop algorithms at this earliest stage of web-based educational inquiry to optimize document selection for learning outcomes. I will demonstrate through the literature review (Chapter 3) that there exists a significant gap in existing work pertaining to specifically designing Web search systems that optimize for an individual's learning objectives, especially personalized systems.

The second part of this paper (Chapters 4 - 12) will be focused on developing a class

of algorithms that are optimized to offer documents that will help a particular individual as well as generic users maximize their learning outcomes along with a study designed to evaluate the effectiveness of such algorithms. Past literature has demonstrated that in Web search users can often lose focus or interest if they are unable to satisfy their goals within the first SERP page of results. This emphasizes the importance of taking into account this limited effort users are willing to expend and the consequent importance of choosing high quality documents that collectively fully cover the material the user needs to know. We will propose such an algorithm that incorporates into its retrieval objective parameters that reward better coverage of the topic's aspects, maximize document quality and penalize reading effort. We then investigate how we can build a data-driven model to learn what document features are strong predictors of learning gains. We demonstrate that our model can show generalized predictive power across multiple independent studies, topics, assessment types and assessment platforms.

The completed studies described in this dissertation form a multi-part research objective aimed at understanding and constructing models of information retrieval that consider optimal learning utility as the end-goal of the user. In addition to the passive objective of document selection, we also investigate interactive interventions that support better learning in documents. These studies are described in chronological order in Chapter 4 to help the reader see the gradual progression of these studies towards this goal.

In total, the completed studies present a compelling retrieval model and sets of regression models for estimating what types of documents are generally better suited for learning goals using high-level document features as predictors. The following are the high-level research questions I will address in this dissertation:

RQ1: Can we apply a model of domain-specific user knowledge state that updates

based on what Web documents they read? Does such a model improve learning outcomes? (Chapters 5 and 6)

RQ2: Can we develop an information retrieval framework that explicitly uses estimated user knowledge gain as its optimization objective? Can such a model outperform a commercial baseline? (Chapter 6)

RQ3: Are there document, user or document set features that are good predictors of knowledge state and knowledge gain in a Web documents context? (Chapters 8 and 9)

RQ4: Can automatic question generation be used to scale the adjunct questions effect to support scalable active learning in Web documents? (In this dissertation, we refer to *active learning* in the pedagogical context not the machine learning context) (Chapter 10)

RQ5: How do learning outcomes differ in the Web context when considering short- vs long-term assessment? Are there user-specific or context-specific factors that influence short- or long-term results? (Chapters 7 and 10)

As discussed, there has been increasing focus on the intersection of Web search and learning but there is a strong lack of principled approaches to personalizing Web search for an individual's learning outcomes. This is further emphasized by the complete absence of any longitudinal studies assessing robust (long-term) learning resulting from the use of specially-designed retrieval algorithms. By better understanding how people interact with general Web documents in the course of an educational learning task, we can better understand what document-level and user-level features are best suited for improving learning outcomes. This

will allow for tremendous benefit for those seeking a free, scalable solution for self-directed and self-paced learning.

Chapter 2

Background

I will split the literature review for this dissertation into a general background review followed by a more specific literature review relating more closely to our studies. The bulk of the research in this dissertation centers on the intersection of information retrieval and education. Related work on this specific area will be explored in depth in Chapter 3. In this chapter, I will go over existing work on how people seek and make sense of information, how different forms and complexities of learning can be categorized and how a person's knowledge can be assessed. These are critical background areas that should directly inform the design of any study seeking to model and optimize resource selection for learning goals. Of particular importance to this dissertation is the first section of this chapter: the Bloom's taxonomy of learning (Section 2.1.1) and the Item Response Theory (IRT) (Section 2.1.2).

2.1 Categorizing, Modeling and Assessing Knowledge

In this section, we focus specifically on knowledge itself, how it can be categorized, represented and evaluated. We will start by looking at how knowledge can be categorized in terms of levels of complexity (e.g. ranging from simple recall to the ability to synthesize new ideas on the topic). Then we will discuss methods that have been used to algorithmically model an individual's current knowledge state as a function of what learning resources they have

been exposed to and the nature of their interactions. Finally, we will discuss methods for evaluating an individual's knowledge state.

2.1.1 Taxonomies of Learning

In the previous section, we looked at one way of breaking down learning: short-term vs long-term. In this section, we consider another breakdown of *types* of learning along dimensions of the cognitive complexity of the learning task using the well-established revision to the Bloom's taxonomy. The well-documented Bloom's taxonomy (Bloom, 1956) and its revision by Krathwohl (2002), suggest that learning can be split into three domains including the cognitive, affective and psychomotor domains. Of particular focus in this study is the cognitive domain of learning. In the revised Bloom's taxonomy (Krathwohl, 2002), this consists of six levels that reflect different forms and complexities of learning, from fact-based recall (remember) to concept-based construction (create). The six levels are organized in terms of the complexity involved in the learning required. In the ideal case, learning should be considered complete when a student is capable of demonstrating proficiency in each of these six dimensions. However, due to the complexity involved in each of these levels and due to the difficulty in constructing a single solution to deal with a very multifaceted problem, we will attempt to tackle these levels one at a time. As there are no prior works we are aware of that have optimized Web search algorithms for personalized educational goals, we will begin at the lowest levels of cognitive complexity and in future work, gradually focus on higher levels on the basis of the results of this work (Figure 2.1). In particular, of focus in this paper is the "Remember" dimension of the taxonomy.

Work by Wilson and Wilson (2013) demonstrated an early approach to operationalizing the revised taxonomy in the form of three separate measures for evaluating different types

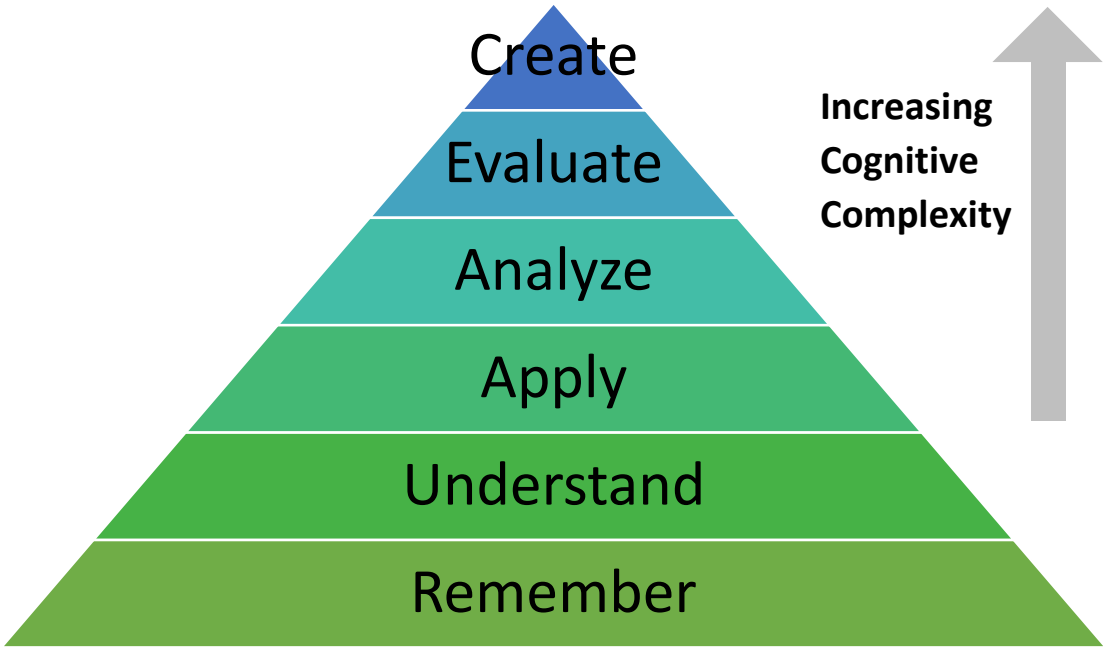

Figure 2.1: Revised Bloom's Taxonomy (Krathwohl, 2002)

of learning. These measures and their links to the taxonomy were: D-Qual (related to "understanding"), D-Intrp (related to "applying") and D-Crit (related to "evaluating") (Wilson and Wilson, 2013). The authors tested these measures in a lab study that involved a pre-task summary, a learning task and a post-task summary. They showed that their measures were valid at differentiating between pre-task and post-task summaries but only D-Qual was able to distinguish between low and high self-reported prior knowledge for a given summary. The authors also showed that the length of the summaries affected the significance of these variables where longer summaries typically yielded better differences and longer summaries showed statistical differences from both the D-Qual and D-Intrp variables. While the results of the study were not conclusive enough to warrant fully adopting their model as a one-size-fits-all solution, it does offer initial insights as to how Krathwohl's taxonomy could be

evaluated.

2.1.2 Knowledge Prediction and Representation

In constructing a search algorithm that will optimize for learning intents, we need to first make assumptions to model how a user learns as they read. Extensive literature has focused on the concept of Bayesian Knowledge Tracing (BKT) which, in its essence considers the student's knowledge to perform Bayesian updates in response to new information the student receives and what information they correctly and incorrectly recall (Corbett and Anderson, 1994). While this concept has been used in developing various Intelligent Tutoring Systems (Koedinger et al., 1997), it has also been proven to be an effective way of modeling Web-based learning (Pirolli and Kairam, 2013). Recently, Zhu (2013) proposed a machine teaching framework for Bayesian learners which could optimally determine the number of instances of each subtopic a learner would need to read about to have fully learned about the subject. Zhu's work also incorporates considerations of how much effort a learner will expend in the learning process.

Another well-established model of assessing learning is the Item Response Theory (IRT) (Junker, 1999; Syed and Collins-Thompson, 2017b). The theory posits that a learner's ability to correctly answer a dichotomous question ($Y_i = \{0, 1\}$) is a function of their latent abilities, often denoted θ_i for topic i and some task difficulty, often denoted β_i. This provides a straightforward, though perhaps costly, way to measure how well a student understood a topic they were trying to learn. In effect, if these latent attributes can be estimated, we get a reasonable estimate of how well they have learned, where "learning" is operationalized by how well they would likely perform on a test on that topic.

The above explanation involves the simplest case of one latent knowledge ability being

measured. However, if there are N topics being tested, there are two categories of modeling the probability of correct answers that can be considered. The first category are the "non-compensatory" models which state that a student's ability to perform well on topic i is only governed by their latent knowledge θ_i and task difficulty β_i:

$$P(Y_i = 1|\theta_i, \beta_i) = \frac{1}{1 + \exp\left(-\left[\theta_i + \beta_i\right]\right)}$$

However, the second category, the "compensatory" model, states that the student's ability to perform well on topic i is governed by a linear combination of their latent abilities with respect to *all* N topics. Thus, if the student has weak knowledge of one topic but has strong knowledge of several others, that knowledge could *compensate* for the weakness (Junker, 1999) (Pirolli and Kairam, 2013):

$$P(Y_i = 1|\theta_1, \cdots, \theta_N, \beta_i) = \frac{1}{1 + \exp\left(-\left[\alpha_1\theta_1 + \cdots + \alpha_N\theta_N + \beta_i\right]\right)}$$

Determining which model of IRT is appropriate thus depends on the specific topic being tested and the latent abilities θ. If the different topics are topically unrelated, the non-compensatory model may be more appropriate. Conversely, if all topics are separate but in the same domain such as "geology", a compensatory model may be more accurate.

Estimating learning is a crucial part of our goal as it is a necessary component in: (1) evaluating the effectiveness of our system and (2) evaluating the student's current state of knowledge for use in a feedback loop to offer documents that can address the weaknesses and leverage the strengths in the student's knowledge (Part 6 of Figure 2.2).

While both Bayesian Knowledge Tracing (BKT) and Item Response Theory (IRT) have

been used extensively in learning modeling literature, there have also been some recent studies that have tried to leverage advantages of both (Wilson, Karklin, Han, and Ekanadham, 2016). However, for the purposes of this dissertation, we will focus on using Item Response Theory for simplicity and to account for the fact that the studies we will be focusing on were not designed to incorporate adaptive content selection as a function of feedback - a task which BKT would have been likely to perform better on.

2.1.3 Assessing Learning

Many methods have been proposed for evaluating a user's knowledge state which in turn could be used at multiple stages to evaluate learning gains (Wildemuth, 2004; Wilson and Wilson, 2013). Some of the more common methods include: (1) multiple-choice questions; (2) sentence cloze tests and (3) free-form responses (Abualsaud, 2017; Frishkoff, Collins-Thompson, Hodges, and Crossley, 2016; Syed and Collins-Thompson, 2017b; Wilson and Wilson, 2013). Each of these various methods have different advantages and disadvantages making some more suitable than others depending on the application. For example, multiple-choice questions are typically more suited when the knowledge being assessed has an objectively correct answer (such as answers to mathematics questions) and where the experimenter is only interested in whether or not the learner is capable of detecting that answer. On the other hand, free-form responses will typically be better for more subjective topics without a clear correct answer (such as topics relating to ethics and morality) and where the experimenter also wants an understanding of the learner's thought process. Multiple-choice questions gave an advantage of easy scalability as there is an objectively correct answer that can easily be detected whereas free-form questions either rely on manual graders or stochastic grading through methods like LSA comparison to a gold standard answer (Franzke, Kintsch,

Caccamise, Johnson, and Dooley, 2005; Graesser, Chipman, Haynes, and Olney, 2005).

While the above methods or some combination of these have been used extensively in many studies of learning (Abualsaud, 2017; Collins-Thompson et al., 2016; Duggan and Payne, 2008; Mao, Liu, Kando, Zhang, and Ma, 2018; Syed and Collins-Thompson, 2017b), other measures have been proposed which are often used in domain-specific settings. For example, through the Betty's Brain teachable agent system, a learner's knowledge is assessed through their ability to express learned concepts through a visual concept map (Leelawong and Biswas, 2008). Similarly, work by Egusa, Saito, Takaku, Terai, Miwa, and Kando (2010) evaluated learning outcomes in a search environment in terms of how participants' pre- and post-search concept maps of the topics changed. In this thesis we will primarily focus on objective measurements of learning to support scalable studies of learning (which may provide better sample sizes for data-driven exploration) but we will also use lab-based free-form graded assessment studies to maintain a balanced understanding of how well the results we find may generalize.

2.2 Models of Information Seeking

In the previous chapter, we discussed how knowledge can be classified, modeled and evaluated. Now we move the focus from classifying and evaluating knowledge to the broader question of how people actually *acquire* knowledge. To design a system that supports learning objectives, we first must understand the basics of how people search and seek information and the complexities of the learning process itself. In this chapter we investigate background theories of models relating to information seeking and learning.

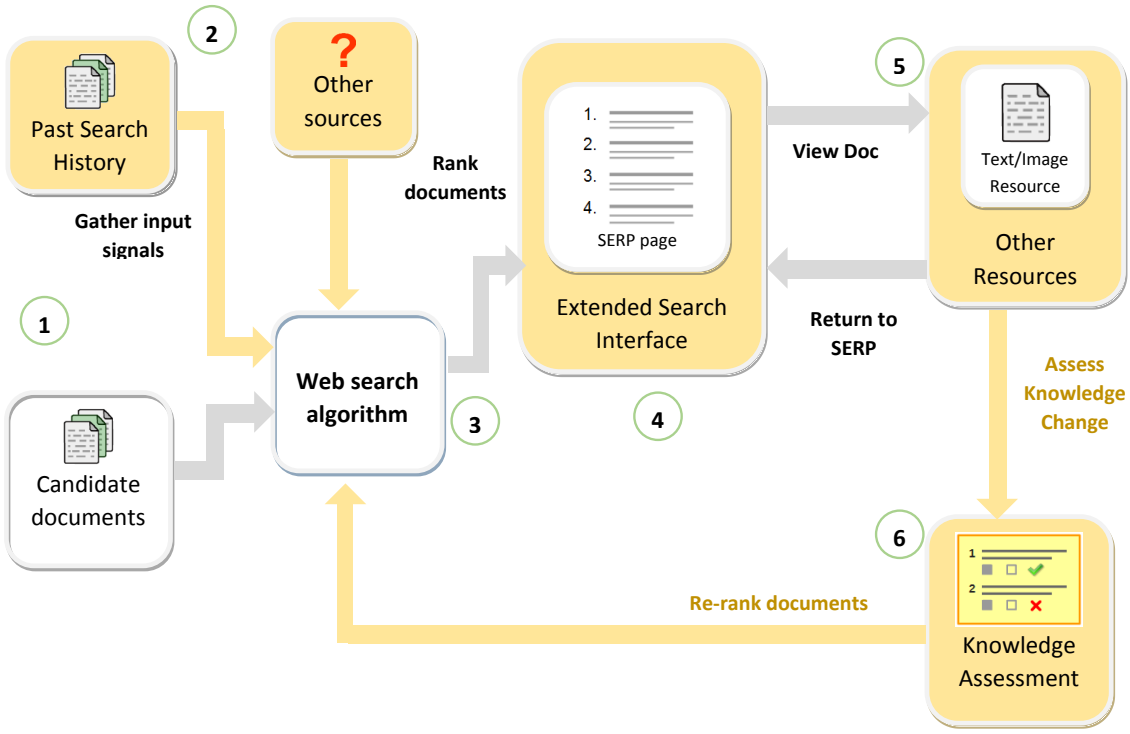

Figure 2.2: High-level description of target Web search algorithm. Orange-shaded entities are areas of possible future work. (interface design, types of resources and feedback loops)

2.2.1 Information Seeking Process

We will start by investigating some of the well-established models of information seeking to identify the different *ways* that people learn. We will first identify the different "levels" of information need that a person can have and then narrow our discussion further. Taylor (Taylor, 1968) proposed a four-level hierarchy of types of information needs, sorted by clarity of the need to the person. In order of increasing clarity, these levels are:

1. Visceral need - characterized by a vague understanding that there is an information need (the true need), often represented by some unclear dissatisfaction.

2. Conscious need - the person has a conscious understanding of what the need is but it is still ill-defined.

3. Formalized need - the person now has clear understanding of their information need and can express it concretely in their own words.

4. Compromised need - the information need as expressed to a search system, typically modified to accommodate the limitations of the search system.

While I acknowledge that the primary information need that anyone has is governed by their visceral need, this level of abstraction is very difficult to operationalize and as such, is out of the scope of evaluation and design in this dissertation. The primary focus in this dissertation will be on the remaining three levels of need which we will express in the context of a Web search system. For each of these levels, there are valid reasons to focus on them when thinking about an ideal search system. We focus on the compromised need because that precisely represents the query that users enter to conduct a search session in the context of Web search (Part 1 of Figure 2.2). We focus on the formalized need because this represents their expectations of the search tool and the thought process that drove the users to construct their compromised need. We focus on the conscious need because a student learning about a new topic is unlikely to be fully familiar with what to search for and will need help resolving possible ambiguities about their true search intent. To help the student with this, we could use information about their prior knowledge about the topic they are interested in learning about as well as information about that topic itself (Chapter 3.5) and (Part 2 of Figure 2.2).

The I-LEARN model developed by Neuman (2011) considers a somewhat more general picture of information need in the "identify" stage of their model. This consists of three parts: (1) activate, (2) scan and (3) formulate. The first part, activate, involves the individual

having a sense of curiosity about something in the world to begin with. The model posits that without this, learning may still happen but is likely to be hindered as individuals tend to learn better when they formulate questions of their own interests (Neuman, 2011). The second part, scan, involves considering something specific in their environment that they have an interest in learning more about. The third part, formulate, directly links to Taylor's Formalized need where the individual now has formulated the questions that they want answered regarding their interest from their scanning.

Now, that we have established the various levels of clarity of information need, we need to determine what are the various models that describe how the underlying need is actually satisfied (i.e. how the knowledge acquisition actually occurs). In particular, we need to develop an understanding of how a student's learning evolves over the course of an information seeking process. At its simplest level, the process of learning can be thought of as the update to an individual's knowledge state in response to new information, as described by Brooke's fundamental equation (Brookes, 1980):

$$K[S] + \Delta I = K[S + \Delta S] \tag{2.1}$$

where $K[S]$ is the individual's current knowledge structure, updated to a modified knowledge structure $K[S+\Delta S]$ by the additional information source ΔI. While this offers a helpful abstract way of thinking of the learning process, it doesn't offer any explanation of *why* people seek information to begin with. Dervin (Dervin, 1983) proposed the well-established sense-making model that attempts to answer this question. The sense-making model posits that people attempt to update their current knowledge state to a new one as a response to situations that can't be adequately explained by the current knowledge state. Comparing to Brooke's fundamental equation, the gap in knowledge states can be thought of as the new

information that would update the current knowledge to a form that can explain the current situation. This knowledge gap is a fundamental basis of the sense-making model.

In particular, the sense-making model can be thought of as consisting of three parts:

1. Situation - this defines the context in which a person encounters an information need.
2. Gap - this defines what the information need actually is.
3. Use - this defines how the person uses the new information they have received in satisfying their information needs.

By the sense-making model, a person only needs to perform a knowledge update as a response to an information need. The model also theorizes that sense-making is not a static process but rather a constantly occurring process of discovery and questions.

Marchionini's Information seeking model. Marchionini (Marchionini, 1997) further specifies how the sense-making model applies in the electronic systems space. In particular, he shows that the information search process using electronic search systems can also be roughly characterized in terms of the situation-gap-use paradigm. In the search system, the analogous stages are: (1) Understand; (2) Plan & Execution; (3) Execution & Use (Marchionini, 1997). Similar to the sense-making model, the Understand phase is centered on the objective of first recognizing and then understanding what the information need is that needs to be resolved. The second stage is in planning and executing the sequence of actions to close the knowledge gap by choosing an appropriate search tool and submitting a query to it. The final stage involves analyzing the results the search system gave in response to the query and determining if it has resolved the information need. If it has not, the process returns to stage two and either reformulates the expressed information need or

chooses different result options. Once the need has been satisfied, the searcher can now "use" the newfound information for whatever intent they had until a new information need arises and the entire process repeats.

Kuhlthau's ISP model. While the above models attempt to explain the Information Seeking Process (ISP) in terms of the cognitive (thoughts and ideas) dimension and, in the case of Marchionini's model, also physical (actual actions, e.g. entering a search) dimension, a third crucial dimension is still missing: the affective dimension (feelings/emotions). These three dimensions are a fundamental basis for Kuhlthau's six-stage model of the Information Seeking Process (Kuhlthau, 1991). In her model, she shows that the process starts with feelings of uncertainty as the person takes on the task of learning something new. The first stage, Initiation, involves a simple recognition of some information need. The next stage, Selection, involves actually selecting the topic of interest to narrow down from. Once the information need has been identified and a topic of interest to satisfy that need has been selected, the actual exploration of resources begins. This occurs in the third stage, Exploration, which involves investigating existing resources on the general topic that had been selected. Up until this point, a searcher is likely to express feelings relating to confusion, uncertainty or anxiety as they are still in the process of figuring out what they need to focus on. This changes in the pivotal fourth stage, Formulation, where the information need is further refined to a more narrowed topic that the searcher feels comfortable with as a specific topic that will satisfy their information need. This will typically be followed by stages of Collection, where the searcher will begin to collect resources that specifically target their focused topic and finally, Presentation expresses the results of their search process (Kuhlthau, 1991). This process, as in the one in Marchionini's model, typically shows a general trend from uncertainty or generalized information need to more certainty and a

focused, concretized information need. However, unlike Marchionini's model, this ISP model considers the affective dimension as well and places a strong importance on including it. For educational search systems, this model is particularly useful as it was largely tested on and built from studies of school and library information seeking tasks which were largely educational in nature. There is also evidence in support of considering the affective dimension of the process, particularly for educational objectives. Kuhlthau found that there was a statistically strong correlation between stronger changes in feelings of confidence at different points in the learning process and stronger actual grades assigned at the end. While this framework of ISP was developed more than two decades ago, a relatively recent large-scale work by Kuhlthau, Heinström, and Todd (2008) found that the ISP model is still valid in recent times.

Information foraging theory. Another approach that has been used in trying to model the human's information seeking process is to approach the problem from a biological point of view. While the previously described models by Marchionini and Kuhlthau address the specific question of how the information seeking process happens, the information foraging approach gives us a more broad picture and a better understanding of *why* humans search in particular ways. It is useful to have this intuition when designing any system that a human will use as the answer to why humans search the way they do gives us insights into what motivates them to search and what might motivate them to abandon their search.

If we frame the human's information seeking process as the user's attempt to acquire certain information in a limited environment, we can look to biological research of how humans more generally attempt to acquire something of interest. This is the underlying concept of the Information Foraging theory, posited by Pirolli and Card (1999) who show that we can model a user's information seeking process in terms of a human's foraging

behavior contextualized in a search task. The theory states that the net information utility, defined as "currency", of a search attempt is the net result of the total currency acquired in the attempt minus the total cost associated with the attempt. Naturally then, optimal information seeking can be thought of as trying to maximize the user's net information utility.

Using the analogy of birds looking for berries (as their "currency") in bushes, the authors suggest that information foraging happens at two levels: (1) the patches, or clusters, level and (2) the within-patch level (Pirolli and Card, 1999). If we consider this in the context of Web search, we can draw parallels to this analogy with the situation of a human looking for solutions to their information need with each possible query they can issue and "hunt" through being the bushes and the SERP pages being the within-patch foraging (See Figure 2.3).

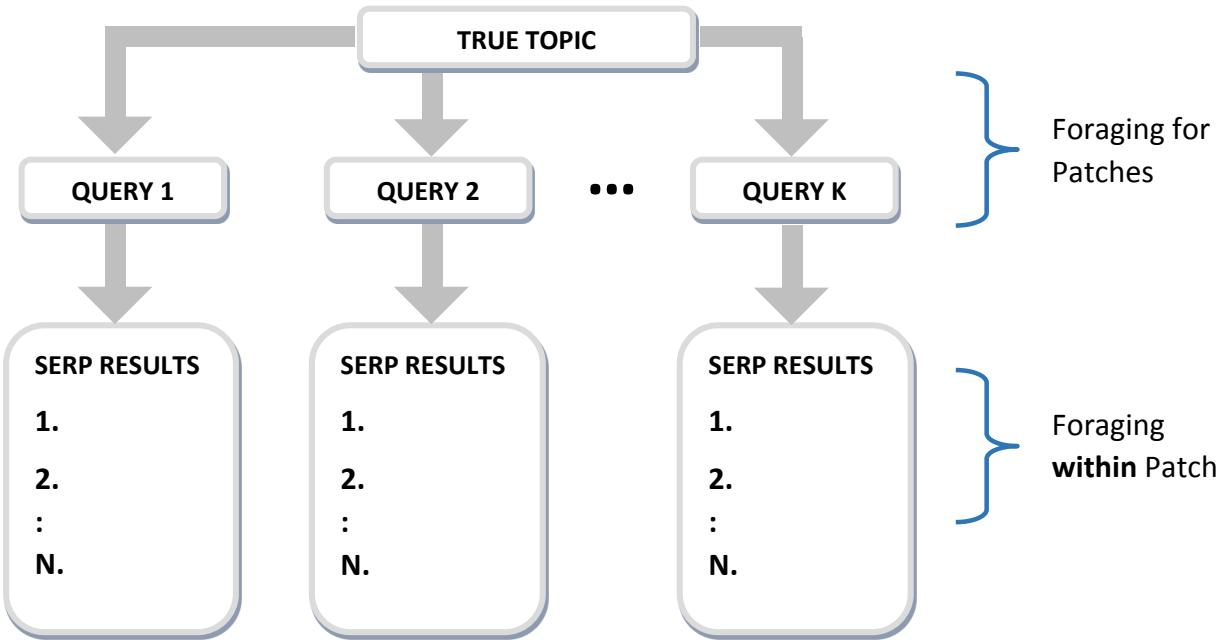

Figure 2.3: Information Foraging theory contextualized for Web search.

The Information Foraging theory also suggests that humans will often use enrichment strategies to improve the utility of the information they find or reduce the cost they have to expend (Pirolli and Card, 1999). This motivates a significant part of the algorithm we will design where we will attempt to reduce the requirement of having to issue multiple queries by providing the user an interleaved single set of results.

The theory also considers the practical scenario of information seeking under risk and uncertainty. It asserts the very real possibility that the searcher does not always know the net utility they will get from a given resource. In particular, if provided a set of documents, the searcher likely does not know how useful they will be. However, studies have found that in these situations, users can make quick judgments of how satisfied they are with the contents of a document just by skimming it to find if it contains what they need (more details in Chapter 3.3.2) before spending more time on it. This is in keeping with the concept that human searchers will not want to expend unnecessary effort fully reading a document that may turn out to be irrelevant.

The Information Foraging theory addresses the very real part of the information seeking process which is the *effort* involved. The theory posits, for instance, that the human searcher will only continue searching in a patch if they determine that the expected utility of expending more effort is not surpassed by other possible options. This is a point that will be very important in our practical design of a retrieval algorithm as we have to assume that the student is going to have finite effort they are willing to expend before, as per foraging theory, they decide to give up and try to address their information need some other way.

We have thus far considered major information seeking models that explain how and why human searchers exhibit certain behaviors and follow certain processes in their information seeking tasks. We will now more specifically focus on the various types of search tasks and

how they pertain to Web search engines.

2.2.2 Search Models and Types of Search

Focusing on the development of an educational search engine, we have to first consider the fundamentals of how search engines are designed to work and the different uses they offer. We first note the high-level distinction between lookup search and exploratory search (Marchionini, 2006). Lookup search refers to searches where the user has a specific search objective in mind and a concrete expectation of exactly what form they expect the results to be in. For example, a search for a specific research article shows a precise search intent, a precise expectation of the format of the result and minimal need to consider multiple results. Similarly, a navigational query to Tesla Motor's website would be a lookup search as there is again a very precise information need, a very clear expectation of what to expect and minimal need to consider more than one document. On the other hand, exploratory search involves an information need that is less precise, usually involves multiple iterations and typically implies less prior knowledge of the topic in question. For example, a physics novice wanting to learn about String Theory might issue an initial query to learn the basics but might issue further queries to understand background knowledge or specific aspects of the theory as their knowledge of the subject develops.

Of particular focus in this paper are exploratory search intents as we consider "educational search" to be a subset of this type of search. Specifically, we define **educational search** to be any Web search with the primary intent of updating the searcher's knowledge about a particular domain-specific subject. We consider educational search to be a subset of exploratory search as the latter can also involve other search intents such as exploratory transactional (commerce intent) search. Now, we see that our definition makes educational

search a subset of all possible knowledge updates an individual can experience as per the high-level Brooke's fundamental equation (Brookes, 1980). We also observe that the concept of knowledge updates being a response to information signals has parallels to the concept of exploratory search evolving as the searcher gets new information from documents they read. As such, we consider educational search itself to be a form of exploratory search so we will focus our paper on the design and evaluation of an optimized algorithm for exploratory search but with educational (learning) goals.

Traditional IR systems were designed on the basis of a linear model of search where a user would enter a query, get a matching document and optionally repeat the process (Bates, 1989) (Figure 2.4).

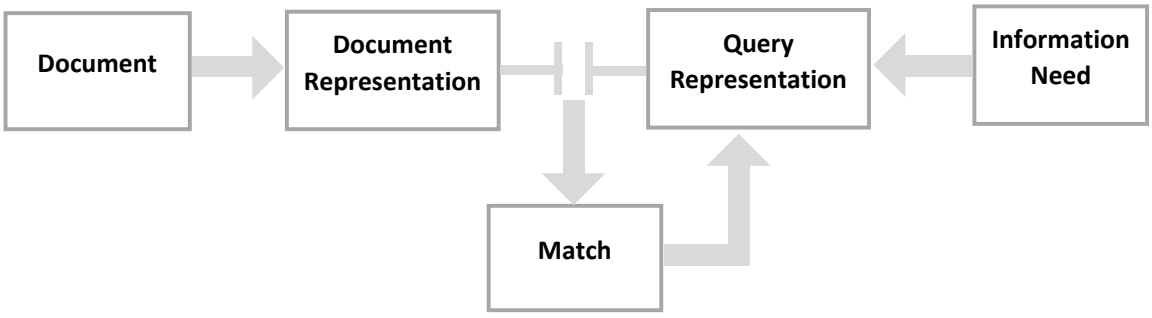

Figure 2.4: The Classic Information Retrieval Model.

This model, while useful for simple lookup tasks was not so useful in explaining tasks that involved far more complex intentions and evolving goals. The Berrypicking model, proposed by Bates (Bates, 1989), offered a new way of considering the search process as an evolving process where the user's goal could change as a function of the resources they read. This brings us closer to the concept of exploratory search (Marchionini, 2006) where there is a

loosely defined goal that updates as the student reads more.

We also note the fact that the learning outcome from using a search system is not simply a function of accommodating the correct cognitive model for the given search task (task domain). The ideal search system would also have to incorporate elements of the task setting (e.g. leisurely or professional) and provide an interface that the user can easily use to accomplish their goals (Marchionini, 1997; Shneiderman and Marchionini, 1988). Marchionini also shows that task types can be defined in terms of three dimensions (Marchionini, 1997). These include:

1. Specificity - explains how much information (*depth* of information) the searcher wants to learn for the given task.

2. Quantity - defines the volume of information units (such as words) the searcher is interested in reading through for the given task.

3. Timeliness - defines the expectations the searcher has for how long the given task will take to complete.

Different task types will naturally necessitate different importance assigned to document features such as redundancy, length and estimated time to read.

2.3 Background Summary

In this chapter I discussed some of the core background that informs the directions of research involved in this dissertation. Specifically, we discussed how knowledge can be represented by the Bloom's taxonomy, can be modeled via Item Response Theory (IRT) and can be evaluated via multiple-choice or free-form responses. As we will discuss later in this thesis,

we use these as foundational basis for some of the core studies. We further discussed about the multiple models of how people seek and integrate new information in general. From these models, it can be shown that people successfully update their knowledge state by taking in new information signals which they can reconcile (i.e. is non-conflicting) with their existing knowledge state. Furthermore, we showed that individuals may have different expectations of how much information they need for a given learning task - that is, there is an importance in personalizing the selection of resources (e.g. Web documents) to accommodate individual preferences and different prior knowledge states.

Chapter 3

Literature Review

In the previous chapter, we discussed models and representations of learning as well as models of information seeking and knowledge acquisition. In this chapter, we investigate related work that addresses the objective of developing an information retrieval model that optimizes for learning goals. Of particular importance to the studies in this dissertation are Sections 3.3 - 3.6 that review literature on the role of effort in search and learning (3.3), how people use Web search for learning (3.4), the role of intrinsic diversity in search (3.5) and what Web document features are indicators of learning outcomes (3.6).

However, before even considering designing such a model, we need to determine whether or not there is even a need for it. Will people use it? How effective are existing systems? If they are not effective, are the reasons known? Will the end result have a steep learning curve? Will the algorithm be suitable for a potentially limitless range of topics? This literature review will be split into three main sections:

1. **Motivation (Part 1).** These sections will investigate existing work that shows just how important Web search is for educational intents and goals and why there are substantial benefits from improving such search systems. We will also look into existing work on Intelligent Tutoring Systems (ITS) and show the enormous potential ITS systems have already and continue to show in specific applications. This will inform an algorithmic approach to constructing a search system that leverages models of

learning.

2. **Effort - The Good and Bad (Part 2).** This section will focus on the multifaceted effects that effort on the part of the end-user can have on their learning outcomes. We show that effort can be both a good and bad variable affecting learning outcomes depending on context.

3. **Search as Learning (Part 3).** These sections will focus on the intersection between search technology and algorithms and what is known and what continues to be investigated regarding learning outcomes using search systems and Web documents.

In surveying the existing literature, we will tackle the fundamental question of whether or not there already exist Web search algorithms designed for optimizing progress towards educational information-seeking goals. We will review existing work in this area and show that the few studies that are designed as such are weakly designed to accommodate generalized Web search for learning. We motivate our own design of an optimized search algorithm toward addressing the flaws in the few existing systems in the area.

3.1 Relevance of Web search engines for education

Over the past few decades, there has been considerable work that has focused on the design, implementation and evaluation of information retrieval models (Junker, 1999; Sanderson and Croft, 2012) . One such system that has gained strong popularity is the search engine, a tool that supports information retrieval using some form of a query as an input and provides a set of resources as output. While traditional search engines limited the input to text and the output to a simple list of links to documents (today, often Web pages), contemporary search

engines may include other media formats for both inputs and outputs such as searching by color to find relevant frames in a video (Lokoč, Blažek, and Skopal, 2014). However, for the purposes of this study, we will focus on text queries as inputs and a list of Web page documents as outputs. Before even considering the implications of designing such a tool that optimizes the retrieval output for educational goals, we have to first question whether or not people even need or want search engines for learning.

Numerous studies have shown that people across different age groups do in fact use search systems as an important part of accomplishing their learning goals. Pew research reports have found over the course of many years increasingly higher percentages of participants use Web search as one of the most popular activities on the internet (Purcell, Brenner, and Rainie, 2018). Several small-scale studies (Bilal, 2000) have demonstrated that students do in fact use and have expressed preference in using Web search engines for educational objectives and at least one large scale study (NetDay, 2004) demonstrated an interest that students show in using technology, more generally, to assist in their learning. Early work by Hölscher and Strube (2000) showed that a majority of users (81%) in their study chose to begin solving their information need by using a search engine. Similarly, a study by OCLC (De Rosa, 2006) showed that an overwhelming fraction of college students (89%) use Web search engines as a starting point for information search and a similarly large percentage (94%) claim that search engines are a "good to perfect lifestyle fit" (De Rosa, 2006) in response to how they would rate search engines based on their "information needs and lifestyles". Similarly, work by Griffiths and Brophy (2005) showed that 86% of students claimed to used search engines at least once a week, with 57% claiming to use search engines everyday. The study also showed that 68% of students used one of several commercial Web search engines as their first step in starting an information seeking task. There is further evidence suggesting that this is a rising trend.

A five-year large-scale study done by Judd and Kennedy (2010) at a university showed that Google search accounted for the largest fraction of internet sessions involving "information seeking sites and services". They further show that this trend consistently increased every year from 2005 (24%) to 2009 (31%) (Judd and Kennedy, 2010). A study by Dutton and Helsper (2007) showed that over the course of two years, there was a sharp threefold increase in the fraction of participants who primarily used Web search engines as their source to look for information on the internet (from 19% of participants in 2005 to 57% in 2007) (Dutton and Helsper, 2007). Even more recently, a lab study by Niu, Abbas, Maher, and Grace (2018) found that 90% of participants reported using a Web search engine as their "primary source for health information" and a study by Salehi, Du, and Ashman (2018) found that 83% of students considered search engines to be an important or very important source of academic information. In addition to individual users using search for learning, there has been work on independent systems using Web search as a back end for a learning application. The REAP project (Collins-Thompson and Callan, 2004) used an intelligent tutoring system for language learning as the client to a richly featured document retrieval system. That system could find authentic practice materials from the Web satisfying multiple constraints related to the student's learning goals.

Furthermore, there is strong evidence that students who use search engines for learning generally find the results to be useful. A study by Henderson, Selwyn, Finger, and Aston (2015) found that nearly all surveyed students used internet search engines to find information relating to their university studies and an overwhelming majority (96.9%) reported the perceived usefulness as either 'Useful' or 'Very Useful'. This is consistent with our own recent study where we asked participants about their perceived usefulness of search engines for learning (Syed et al., 2020). We found that 91% rated the usefulness as 'Useful' or

'Very Useful'. However, both studies found significantly lower fraction of students giving the higher rating of 'Very Useful' (68% in the study by Henderson et al. (2015) and 35% in our own study). This suggests that there still remains substantial room for improvement in the search as learning experience.

Though Web search engines can be used for more than just information seeking (e.g. entertainment, games, shopping, banking), prior work has shown that a significant portion of queries in a major Web search engine were indicative of some form of information acquisition. Broder (2002) investigated the distribution of types of queries on a major Web search engine where the query types were categorized under one of three categories: navigational, informational and transactional. A log analysis and survey were conducted on actual Web search data and showed that the prevalent type of search was informational, characterized as "The intent is to acquire some information assumed to be present on one or more Web pages" (Broder, 2002). As a starting point, this at least confirms that people who used Web search engines as of 2002 were largely using it for some type of learning purpose (e.g. as compared to navigational or transactional purposes). A later study using the same search engine and breaking the taxonomy down to a finer granularity found similar results (Rose and Levinson, 2004). The study more specifically found that the predominant type of informational queries were undirected informational queries, characterized as: "I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X." (Rose and Levinson, 2004). The trend towards an increasing fraction of informational search traffic is further shown in a study three years later (Jansen, Booth, and Spink, 2007) where informational queries account for about 80% of a sample of Web traffic and much more recently in a study which found 49.7% of queries where informational and more particularly had "specific learning intent" (Yu, Gadiraju, Holtz, Rokicki, Kemkes, and

Dietze, 2018).

A large-scale study of search queries on commercial search engines by Bailey et al. (2012) demonstrated that queries that have the purpose to "discover more information about a specific topic" account for the second-highest fraction of queries issued per session (approximately 14%) over a two-month period in 2009. Furthermore, such educational tasks are shown to involve multiple queries (6.8 on average) and significant time spent (13.5 minutes on average) (Bailey et al., 2012). Therefore, there is evidence not only of significant informational intent in Web search queries but also evidence that this intent is of an exploratory nature (Marchionini, 2006).

Despite the use of search technology to assist in learning, studies have also found that there are some limitations that may discourage more dependence on search engines for learning (De Rosa, 2006; Fox and Jones, 2009; Ng and Gunstone, 2002). For example, while students in classroom settings claim to prefer using online tools to assist in learning, a majority of them reported the presence of a teacher to be crucial as well (Ng and Gunstone, 2002). There is also evidence that students don't trust online resources nearly as much as they trust the opinions of their teachers. An OCLC study (De Rosa, 2006) found that college students consider a teacher or professor to be the most trusted source for validating information. Neuman (Neuman, 2011) also shows that in a study with seventh-grade children, many demonstrated a strong dependence on a teacher's instructions to know what to search for and showed weak ability in being able to independently synthesize ideas to form a collective whole. Addressing such issues of dependency in a Web-based learning environment is then naturally an important area of focus. While solving this problem in a Web-based learning environment is nontrivial and will not be the focus of this paper, it will naturally be an important area of focus in the development of an ideal self-supported Web search system for

learning.

For sensitive information, such as health information, people still strongly use trusted sources such as health professionals or doctors (Fox and Jones, 2009) although a significant, and growing, fraction (57%) also claim to use the internet for health concerns. A more recent study, also by Pew, shows that the fraction of those who use Web search for health concerns has jumped up to 72% (Fox, 2014), suggesting a greater interest in using the Web even for more serious topics such as medicine. It is not, however, clear if this increase in use of the Web searching translates to increase in using Web search *engines*. Earlier work by Spink, Jansen, Wolfram, and Saracevic (2002) showed that over the course of six years, the fraction of health-related queries to a major search engine of the time steadily declined from a high of 9.5% to a low of 7.5%. A later study by Spink, Yang, Jansen, Nykanen, Lorence, Ozmutlu, and Ozmutlu (2004) showed that this fraction dipped even further when comparing another search engine down to 3.2%. More recent work by White and Horvitz (2009) showed that only about 2% of all queries from a sample of user queries in a large-scale query log were health-related. It is worth noting, however, that approximately 25% of all users in the large-scale log issued at least one health-related query at some point in the log results. Furthermore, results from a large survey in the same study showed that on average, participants report issuing around 2 health-related Web searches per week. This suggests that while health-related queries may not be very prevalent, relative to other types of queries, they are still frequently used by many and it becomes imperative that the search engines servicing such queries are able to provide relevant and correct information on the topic.

As we have now clearly established, there is a strong demand for Web search engines for educational or learning purposes that transcends the demographic of only school students

and also includes the general population as discussed in the general population log studies above. We now question whether or not existing Web search engines are adequately satisfying this demand already. A study by Brophy and Bawden (2005) showed that in four exploratory educational search tasks, the precision of the top 10 results returned by Google was, on average 56%, indicating that of the results being offered, there is significant room for improvement as far as educational relevance is concerned. A study by Griffiths and Brophy (2005) similarly shows that in the academic search tasks students were assigned using various Web search tools, 30% reported being unable to find the information they needed and 12% simply gave up, giving reasons along the lines of "frustration; all sites were irrelevant". Even of those who did find the required information, only 50% claimed it was easy. This suggests that half of searchers find it difficult to locate required information using existing Web search engines and nearly one-thirds of searchers are unable to find the information they need.

An emergent theme in the educational space over the past few decades has been the concept of "flipped learning" where students learn or study the passive content of the lectures at home and engage in active learning, discussions and activities in the classroom (Bishop and Verleger, 2013). Whereas in the past, students had relatively more structured and directed information goals, the flipped learning paradigm gives students more options for open-ended goals, guided by their own active thinking and discussions with their peers. Earlier we showed how Web search plays an important role in self-directed learning. As such, the flipped learning paradigm would be heavily improved if self-directed learners had access to a search algorithm designed to support a variety of learning-oriented goals.

A recent comprehensive review of existing literature in this space shows that little research conclusively measures how effective flipped learning is in terms of some measure of actual learning improvement (Bishop and Verleger, 2013). However, the literature does show that

students generally show positive attitudes towards the idea of flipped learning and the idea of more active engagement in classroom activities. A further benefit of our work in the area of search as learning is that we could add to the existing literature on investigating how students may benefit in actual learning improvements from at-home learning. While a search engine customized for self-directed learning objectives covers only one half of flipped learning, it would be a valuable contribution and further motivates the need for our work as the emergence of flipped learning paradigms necessitates the need for students to easily access educational material that offers high learning utility.

3.2 Intelligent Tutoring Systems

The concept of providing optimal educational resources to a user in an educational information seeking context has been well-established through the varied implementations of Intelligent Tutoring Systems (ITS), both offline and on the Web (Brusilovsky, Ritter, and Schwarz, 1997; Kazi, 2005; Keleş, Ocak, Keleş, and Gülcü, 2009; Koedinger et al., 1997; Wolfe, Reyna, Widmer, Cedillos, Fisher, Brust-Renck, and Weil, 2015). Such systems often involve several key components: (1) a student model: a representation of the student and how their knowledge can be estimated; (2) an expert model: a representation of the topic to learn, how it can be represented and what rules it follows; (3) a tutor/optimization model: a model that decides how best to connect a pool of potential resources with the student to optimize their expected learning outcomes. ITS systems have enjoyed significant popularity due to their impressive results in real-life learning outcomes (Koedinger et al., 1997; Wolfe et al., 2015). An early and powerful result in a non-automated setting by Bloom (1984) found that personalized tutoring instruction and mastery learning could significantly improve real-life learning outcomes for students, following which many intelligent systems have

tried to leverage the potential that personalized pedagogical systems could provide.

It should be observed, however, that despite the remarkable results from many of these systems, there are significant challenges to adapting them to the open Web search environment. Firstly, the expert models that such systems use need to have a set of rules that govern correct knowledge of the subject. The ITS systems then trace the student's progress by providing the student opportunities to apply these rules and evaluating their success in doing so. One of the most popular paradigms for this approach is the Knowledge Tracing (KCT) method, proposed by Corbett and Anderson (1994) and used quite extensively since ((Huang, Yudelson, Han, He, and Brusilovsky, 2016; Koedinger et al., 1997). Unfortunately, the very nature of these models often requires explicit or manual coding of the rules of a particular domain, sometimes even in different symbolic language (e.g. calculus will be governed by very different rules and language compared to organic chemistry). As such, traditional ITS systems are very often limited by being domain-specific, limiting their ability to scale to teach arbitrary topics, newly-formed topics or topics of little general interest to most people.

Recent work by Huang et al. (2016) investigated an automated approach to estimate changes in a student's knowledge as they read an online textbook using a more flexible approach to knowledge tracing. For example, rather than explicitly requiring students to apply domain-specific rules, they hypothesized that students who spent relatively less time on documents covering a particular *knowledge component* were more likely to have understood that component, thus leading them to spend less time. Through this approach, the authors demonstrated a potential approach to large-scale knowledge tracing for students reading any textbook. However, a critical assumption they made was that "knowledge level is the only factor that affects reading time", which ignores an arguably crucial variable of tiredness - as people read more, they may get tired and be less willing to spend more time on future

documents. It is thus possible that such an approach might be modeling a user's tiredness rather than learning ability. While the approach investigated by Huang et al. (2016) was an innovative approach to applying ITS concepts at scale, this does show that we need to be careful to choose variables indicating knowledge levels that are not likely conflated with unrelated measures. Similarly, work by Pirolli and Kairam (2013) investigated how to apply the concept of knowledge tracing to the open Web. They allowed participants in a small user study to browse the Web in a learning task and tested the users in a post-test afterwards. In their study, the intermediate variable for knowledge estimates was a function of the proportion of relevant words read by the user, along with user-specific weights assigned to each knowledge component assessed. The authors demonstrated a strong ability for their model to predict actual learning gains, suggesting their approach is a feasible method for some types of learning tasks at scale. However, it should be noted that though there were only five LDA topics the authors modeled on, there was a relatively small sample size, possibly suggesting the results may not generalize for larger numbers of users.

In aggregate, there is strong evidence in favor of the use of ITS systems to support domain-specific learning at scale for self-paced instruction. However, one of the key limitations to this approach lies in the fact that most ITS systems often use custom-designed domain-specific rules which are usually manually coded. This may limit the scalability of ITS in terms of number of topics supported along with their ability to rapidly adapt to new and emerging topics of interest. Furthermore, there may be a usability factor involved. While most people are familiar with using Web search engines to acquire information, it is very likely that most are *not* familiar with using specific intelligent tutoring systems nor is there any particular standard for what design and usability features ITS systems should follow.

As such, while there is definitely strong potential for learning gains through ITS systems,

we will not be using such architectures in this work. In the ideal case, users would learn on the Web using an ITS system that could be adaptive to any arbitrary topic, accurately estimate the user's current knowledge state at any given time and adaptively choose which documents to provide next. As discussed, there are several limitations hindering such an open-ended system but the models and retrieval algorithms that will be detailed later in this paper lay an important foundation for both specific and general-purpose learning on the Web. It is our hope that later work could build on the models we provide to move closer to this idealized objective.

3.3 Difficulties in Learning - The Good and Bad

In the previous section, we have discussed the prevalent theories and models of search and information seeking both in the general context and in the specific context of search systems. In the current section, we will focus on another dimensions of the learning process: *difficulties* (which we will interchangeably also call as *effort*). We will show that while some difficulties in the learning process are indeed harmful (leading to intuitively worse learning outcomes), there are other cases where *appropriate* difficulties are actually helpful for learning improvements that endure over time.

3.3.1 The Good - Desirable Difficulties

While thus far we have talked about learning in a largely single-dimension form, we will now start breaking down "learning" into different forms. In this section, we distinguish between *recall* (which we will define as short-term learning) and *robust learning* (which we will define as persistent, or long-term learning) (Schmidt and Bjork, 1992). While many techniques have

been developed and many training techniques have been designed for short-term training, the long-term impact is often not considered as strongly (Bjork, 1994). A concerning and consistent finding is that teaching programs that offer students better immediate learning, either perceptually or actually, often tend to offer significantly weaker long-term proficiency (Bjork, 994a).

Bjork first introduced the concept of desirable difficulties as a necessary component for long-term learning where the student must be tested above and beyond what they may be comfortable with in order to facilitate more active learning (Bjork, 1994)(Bjork and Bjork, 2011). Numerous studies (Little and Bjork, 2012)(Adams, McLaren, Mayer, Goguadze, and Isotani, 2013)(Bjork, Little, and Storm, 2014) after this concept was initially introduced have demonstrated that concepts such as spaced learning (Dobson, 2011) and interleaved assessment (Rohrer, Dedrick, and Stershic, 2015)(Kornell and Bjork, 2008) do in fact offer improved robust learning. Similarly, a recent work by Vakkari and Huuskonen (2012) investigated the relationship between search effort and task outcome in a web-based study and found that variables indicating more effort positively correlated with improved learning scores. A more recent study by Tang, McBride, and Pardos (2015) found that in a template-based Intelligent Tutoring System, problem sets that showed a question with more learning difficulty first, led to a higher learning rate for that problem set overall. The authors demonstrated how the concept of desirable difficulties explains their finding.

In constructing the ideal search tool for learning, we must take into account the fact that, counterintuitive as it may be, a difficult learning experience is ultimately an optimal one, insofar as the difficulty is “desirable" (Bjork, 1994).

Prior work also demonstrates the importance of considering errors and mistakes in the learning process not as indications of some intrinsic failure but rather as opportunities for

better learning (Bjork, 1994)(Ohlsson, 1996). Ohlsson, for example, shows that errors are simply a manifestation of existing knowledge deficiencies and can therefore be useful in helping to diagnose the problem and correct the learner's understanding. In particular, he considers all practical knowledge to be composed of methods which themselves may contain one or more *production rules*, the most fundamental unit of practical knowledge. Each production rule is governed by a specific goal G, a situation S which the person currently is in and a corresponding action A that they take to accomplish G in situation S. Ohlsson posits that while errors can stem from many sources, their fundamental effect is for the production rules to incorrectly assign action B as appropriate when action A was actually correct (Ohlsson, 1996). He shows the importance of errors in learning as they can offer appropriate feedback that addresses the *cause* of the errors and can be used to incorporate the appropriate constraints on the production rules to prevent future mistakes and hence improve practical knowledge.

We note that while Ohlsson's work is largely focused on practical knowledge for practical skills, the concept of using mistakes for learning applies to the theoretical domain as well as we have discussed in terms of desirable difficulties. By making mistakes and by having the correct feedback mechanisms, students can acknowledge defects in their production rules, adjust those rules accordingly and perform better. If a student is given very easy assignments and tasks to perform, they are far less likely to make mistakes but in doing so, may be unaware of fundamental errors in their production rules that may only manifest in more difficult situations.

The concept of incorporating effort in the learning process predates the work of Bjork (Bjork, 994a) with one of the earlier concepts of the Generation effect, posited by Slamecka and Graf (1978). This effect essentially claims that human learners acquire knowledge better

when they actively have to *generate* that knowledge in some form, as compared to accepting passive input. This effect was observed in a vocabulary test across multiple conditions where those who had to engage in active learning showed better learning than those who engaged in passive learning (Slamecka and Graf, 1978). A later study by Davey and McBride (1986) similarly found that students who had to generate questions for reading passages generally scored better on later assessment of both literal and inferential questions as compared to those that didn't have to generate questions. A relatively new model of information literacy and learning called I-LEARN (Neuman, 2011) similarly posits that students really learn rather than just acquire information when they apply, reflect and generate knowledge in the information seeking process. We note that even something like a test at the end of a reading task can be helpful in engaging the students in active learning. In developing a system for teaching, it is imperative then, that we must incorporate some form of the Generation effect to optimize the student's ultimate learning outcomes.

As the focus of this paper is centered on developing a Web search algorithm, specifically with the constraint of a static SERP page, we will not focus on actually incorporating a constant feedback loop to re-rank the document set as a function of which documents in the ranking the user selects. However, as per the above discussion, we do point out that incorporating such a feature could certainly be an extension of the Web search algorithm in the development of a novel Web search *system*. As we discuss in further sections as well, while our focus is primarily on developing a Web search re-ranking algorithm, there are many ways that the model we construct could be extended for potentially better outcomes in future work. Figure 2.2 shows this concisely where the unshaded entities represent the objective of this study and the orange-shaded entities represent possible expansions for future work.

3.3.2 The Bad - Impact of Effort and Difficulty in Web Search

We have now discussed the positive effect of *desirable* effort on learning in supporting long-term learning outcomes. However, as was also mentioned in the preceding section, even desirable effort can hurt short-term learning and as many people use Web search for short-term learning objectives, we must consider this. In particular, in the Web search context we consider how more general forms of effort, unrelated to learning any particular topic, can influence how people search and what documents people are likely to read. Although "effort" is an overloaded term in the search literature, it often refers to the amount of text read or contained in a document (Syed and Collins-Thompson, 2017b) and/or the amount of time spent reading such documents (Smucker and Clarke, 2012). Work by Granka, Joachims, and Gay (2004) demonstrated, through an eye-tracking study, that people's interest in perusing documents further down in a SERP ranking falls rapidly as an almost exponential decay. A similar result was confirmed in a study by Joachims, Granka, Pan, Hembrooke, and Gay (2005) as well as another study by Pan, Hembrooke, Joachims, Lorigo, Gay, and Granka (2007) several years later. This points to a delicate tradeoff between wanting users to enjoy long-term learning gains but having a target audience that seems to be very unwilling to expend much effort in the learning process.

An understanding of the problems of limited effort that users have is well-established with common measures (Järvelin and Kekäläinen, 2002) like normalized Discounted Cumulative Gain (nDCG) used to incorporate the weakening interest and accordingly, general gains, that a user gets as they move down a list. More recent work has suggested that the more generalized approach to these cumulative gain measures might not be as effective as thought as they don't incorporate the effort a user must exhaust in actually reading each document.

Smucker and Clarke (2012) developed a time-based measure to better estimate the true effort that is being spent per document the user reads in a list. This measure is a function of the total words contained in the document and an estimate of how much time a user spends per word.

While relevance of documents in a search task is important, the actual link between document relevance and search session satisfaction is also necessary to investigate. Prior work by Huffman and Hochster (2007) shows that relevance of documents in a search session shows very strong correlations to user satisfaction at the end of the session. In particular, they found that even considering only the relevance of the first document of the first query in the session yielded a very strong correlation (r=.722) with the session-level user satisfaction score (Huffman and Hochster, 2007). We now consider how user effort might explain the gap in the correlations between document relevance and user satisfaction.

Yilmaz, Verma, Craswell, Radlinski, and Bailey (2014) conducted a recent study in investigating the appropriateness of existing relevance measures for assessing the usefulness of a document for users. They show that existing measures are not fully measuring document utility as they don't incorporate an element of effort in defining the true "relevance" of a document. As *effort* itself can be defined in different ways, the authors carefully define effort, or high-effort documents to be those "where people need to work relatively hard to extract relevant information" (Yilmaz et al., 2014). In their work, the authors operationalize this definition with two general measures (document length and readability) containing nine specific features. Regression analysis shows that a gap between coded relevance judgments and implicit document utility can be explained, with statistical significance, by both readability features such as the LIX index and by document length features such as the total words in the document. Supporting this position, a more recent study by Liu, Liu, Mao,

Luo, Zhang, and Ma (2018) found that users showed significantly higher perceived usefulness in an exploratory search task when readability was better.

A Web document will not be considered useful, even if relevant, if it is incomprehensible to the specific user visiting the site (Akamatsu, Jatowt, and Tanaka, 2015; Yilmaz et al., 2014). The problem of incomprehensibility in educational Web search goes back at least more than a decade where Ng and Gunstone (2002) found that the most common negative response to an educational task performed on the Web was difficulty in understanding the content. While a certain degree of difficulty in the learning process is desirable, this only holds true if the learning material is still comprehensible.

Verma, Yilmaz, and Craswell (2016) more recently built on the work by Yilmaz et al. (2014) by directly getting "effort" judgments from crowdworkers rather than only getting relevance judgments as was done earlier (Yilmaz et al., 2014). They further specify their definition of effort as consisting of three components: (1) findability - how easy it is to quickly find what you were looking for in a document, (2) readability - how easy is the vocabulary in the document to understand and (3) understandability - how easy was it to actually learn something from the document. They show that of these factors, findability and relevance both predict user satisfaction with statistical significance, thus bolstering the earlier claim that effort does impact the user's "true" relevance judgment. Furthermore, they find that the CLI readability index over a document was a strong and negative predictor of findability. This suggests that documents using more difficult vocabulary typically made it difficult for users to find what they were looking for, thus lowering their overall utility. The authors also found the the document length, measured as total words, was a strong and negative indicator of relevance, possibly suggesting that we should avoid longer documents where possible.

Another recent study by Jiang, Hassan Awadallah, Shi, and White (2015) again shows

that user effort is negatively linked to user satisfaction where the authors consider effort to be defined as a function of total queries Q issued during a search session. They consider effort to be a Q-weighted linear sum of four types of effort: (1) effort in issuing the query (measured by query length) (2) effort in assessing results (measured by average clicks per query) (3) effort in assessing result snippets (measured by the deepest rank of previously clicked snippets) and (4) effort in viewing documents (not defined in their study). While the last type of effort - effort in viewing documents - was not defined in their study (Jiang et al., 2015), the previous two works (Verma et al., 2016; Yilmaz et al., 2014) did consider features that define effort at the document level. A combination of these three results can give us a good indication of effort incurred by users both at the document level and at the more general query level. Another recent study by Akamatsu et al. (2015) investigated the problem of balancing relevance and comprehensibility and showed that their solution to the problem offered documents with more relevant comprehensibility compared to a major search engine. Considering all these findings, when we incorporate effort in our work, we will consider both vocabulary difficulty and document length as factors affecting the user's learning outcomes.

3.3.3 Personalized Difficulty - Difficulty is relative to the User

A key component in developing an optimal search system for education will be to identify the individual learner's current knowledge level and choose teaching resources that challenge that level without becoming too challenging. There is evidence that a significant fraction of Web content for technical topics falls in groups of either high or low reading level difficulty (Kim, Collins-Thompson, Bennett, and Dumais, 2012). The study also found that for some topics, user search preferences indicate that different people visiting sites of the same topical domain

can show significant differences in reading level preferences. As such, content of appropriate difficulty levels are available; it is only a question of how best to use such resources for optimizing learning.

Prior work by Collins-Thompson et al. (2011) and Tan et al. (2012) have investigated the effect of a user's estimated knowledge level on search behaviors where the knowledge level was contextualized as a distribution over reading comprehensibility levels. These studies both showed improvements in standard IR measures when re-ranking documents either according to desired reading level as in (Collins-Thompson et al., 2011) or by re-ranking according to desired difficulty level as in (Tan et al., 2012). The work by Kim et al. (2012) expands on this by investigating why a searcher might be interested in visiting documents far above (at least 4 levels higher) their own typical reading level. They found that this "reading stretch" behavior was indicative of high-motivation tasks like seeking out legal forms, test prep resources or medical information. It is important, then, that an educational search engine should adaptively change its knowledge level criteria to reflect the estimated motivation of the user's queries.

3.4 Search behavior during Learning Tasks

Now that we have a general understanding of how people learn and are impacted by different forms of effort, we need to better understand how learners use existing search tools for their learning tasks. Recent work has shown that the intersection of Web search and learning is a complex and multifaceted domain, which can involve elements of different types of learning, different levels of motivation and variations in levels of expertise (Rieh, Collins-Thompson, Hansen, and Lee, 2016). A thorough understanding of the current research in identifying search patterns and behaviors in learning tasks is a crucial prerequisite step in developing

an optimal search algorithm, as we need to know what to optimize for and what advantages and disadvantages existing search techniques offer.

3.4.1 Search patterns and behaviors

Prior work by Jansen, Booth, and Smith (2009) tested Bloom's revised taxonomy of cognitive learning in a real-life search task, allowing participants to search using any online tool they wanted and see if there were differences in various metrics of search across the six types of cognitive learning. The authors found the interesting result that as they tested tasks of higher cognitive complexity by the revised Bloom's taxonomy, the computed search difficulty did not monotonically rise but rather showed an inverted U shape peaking at the "apply" type of learning. The results of the study do indicate the importance of not considering all intents of queries equally. For example, those who are performing "remember" tasks are likely to just want the simple, easy-to-recall facts whereas those who are doing "evaluating" tasks may benefit from getting search results that offer different sides of a topic (Jansen et al., 2009). A later study by Wu, Kelly, Edwards, and Arguello (2012) also investigated how different search interactions might manifest with tasks of different cognitive complexities using the same revised taxonomy. They found that search interactions in terms of time spent, queries issues and links selected nearly all did monotonically increase, on average, as the cognitive complexity increased. A more recent study by Kalyani and Gadiraju (2019) also investigated search behaviors for learning tasks addressing the six levels of complexity. Their study was conducted via crowdsourcing but similar to the findings by Wu et al. (2012), they too found that the time spent increased nearly monotonically as task complexity increased and found partial evidence that higher task complexity results in more queries issued. These results show that different complexities of search tasks even for the same topic will involve users

following different information trails. As such, there is also an importance in optimizing selection of search results not just for educational intents but also for the particular complexity within such intent. In this dissertation, we will mainly focus on the simplest level of cognitive complexity and introduce a model that adapts its optimization to this form of learning.

Prior work by White, Dumais, and Teevan (2009) have found that Web search behavior between domain experts and non-experts can vary quite significantly in various aspects, suggesting that those with better or worse domain knowledge show different search patterns. In particular, the study found that experts tend to issue more queries, spend more time on average per search session, visit more unique domains and exhibit more "branchiness". They also found that one of the strongest indicators of difference between experts and non-experts was in the fraction of queries issued that contained at least some domain-technical terms. On the basis of this metric as an indicator of expertise, the authors found that over a three-month period, a significant portion of originally non-expert users showed increasing signs of expertise, showing that domain expertise cannot be treated as a static quality and algorithms that use it must treat it as an evolving variable. Eickhoff, Teevan, White, and Dumais (2014) conducted a similar study several years later which was also a post-hoc log analysis where the focus was more on delineating between procedural learning behavior and declarative learning behavior. The authors did consider many of the metrics White et al. (2009) used although they introduced several other measures as well. Unlike the method in (White et al., 2009), the authors looked at changes in domain expertise as a function of six metrics over the much finer session-level granularity. They also found a significant indicator of what seemed to cause changes in expertise which turned out to be viewing a document. They proposed that a Bayesian-style update process was happening, consistent with Bayesian

learning models discussed before (Corbett and Anderson, 1994). Finally, they also identified several document features that could be used in estimating its potential for learning. These two studies show that there are differences in how both the types of learning tasks and the domain expertise can influence search patterns. A study by Kim et al. (2012) also looked at the set of expert and non-expert documents used by White et al. (2009) and found that expert sites have the operational attributes of having higher reading level difficulty and tend to be more topically focused.

Wildemuth (2004) also investigated search patterns in a more direct educational context. They demonstrated that, over the span of an educational course on microbiology, learners at different stages of the learning process show evidence of learning through the nature of their query formulation and reformulation patterns. At the start of the course, most learners were using more terms and were less skilled at knowing exactly what to search to get good results. At the end of the course, students' knowledge assessment scores were almost doubled and there was evidence that they knew what to look for - there was a much higher percentage of search patterns that involved specifying a new concept, viewing the search results and ending the session without further iteration needed (Wildemuth, 2004). Furthermore, there was evidence that at the start of the course (low knowledge state) students tend to issue less searches overall but at the end of the course (high knowledge state), students issued above 20% more searches on average.In developing a search system capable of generic educational goals, the system must be capable of detecting non-expert users and providing them resources that are relatively easier but which also mix in elements of more technical content and websites as desirable difficulties.

We do see a superficially contrasting picture when comparing some of these findings with those of Duggan and Payne (2008). They investigated how prior knowledge of two specific

topics would influence search patterns and post-search knowledge in a controlled lab study. The authors first explicitly tested the participants' knowledge of two topics (football and music) by having them answer fifteen trivia questions about each. The authors then provided them with a Web browser and asked them to search for the answers for the same questions and enter their new responses (post-search knowledge). The authors did show the intuitive finding that prior knowledge scores positively correlated with post-search knowledge scores. However, they also found that prior knowledge scores (arguably indicative of higher domain expertise) were strongly correlated to *less* time spent per page visited, queries issued and number of pages visited, although this was only for the football topic. This is interesting as it directly contrasts what White et al. (2009) found in their study regarding search behaviors of domain experts. Furthermore, a much more recent study that was also lab-based like the study by Duggan and Payne (2008), found similar results regarding search behavior of experts versus non experts (Mao et al., 2018). We hypothesize that this contradiction could be for several reasons: (1) In the lab study, participants were given very specific fact-finding objectives, which may involve very different search patterns as compared to exploratory search; (2) In the lab study, the search activities were "artificial" in that they were not topics that were necessarily reflective of the participants' own information needs. The search behavior in the large-scale study, on the other hand, was organic and had no such constraints and (3) the lab study required all participants to perform the *same* task whereas in the search engine study users were very likely engaged in different tasks, depending on their own personal requirements. While there is evidence that experts tend to perform tasks faster than novices, this tends to be for tasks that are the *same* for all participants (Ohlsson, 1996). The study by Duggan and Payne (2008) also found that prior knowledge of football correlated strongly and negatively with lack of prior knowledge of music, possibly indicating

that some topic-independent traits could be driving domain knowledge of unrelated topics.

Other prior works have investigated how differences along other variables about an individual can affect their search patterns and learning behaviors. For example, gender differences have been found to influence an individual's perception of self-performance during a learning task in a school setting (Lamoureux, Beheshti, Cole, Abuhimed, and AlGhamdi, 2013). Boys typically rated their confidence levels higher than girls during the time period of the project they were assigned whereas girls rated significantly higher confidence levels after the project completed. This may indicate that gender differences in perceptual learning performance not only exist but may vary based on the different stages of the information seeking process (Kuhlthau et al., 2008).

There is also evidence that children learners tend to show different search behavior when it comes to using the Web for learning tasks. Druin, Foss, Hatley, Golub, Guha, Fails, and Hutchinson (2009) investigated how children use the Web to conduct searches and found that while a majority already know about and use Web search for learning goals, the participants also showed signs of incorrect spellings that the search engine could not correct and frustration when they couldn't find what they were looking for. There was also further evidence that even children don't pay much attention to search results past the first page of 10 results, consistent with prior findings of rapidly declining interest (Granka et al., 2004)(Pan et al., 2007). Duarte Torres, Hiemstra, and Serdyukov (2010) also investigated childrens' information seeking behaviors through log-analysis of a major search engine. They found evidence that kids typically entered more wordy queries, issued more queries overall and spent more time per session overall when compared to all users. These studies illustrate a clear problem that children have when it comes to Web search which is that the search engines aren't fully equipped to accommodate the different ways that children express their information needs.

It is of interest to note that another emergent theme in these findings is that certain objective metrics of search success can easily show conflated meanings when taken out of context. In particular, we emphasize how work by White et al. (2009) suggests that higher average time per search session is indicative of greater domain expertise of the user but work by Duarte Torres et al. (2010) shows that the same metric is actually indicative of more child-like search behavior, relative to all users. These two seemingly conflicting results could be explained by the fact that although the time spent on tasks by domain experts and children tend to be similar, the *nature* of the tasks performed are very different. Similarly, work by Odijk, White, Hassan Awadallah, and Dumais (2015) found that evidence of "struggling" search sessions could actually be indicative of either search sessions that were successful or unsuccessful, depending on other variables such as the total queries issued and types of query reformulations. We emphasize the crucial importance of contextualizing these measures of search behavior to avoid issues of conflation.

Differences in search behavior aren't just limited to gender and age group, however. Heinström (2006), for instance found that information seeking techniques could be categorized into roughly three groups: Broad scanners, fast surfers and deep divers and that the behaviors of each group were strongly linked to the psychological profile of the searcher. Broad scanners were more likely to be show exploratory behaviors in their search whereas fast surfers and deep divers were more likely to show specificity in their search. Some of the qualities determined were that broad scanners tend to be "open, curious, competitive" whereas deep divers tend to be "motivated, conscientious, focused". Ford, Miller, and Moss (2003) also investigated how different search behaviors in educational tasks could be clustered and whether or not the particular search task would affect the search behavior. They indeed found that the top three PCA clusters from the Study Approaches Inventory showed

evidence of the three types of study approaches identified in earlier works as "deep approach", "surface approach" and "strategic approach". The study also allowed participants to choose from three different search approaches which were: (1) Boolean search; (2) best-match search (3) combined search. The authors found that the choice of search approach also showed relationships to study approaches. For example, they showed that those who used a Boolean approach were more likely to show active interest and be more anxious (higher fear of failure) whereas those in the best-match approach showed the opposite for two out of three tasks (Ford et al., 2003).

As the information seeking behavior that is most suitable to an individual is linked to their psychological profile, it would be prudent to incorporate some element of their psychological preferences in the educational search system.

3.4.2 Learning Outcomes in Web Search

We have thus far discussed the existing literature on the various patterns that different types of searchers show in educational search. We will now discuss the various indicators that link to measurable educational success.

A recent study by Collins-Thompson et al. (2016) investigated how different search strategies would affect both perceptual and actual learning outcomes and what variables influenced these outcomes. They found that self-reported perceived task difficulty at the start of the task was correlated with lower actual learning outcomes at the end of the task, indicating that a student's perception of how easy a task is can have a strong influence on their educational search outcomes. Prior work by Wu et al. (2012) showed that students typically do have strong perceptual understanding of the actual difficulty of an educational search task.

The study found that in an experiment with college students, most participants' perception of expected difficulty in search tasks were consistent with the designed difficulty as per the revised Bloom's taxonomy.

The study also showed that the time spent reading each document, regardless of the search strategy had a significant positive correlation with actual learning outcomes, thus showing that the time spent per document could be a good implicit indicator of ultimate learning outcomes. This is consistent with the concept of achieving better learning when there are desirable difficulties (Chapter 3.3.1).

3.5 Intrinsic Diversity and Learning

Early work in optimizing information retrieval systems (Robertson, 1977) offered the straightforward principle that an optimal IR system would offer documents ranked in order of decreasing relevance to the user.Robertson (1977) demonstrated that this principle, the Probabilistic Ranking Principle, could be shown to be optimal only on two assumptions: (1) the relevance of document A in a ranking is independent of the relevance of all other documents and (2) the usefulness of relevant documents may change as a function of how many relevant documents have been read. The second assumption seems to have held up well with studies mentioned earlier showing how user interest wanes the further down they go in a SERP page and well-tested measures designed to incorporate a general loss in utility irrespective of document quality (Järvelin and Kekäläinen, 2002). However, the first assumption has been challenged by multiple studies, particularly in the area of intrinsic diversity in Web search. We consider tasks in Web search to be "intrinsically diverse" if they are multifaceted - requiring multiple queries that cover different aspects of the main information goal, to complete (Raman, Bennett, and Collins-Thompson, 2013). Zhai, Cohen, and Lafferty (2003) were one

of the first to introduce the concept of "subtopic retrieval" and importantly demonstrated that optimizing results for such topics required incorporating an assumption that the relevance of a document was in fact dependent on what other documents the user already saw.

Considering that educational search is a form of exploratory search, it follows that optimizing search systems for exploratory search with educational intent will likely improve the quality of results for educational search. In particular, many exploratory search topics can be considered to be intrinsically diverse (Raman et al., 2013) - meaning that these topics, once disambiguated, can be thought of as consisting of multiple subtopics. Early work by Carbonell and Goldstein (1998) has demonstrated how optimizing search rankings for generic relevance is not a sufficient criteria in general as it can result in many topically relevant but redundant documents, leading to no new information for the learner. Carbonell and Goldstein proposed a re-ranking technique, Maximal Marginal Relevance (MMR), to deal with this by incorporating both topical relevance and the marginal novelty offered by each document in the search ranking algorithm (Carbonell and Goldstein, 1998). This technique has been used extensively in search diversity work e.g. by Radlinski and Dumais (2006), Zhai et al. (2003) and the recent large-scale study by Raman et al. (2013). More recent work by Collins-Thompson et al. (2016) demonstrated the potential usefulness of intrinsic diversity in search for improving learning outcomes and perceived search outcome satisfaction. Building on results from (Collins-Thompson et al., 2016), a larger-sized study by Syed and Collins-Thompson (2017a) investigated a tradeoff of using intrinsic diversity as a retrieval objective versus optimizing for reduced user effort via keyword density maximization and found that interesting tradeoffs between learning gains and learning efficiency. Work by Syed and Collins-Thompson (2017b) built further on this by doing an even larger study with a

larger variety of topics assessed to evaluate the effect of optimizing for reduced effort with and without personalizing the selection of documents based on a user's prior knowledge. Later work by Abualsaud (2017) also found that self-reported novelty in Web documents correlated strongly with both user's knowledge gains and user satisfaction in a learning task.

Further work in using MMR-like retrieval models showed that incorporating such novelty could actually result in documents with more relevance to the user. Work by Radlinski and Dumais (2006) demonstrated that a search re-ranking algorithm that incorporated diversity through MMR offered the document set with the highest document relevance to the user. The study tested three algorithms for incorporating diversity in a document set and found that the Maximum Result Variety (MRV) algorithm that incorporated MMR performed the best where the relevance of a document to the user was given by a variant of the BM25 measure (Radlinski and Dumais, 2006). In a post-hoc study, Raman et al. (2013) investigated the performance of an algorithm that incorporated MMR in a greedy optimization for intrinsic diversity. They show that with an idealized source of subtopic/related queries, their algorithm was able to significantly outperform a baseline document ranking in terms of both precision and DCG measures (Raman et al., 2013).

It is important to note that the MMR criteria, as applied in an educational context, would be designed to reward different aspects of the query topic, rather than different *interpretations* of the query topic. The distinction between ambiguous and underspecified queries (Clarke, Kolla, and Vechtomova, 2009) is important as the primary focus of this work is on optimizing for query topics that can be assumed to be unambiguous. Resolving topical ambiguity is itself a separate field of work.

Clarke, Kolla, Cormack, Vechtomova, Ashkan, Büttcher, and MacKinnon (2008) developed a general algorithm for incorporating subtopic diversity which forms their modification

to the commonly-used nDCG measure in the form of the α-nDCG measure. The variant proposed in (Clarke et al., 2008) computes the discounted cumulative gain while considering the topic as a collection of subtopics compared to the original measure that considered atomic topical relevance. The authors refer to subtopics as "nuggets" - quantifiable properties of a document, often as dichotomous variables pertaining to aspects of the main query. Clarke et al. (2009) distinguish nuggets from aspects where the first is operational and the second is conceptual. Agrawal, Gollapudi, Halverson, and Ieong (2009) have also proposed variants of some common ranking evaluation measures such as MAP, nDCG and MRR by making them "intent-aware". These modifications retain the basic principles of the original measures but weight them by the distributional probability of the different categories that the query could have referred to. Unlike the example in (Clarke et al., 2008), the study by Agrawal et al. focuses on ambiguous queries rather than underspecified ones.

We note that a common requirement in most diversity-based work is the first step of detection and retrieval of subtopics for the given main topic. Although much prior work in the area has focused on diversification at the subtopic level (treating each subtopic as an atomic unit), recent work by Dang and Croft (2013) shows that the grouping of terms into subtopics is unnecessary. They demonstrate that they can achieve comparable results by standard IR measures when breaking the subtopic set of terms down to more fundamental units of simply keywords. A major contribution of this work is that this makes practical algorithms for diversity far easier to design and implement as subtopic extraction can be significantly more difficult than keyword extraction (Dang and Croft, 2013).

While some work have focused on subtopic diversity algorithms at a post-hoc level (Raman et al., 2013; Zhai et al., 2003), we need to consider methods for detecting these subtopics in real-time in a "cold start" situation. In particular, our focus is on methods to determine

and extract subtopic queries without any dependence on the user's prior search or Web browsing activity. Several techniques have been developed to accommodate this in recent works. Zhang, Lu, and Wang (2011) developed a subtopic ranking algorithm that extracted subtopics from three sources: (1) post-hoc query log analysis, (2) subtopic extraction from related encyclopedia entries and (3) from related search suggestions from major search engines. Later work by Raman et al. (2013) also tested subtopic selection techniques that were largely focused on post-hoc browsing log analysis but they also did evaluate the option of using related search suggestions from a major search engine. Their study showed that that option offered the worst performance by several measures of information retrieval. A more recent work by Collins-Thompson et al. (2016) extends the algorithm proposed in (Raman et al., 2013) for real-time application and used the Wikipedia article for the corresponding main query as a source for subtopics. The subtopics were constructed by augmenting the main query with the main headers in the Wikipedia article.

3.6 Document and Search Features that Improve Learning Outcomes

In the previous sections, we have discussed existing literature that has investigated how people learn during a search task along with their browsing and search session behaviors. In the previous section, we talked in more detail about how intrinsic diversity can be a useful feature for providing a set of documents that cover multiple aspects of a given topic while avoiding the problem of redundancy. In this section we will focus more specifically on the question of what type of document- and document-set-level properties are good indicators of learning outcomes.

Session features. Most work in the space of search as learning has focused on session-level and browsing-level activity that happens during search but few studies have put emphasis on what types of document properties are good predictors of learning outcomes in search. A large-scale log study by Eickhoff et al. (2014) investigated not only session-level behaviors but also some document properties that were trained as good predictors of learning outcomes (i.e. the user shifted from being an estimated novice to an estimated expert). A later lab-based study by Mao et al. (2018) investigated differences in search behavior, amongst other variables, when comparing between domain novices and experts. In investigating browsing or query features for estimating knowledge state, there have been plenty of other studies as well. Preliminary results by Palotti, Hanbury, and Müller (2014) showed that they could train a random forest classifier with only two features to get substantial improvement in classifying domain novices vs experts in the medical space. Later work by Zhang, Liu, Cole, and Belkin (2015) and Yu et al. (2018) also investigated session-level features that were good predictors of user domain knowledge and knowledge gains respectively. Besides for just predicting knowledge gains, other recent studies have looked at other variables in search as learning activity. A study by Liu et al. (2018) investigated the relationship between user satisfaction and search success during exploratory search as learning where they showed that people's perspectives of their learning success don't always align with their actual changes in knowledge state. Another study approached the question of learning in search from a different angle, investigating how people's search outcomes - both in terms of satisfaction and information gain - were affected by their choice of search service to use (Li, Liu, Cai, and Ma, 2017). The authors found that search behavior on Community Question Answering (CQA) sites were a strong indicator of positive search outcomes during a learning task (Li et al., 2017).

Document features. However, there may be use cases where session-level data is unavailable or not useful and where document properties alone need to be used. A recent study by Bulathwela, Yilmaz, and Shawe-Taylor (2019) introduced five general dimensions that affect the quality of a document, including Understandibility, Topic Coverage, Freshness, Presentation and Authority. Similarly a lab study by Abualsaud (2017) proposed a framework of six 'Learning factor' document features that could be predictors of knowledge gain in a learning task. These include dimensions of 'Understability', 'Readability', 'Broadness', 'Detailedness', 'Novelty' and 'Reliability'. In their study, only the factor of 'Novelty' showed significant correlation with knowledge gain. In this dissertation, I am primarily focusing on optimization towards the general metrics of understandability and topic coverage/novelty. Addressing the remaining components such as authority/reliability and presentation will be an area for future work.

A study by Syed and Collins-Thompson (2017b) demonstrated that for a vocabulary learning task, documents with a high keyword density feature were likely to result in stronger learning gains. A follow-up to that study by Syed and Collins-Thompson (2018) investigated a much larger set of features including features like word count, paragraph length and image count and how these features could predict learning outcomes through regression models. Earlier studies have also focused on isolating the specific effects of particular features on learning outcomes in a Web environment. Work by DeStefano and LeFevre (2007) and Zumbach and Mohraz (2008) showed that non-linearities in text content in Web resources could hurt the learning process. DeStefano and LeFevre (2007) specifically investigated how elements like hyperlinks on pages could hurt learning outcomes by adding extra and unnecessary cognitive processing. Zumbach and Mohraz (2008) investigated how linearity in the form of navigational structure and narrative style in expository text could influence

learning outcomes. There has also been work investigating whether the use of images in Web content helps or hurts learning outcomes. Early work by Mayer (1997) suggest that if the choice and selection of media elements is done properly, images could show positive association with learning outcomes. However, a more recent study by Freund, Kopak, and O'Brien (2016) compared learning outcomes between participants who got a plain text page and those who got the same text content but also with images and found that those who got images performed worse. Regression weights from models trained by Syed and Collins-Thompson (2018) also found an interesting relationship: when considering image count in terms of *all* images (by HTML tags), there was a negative coefficient with learning gains but this turned positive when manually excluding images that were either ads or navigational in nature. Thus, it is likely not the use of images itself but rather *which* images and *how* they are used that will influence learning outcomes.

It should also be noted that most of the studies in the space of search as learning don't investigate the phenomenon from the perspective of solving an optimization problem but rather one of understanding search behaviors and strategies and understanding what type(s) of document content or structure is better suited for supporting learning on the Web. However, it is also important to consider the perspective of optimization in order to proactively provide better results to begin with. Syed and Collins-Thompson (2017b) did some early work in this direction by re-ranking documents presented to users based on the keyword density feature they had identified. Their model for choosing how to re-rank and how many documents to provide were informed by greedily solving an optimization problem based on the Item Response Theory objective function (Syed and Collins-Thompson, 2017b). There remains substantial opportunity to test what has been found about learning in search through actively re-ranking content from existing commercial Web search engines and empirically

Feature	Type	Description
Non-linear content flow	Design	Implies heavier cognitive load (Zumbach and Mohraz, 2008). Negative relation to learning.
Non-Text Elements	Design	Can create extra cognitive load. Potentially negative relation to learning (Freund et al., 2016). However, multimedia does have potential to improve learning when applied correctly (Brock and Smith, 2007; DeStefano and LeFevre, 2007; Guo, Kim, and Rubin, 2014; Mayer, 1997).
Embedded links	Design	Can imply heavier cognitive load (for both deciding whether or not to click and the interruption creating by clicking) as a non-linear feature (DeStefano and LeFevre, 2007). Negative relation to learning.
Reading difficulty	Content	Generally relates negatively to learning outcomes (Liu et al., 2018; Marks, Doctorow, and Wittrock, 1974; Ng and Gunstone, 2002)
Difficulty-weighted Keyword Density	Content	Positively associated with learning outcomes. Tested as a retrieval objective for learning task (Syed and Collins-Thompson, 2017b). Outperformed commercial baseline rankings.
Document content novelty	Content	Positively correlated with learning outcomes Abualsaud (2017); Syed and Collins-Thompson (2018).
Perceived learning	Subjective	Positively correlated with actual learning outcomes Abualsaud (2017); Collins-Thompson et al. (2016).
Time spent reading document	Subjective	Positively correlated to actual learning outcomes Collins-Thompson et al. (2016).
Boredom	Subjective	Negatively associated with learning outcomes and task satisfaction (Baker, D'Mello, Rodrigo, and Graesser, 2010; Cordova and Lepper, 1996; Craig, Graesser, Sullins, and Gholson, 2004).

Table 3.1: Set of features found by prior studies to influence learning outcomes or predict knowledge level.

evaluating how well the re-ranked results improved learning outcomes relative to the baseline. There have been other studies that have trained models for predicting knowledge state in a search task (Zhang et al., 2015) as well as predicting knowledge gains in a search task

that also investigates perceived usefulness (Li et al., 2017; Liu et al., 2018). It is yet to be determined, however, if these regression models could be used to re-rank SERP results in ways that practically yield superior results to commercial baseline results by some metric of success.

In the following chapters, we will discuss a set of studies that address the principle objectives outlined in the Introduction (Chapters 4 - 10).

Chapter 4

Dissertation Overview

We have now reviewed existing work covering: (1) why a learning objective is so important for modern Web search engines; (2) what document features may affect learning outcomes; (3) the difference between short-term and long-term learning and the importance of being able to optimize for either type. In this chapter, we will detail completed studies that have already addressed a significant number of the research objectives we outlined in the Introduction (Chapter 1). The overarching goal that we aimed to accomplish in the following completed studies could be visualized in Figure 4.1. This target system would comprise multiple steps:

1. **Estimate User Prior Knowledge.** Capture the user's prior knowledge of the topic.
2. **Resource Selection.** Choose a subset of resources from a pool of resources for the user to learn from. In this thesis, the scope of the resource type is limited to Web documents but in future work other resources may include videos, pdfs and interactive applications.
3. **Maximize Knowledge Gain.** The overarching goal is to find the set of resources that maximizes the expected knowledge gain the user will achieve.

The following chapters describe various algorithms and frameworks that gradually build towards this goal. Broadly speaking, the completed studies provide the following contributions:

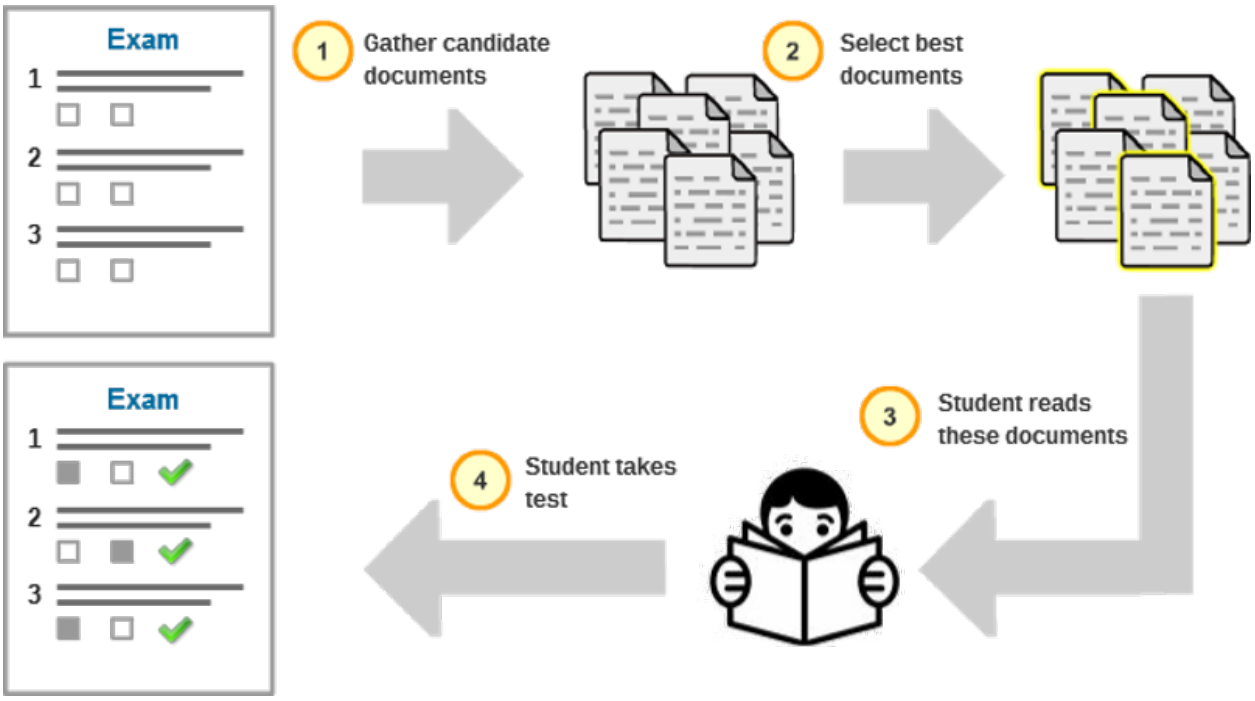

Figure 4.1: High-level overview of intended solution. The user first provides information about their prior knowledge. The system then chooses a subset of optimal candidate documents to provide the user. The user reads this material and takes a final test. Ideally, we want to find the best way to choose the documents subset such that the user's final test performance is maximized.

1. **Investigation of the Effect of Intrinsic Diversity on Search as Learning.** The first two studies presented here started by investigating how intrinsically diverse ranking of search results affected learning outcomes. This was first tested in a more open-ended learning task and then later tested in a more controlled vocabulary learning task (Chapter 5). We show that intrinsically diverse search results do show the potential to achieve better learning outcomes along with better learning gains per unit of content provided. The majority of these two studies were published in CHIIR (Collins-Thompson et al., 2016) and SIGIR (Syed and Collins-Thompson, 2016) respectively.

2. **Novel theoretical framework for optimizing Search as Learning.** We later

introduced a theoretical search results ranking framework explicitly designed to optimize expected learning gains. This framework optimizes document selection based on the expected learning improvements it would provide a user, calculated using Item Response Theory (IRT) models. We demonstrate that this model showed significant improvements in both learning gains as well as learning gains normalized by document length (Chapter 6). This second result was particularly impressive as the magnitude improvement of the normalized learning gains was nearly three times that of the Google Custom Search API baseline. This suggests that participants were able to accomplish the same learning gains even when being provided less than 1/3 the content that baseline participants got. This has significant implications for educational website design and search engine re-ranking principles. The majority of this study was published in SIGIR (Syed and Collins-Thompson, 2017b).

3. **Long-Term Analysis of Learning Gains.** We also extended the novel framework's results by conducting the first longitudinal crowdsourced study of learning from Web documents (Chapter 7). This study was conducted after a substantial time delay (nine months) and further strengthened earlier findings regarding our personalized algorithm. In particular, we found that long-term retention of more difficulty vocabulary terms was substantially higher in participants who got documents from our algorithm as compared to the baseline participants. This has strong implications in favor of our algorithm for not just supporting short-term but also long-term learning gains. The majority of this study was published in CHIIR (Syed and Collins-Thompson, 2018).

4. **Robust Regression Models of Learning.** We extended further on our novel framework by considering a larger possible feature set to investigate what additional document features influenced learning gains (Chapter 8). These included vocabulary-

specific features but also general structure and stylistic features of the documents and their content presentation. We found very promising results in terms of cross-validated predictive strength and further found that the strength of these findings persisted even after removing user-specific features. As almost all features can be computed efficiently and automatically at scale, these trained models can have great potential application for optimizing document selection for search as learning. The majority of this study was published in CHIIR (Syed and Collins-Thompson, 2018).

5. **Validation of Regression Models on Independent Datasets.** We further investigated whether or not we could train binary models of learning outcomes on one dataset of search as learning and evaluate it on independent datasets constructed from independent research. In general we found positive results in this direction where we found strong and significant test-set correlations.

6. **Investigation of Personalized Active Learning via Gaze Tracking and Automatic Question Generation.** Finally, we investigated the benefits of a form of active learning - the adjunct questions effect - when applying personalization and using automatically generated questions. We found strongly promising results from this study, suggesting the potential for scalable application of the adjunct questions effect to arbitrary text material. The majority of analysis in this chapter will be published in the Web Conference 2020 proceedings (Syed et al., 2020).

The above summary of results is now expanded on in deeper detail in the following sections.

Chapter 5

Role of Intrinsic Diversity on Learning in Web Search (Study 1)

As an early direction of research, we investigated how the concept of intrinsic diversity in a collection of Web documents could influence actual learning outcomes. Prior work by Raman et al. (2013) showed that many exploratory search topics could be represented by a discrete set of intrinsic subtopics and that an intrinsically diverse (ID) search ranking could provide users access to documents covering a range of subtopics earlier. However, it was unclear whether or not such a re-ranking would actually influence learning outcomes in a learning-oriented search sessions. In this chapter, I describe two studies we conducted that investigated how intrinsic diversity affected actual learning outcomes in two separate contexts. In the first study, we investigated the role of ID in an unconstrained Web search environment compared with standard search results. In the second study, we considered a special case of applying ID in a vocabulary learning context subject to an effort constraint.

5.1 Intrinsic Diversity in Web Search (Study 1a)

This study, detailed in (Collins-Thompson et al., 2016), was one of the first to investigate how multiple search query models, including one focused on intrinsic diversity in Web search results, affected user's actual learning outcomes. The choice of intrinsically diverse search in

that study was based on the idea that many exploratory search tasks often involve multiple queries which translates to extra effort and this could be reduced if multiple sub-topics of interest could be represented in one set of search results (Raman et al., 2013). While earlier work by Raman et al. (2013) which did investigate intrinsic diversity in Web search, their analysis was a post-hoc analysis on a large-scale query log. In practice, however, we don't have future knowledge of which documents the user will click at later timestamps. Furthermore, we didn't have a source of future queries the user would enter, from which we could acquire signals of important subtopics. As such, in our study (Collins-Thompson et al., 2016), we extracted the subtopics for a given query using the closest Wikipedia entry for that query string and extracting the main content headers as important subtopics. We could then apply the same algorithm as proposed by Raman et al. (2013) in a real-time context. Our implementation of intrinsically diverse search solved the following optimization problem, based on that proposed by (Raman et al., 2013):

$$\arg\max_{\mathcal{D}} \sum_{i=1}^{|\mathcal{D}|} Rel(d_i|q) \cdot Rel(d_i|q_i) \cdot e^{\beta\eta_i} \tag{5.1}$$

where $\mathcal{D}$ is the result set of documents to provide the user, $Rel(d_i|q)$ is the relevance of document d_i to the topic query q, $Rel(d_i|q_i)$ is the relevance of d_i to the subtopic query $q_i \in \mathcal{Q}$ (augmented queries from Wikipedia headers) and η_i is a maximal marginal relevance (MMR) tradeoff between relevance and novelty, specified in more detail in (Raman et al., 2013).

Results from this study indicated that the intrinsic diversity search condition (ID) did outperform two other models of searching that both involved simply providing default Google search results for a given query. Specifically, the knowledge gains resulting from the ID condition were generally stronger as compared to the other two conditions even when compared

across two search tasks that users could have been assigned to. Furthermore, users who got ID search results reported significantly better perceived search outcomes for one of the tasks. They also reported being able to "synthesize the various pieces of information together" significantly more in the ID condition (Collins-Thompson et al., 2016). This is consistent with what we would expect as the whole point of intrinsic diversity in this context was to provide the user content covering multiple aspects of a particular topic. Overall, the results from this study were very promising in indicating the potential learning usefulness of incorporating a model that had better coverage of topic components in Web search.

5.2 Intrinsic Diversity under Effort Constraints (Study 1b)

Building on these results, we investigated how intrinsic diversity in search, subject to effort constraints and an effort reduction extended model affected learning outcomes in a variety of topics (Syed and Collins-Thompson, 2017a). In this study, we investigated learning in the context of vocabulary learning (how well a user knows the definition of topic-specific vocabulary terms).

5.2.1 Teaching content representation and extraction.

We extracted the top 10 most representative unigrams for each topic ranked by a measure of weighted term frequency which rewards frequent term occurrences in a representative document $\mathcal{D}^*$ and penalizes their frequencies in a global corpus (GC)[1]. Specifically, we had a scoring function for each unigram in the representative document:

[1]We used the British National Corpus (BNC) as the global corpus

$$Score(u_i, \mathcal{D}^*) = \frac{\text{TermFreq}(u_i, \mathcal{D}^*)}{\log\left(\text{TermFreq}(u_i, GC)\right)} \tag{5.2}$$

We kept only the top $N = 10$ most representative keywords $K = \{K_1, \ldots, K_N\}$. Once these were extracted, we theorized that the more instances of a given keyword the user sees in some relevant sentence, the more likely they can triangulate the meaning of the keyword. By this theory, ideally the user should be exposed to ∞ instances of each keyword for maximized learning but this requires intractable effort. So we solved a simple effort-constrained optimization problem to determine a finite number of instances S_i of keyword K_i to provide the user.

Let each keyword have an associated weighted importance W_i such that keywords with greater weight are deemed more important to learn. These weights were computed as the number of occurrences of the keyword in a representative document and normalized as a multinomial distribution where: $\sum_{i=1}^{N} W_i = 1$. Further, to avoid providing an unreasonable number of instances of each keyword, we constrain the total sum of instances of all keywords to T where $T = \sum_{i=1}^{N} S_i$. In this study (Syed and Collins-Thompson, 2017a), T was manually chosen for each topic to avoid getting too many documents that might cause the participants to get frustrated (in a following study, this became partially automated). Finally we distributed the T keywords proportionally by weight. That is, we had: $S_i = T \cdot W_i$.

5.2.2 Document retrieval criteria.

As mentioned, we used an extension of the intrinsic diversity algorithm described in optimization problem (5.1). Specifically, we added another term ϵ_i to the objective that incorporated effort reduction via keyword density. The hypothesis was that if we gave preference to documents that had a higher ratio of instances of keywords to instances of any term, the user

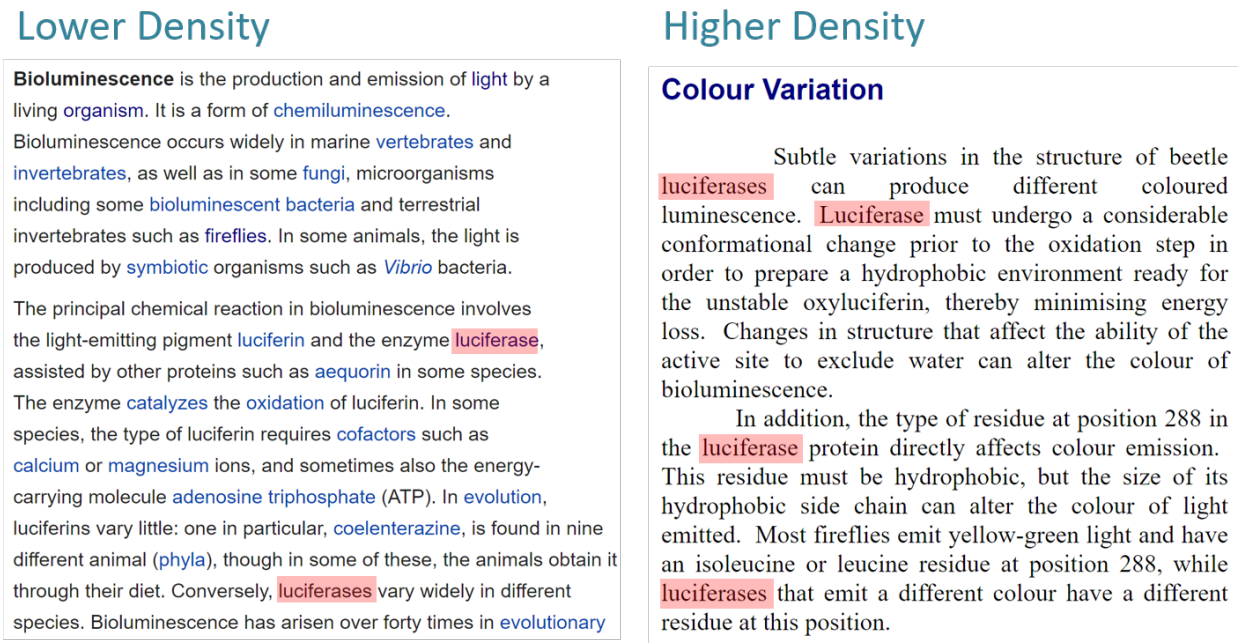

Figure 5.1: Two documents with different keyword density for keyword 'luciferase' (considering both singular and plural tenses). Left document has lower density; Right document has higher density.

would be exposed to more instances of relevant learning material with as little extraneous text, thus reducing their total unnecessary cognitive load (See Figure 5.1 for an example). Formally, this new objective was given as:

$$\arg\max_{\mathcal{D}} \sum_{i=1}^{|\mathcal{D}|} Rel(d_i|q) \cdot Rel(d_i|q_i) \cdot e^{\delta\eta_i} \cdot e^{\alpha\epsilon_i} \tag{5.3}$$

There are two main differences between problem (5.3) and problem (5.1). Firstly, the novelty measure is different: we now consider the cosine similarity between documents instead of only SERP snippets (Syed and Collins-Thompson, 2017a). Specifically, we compute η_i as:

$$\eta_i = \lambda \left[\cos(\text{snip}(q_i), \text{snip}(q))\right] - (1-\lambda) \max_{j<i} \left[\cos(d_i, d_j)\right]$$

where cos(a,b) is the cosine similarity of a and b and snip(x) is the bag of words representation of the top 10 snippets returned by query x (Syed and Collins-Thompson, 2017a). Secondly, we've added the ϵ_i term, with α as a parameter to control how much weight to give the keyword density term as compared to the intrinsic diversity score. Observe that $\alpha = 0$ reduces the problem to almost the same as problem (5.1).

The ϵ_i parameter is actually more nuanced than a simple keyword density calculation and is actually the normalized contribution that document d_i offers in terms of how much closer it brings the student towards reading the total required number of keyword instances (the S counts for each of the N keywords). Let $C_{\mathcal{D}} = \{C_{\mathcal{D}1}, C_{\mathcal{D}2}, \dots C_{\mathcal{D}N}\}$ be the set of keyword counts the student has cumulatively seen so far from documents in set $\mathcal{D}$, let $C_i = \{C_{i1}, C_{i2}, \dots C_{iN}\}$ be the set of keyword counts in document d_i and $|d_i|$ be the total word count of d_i. Then we have:

$$\epsilon_i = \frac{1}{|d_i|} \sum_{j=1}^{N} \begin{cases} C_{ij} & C_{ij} + C_{\mathcal{D}j} \leq S_j \\ \max(0, S_j - C_{\mathcal{D}j}) & \text{otherwise} \end{cases} \tag{5.4}$$

Now that we have established the optimization objective to aim for, we greedily select documents that maximize the objective, adding them to the set $\mathcal{D}$ in each iteration. Our stopping criteria is determined by the cumulative counts of each keyword, given by vector $C_{\mathcal{D}i} \ \ \forall i$ and the required minimum counts, given by vector $S_i \ \ \forall i$. In particular, at the start of each new iteration, we terminate when the following logical check yields True:

$$\nexists \, C_{\mathcal{D}i} : C_{\mathcal{D}i} < S_i \ \ \forall i$$

The details of this document selection and retrieval process are more formally expressed in Algorithm 1. From all of this, we can now construct a finite set of documents $\mathcal{D}$ that meet

the minimum necessary keyword instances constraints S and which are greedily optimal for intrinsically diverse, effort-reduced retrieval. Now, we discuss our evaluation of the results.

Algorithm 1: IntrinsicTeacher algorithm that ranks documents for the vocabulary learning task. First developed in Syed and Collins-Thompson (2017a).

Input: D_i as Google search results for subtopic query q_i for all $\mathcal{Q}$
C_{dk} given as vector of keyword counts in document $d_k \in D_i$.
$C_{\mathcal{D}}$ given as cumulative vector of keyword counts for each of keywords K covered in $\mathcal{D}$.
S given as vector of required keyword counts for keywords K.
Output: $\mathcal{D}$ as output document set

$\mathcal{D} \leftarrow \emptyset$
$C_{\mathcal{D}j} \leftarrow 0 \ \forall j \in C_{\mathcal{D}}$
while $\exists C_{\mathcal{D}j} : C_{\mathcal{D}j} < S_j$ **do** ▷ exit when all $C_{\mathcal{D}j} \geq S_j$
  $bestV \leftarrow 0$
  $bestD \leftarrow \emptyset$
  $C_D \leftarrow \emptyset$
  forall $q_i \in \mathcal{Q}$ **do**
    forall $d_k \in D_i, d_k \notin \mathcal{D}$ **do**
      $docV \leftarrow Rel(d_k|q) \cdot Rel(d_k|q_i) \cdot e^{\delta \eta_i} \cdot e^{\alpha \epsilon_k}$
      if $docV > bestV$ **then**
        $bestV \leftarrow docV$
        $bestD \leftarrow d_k$ ▷ document with highest $bestV$
        $C_D \leftarrow C_{dk}$
      end
    end
  end
  $\mathcal{D} \leftarrow \mathcal{D} \cup bestD$ ▷ append $bestD$ to output $\mathcal{D}$
  forall $C_{\mathcal{D}j} \in C_{\mathcal{D}}$ **do**
    $C_{\mathcal{D}j} \leftarrow C_{\mathcal{D}j} + C_{Dj}$ ▷ update keyword counts in $\mathcal{D}$
  end
end

5.2.3 User Study Design

Now that we have established the general flow of how to extract a set of representative keywords, how to choose aspect queries and what criteria to score candidate documents on, we can now evaluate how well the resultant document sets improve actual learning outcomes. In this study, the primary focus was on the effect of tweaking the effort penalty variable α. We assessed four levels of α in a partially between-subjects experiment design (partially, because for a particular topic, all participants had to be non-repeating but a participant could take part in single tasks on multiple topics, as they had not yet been exposed to that topic's learning material or question set). The four levels of α were $\alpha = [0, 80, 120, \infty]$. Note that $\alpha = 0$ largely restores the pure intrinsic diversity algorithm whereas $\alpha = \infty$ *removes* the entire intrinsic diversity part of the optimization and purely optimizes for keyword density ϵ_i. The specific values of $\alpha = 80$ and $\alpha = 120$ were chosen based on manual observations of the average maximum variation in the document sets produced by different levels of α across multiples of 40 when compared with the $\alpha = 0$ condition. We selected five distinct science topics, covering a range of domains: Igneous rocks (geology), Tundra (environmental science), DNA (genetics), Cytoplasm (biology) and GSM (telecommunications).

We prepared document sets for each of these five topics and for each of the four α conditions, resulting in a total of 20 unique conditions. To get a sufficient number of participants, we chose to use the Crowdflower platform to run our experiment where each unique condition was assigned 35 participants, yielding a total of 700 participants. Participants were offered US$0.04 per page (the equivalent of US$3.20/hr) for completing the tasks. For quality control, in addition to Crowdflower's proprietary mechanisms and 'gold standard' questions, we limited the participant pool to users from the U.S. and Canada, given the vocabulary-centric

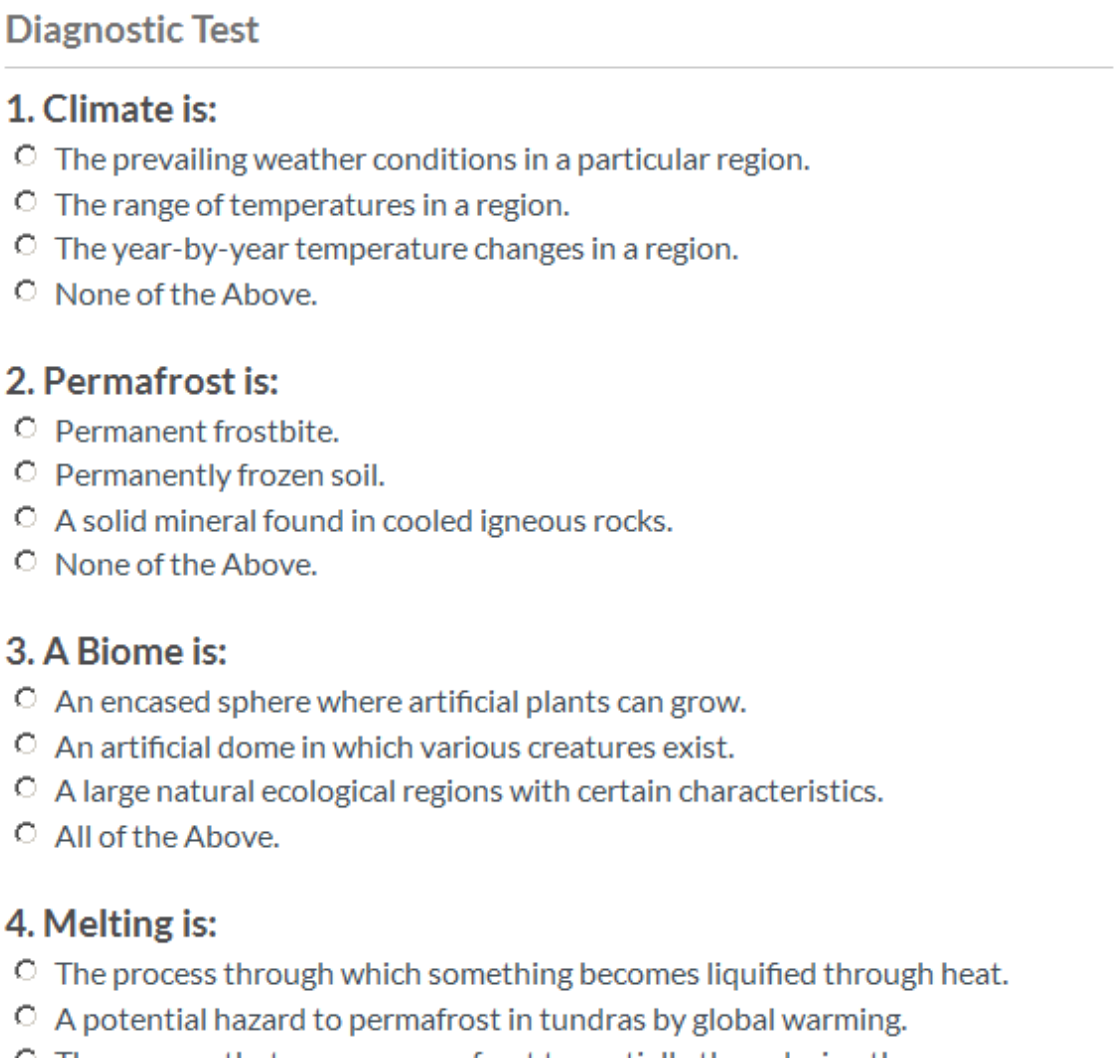

Figure 5.2: User study pre-test. Knowledge of each vocabulary term assessed through multiple-choice questions (Syed and Collins-Thompson, 2017a).

nature of the task and reliance on English reading skills. We also offered the tasks only to workers in the highest quality (level 3) pool, and only kept responses from those workers who spent at least four minutes on the task.

The task consisted of three stages: (1) Participants first completed a multiple-choice pre-test to assess their existing knowledge of the keywords; (2) then, based on the condition, read

through a provided retrieval set of documents containing the keywords to be learned; (3) finally, they completed an immediate post-test to assess their updated keyword knowledge. The design of the pre- and post-test is shown in Figure 5.2. Participants had to complete these stages in this ordered sequence and after progressing to the next stage, could not return to a previous stage. In the reading stage, participants had to click on and read all the links they were provided. There was no time limit explicitly provided to the participants but we manually excluded any who spent less than four minutes on the entire task as they likely did not take the task seriously. After applying all of our quality control filters, we ended up with a total of 447 participants out of the total 700. The following analysis is based on this subset of participants.

5.2.4 Results

In this section, we discuss the main results of the user study. In particular, we will discuss two measures of learning outcomes and how they differed by topics and by conditions. The pre- and post-test scores were recorded as binary responses to the multiple-choice definition questions ($Pre_k = 0$ or $Post_k = 0$ if the pre-test or post-test answer respectively for keyword k was wrong and $Pre_k = 1$ or $Post_k = 1$ otherwise). Then, we investigate the following two measures:

1. **Learning Gains.** Computed as the sum of improvements in knowledge over all keywords where one unit of learning gains is awarded when a participant learns a keyword ($Post_k = 1$) which they previously didn't know ($Pre_k = 0$). Specifically, the total learning gain (LG) is given as:

Topic	**α=0**	**α=80**	**α=120**	**α=∞**	**p-value**
Igneous rock	**1.55**	1.20	1.38	**1.55**	p=.727
Tundra	1.44	**1.852**	1.815	1.37	p=.473
DNA	1.71	1.55	**1.76**	1.57	p=.938
Cytoplasm	1.86	**2.90**	1.45	1.58	p=.012*
GSM	1.60	**2.50**	1.45	2.33	p=.064.
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 5.1: ANOVA analysis for learning gains across different α conditions. Bold values are maximum across conditions (Syed and Collins-Thompson, 2017a).

$$LG = \sum_{k=1}^{N} \begin{Bmatrix} 1 & Pre_k = 0 \text{ and } Post_k{=}1 \\ 0 & \text{otherwise} \end{Bmatrix}$$

2. **Learning Gains per Word Read.** Computed the same as Learning Gains but normalized by the total number of words the user was exposed to in the document set they were provided. This gives us a measure of the learning improvement as a function of how much effort was required to achieve that improvement. If the total words is given as $WordsTotal$, we have:

$$LGPW = \frac{LG}{WordsTotal}$$

Learning Gains. Firstly, in evaluating Learning Gains, we found that none of the four conditions were consistently better or worse across the five topics (Table 5.1), suggesting that even if certain settings of α were characteristically better for some topics, the effect was clearly not generalizable. We observe that only two topics showed statistical significance (only one if we strictly cutoff significance at the p=.05 level). As the remaining topics didn't show statistically significant differences across conditions, we consider these two topics for

this analysis.

Both of these topics showed a peak learning gain at the $\alpha = 80$ condition, suggesting that a combination of lowering effort via the keyword density parameter and rewarding intrinsic diversity in documents offers better learning gains than either factor alone. However, we also found that the setting of $\alpha = 120$ yielded the worst learning gains in those same topics. This suggests that the learning gains are quite sensitive to the particular choice of α and that choosing an α that combines both the ID objective and the keyword density objective is not always going to improve learning utility. It's not entirely clear why the specific value of $\alpha = 80$ offered better performance and it is a possible direction of future work to investigate this further and determine an algorithmic approach for determining the optimal α setting for a given topic. However, it is certainly concerning that most topics did not show significant differences, and arguably only the topic "Cytoplasm" showed true differences. As such, the overall result from this analysis is that the results may improve as a result of integrating keyword density but it appears to be a very sensitive tradeoff and does not appear to generalize well.

Since the target keywords ranged from more familiar to more technical, and learning gains could be expected to interact with keyword difficulty, we faceted the learning gain results by low- and high-difficulty keyword categories[2]. Figure 5.3 shows the result of averaging the learning gains for each keyword in the two difficulty categories and then averaging the results across the five topics. We see that there were learning gains in all conditions for both low- and high-difficulty keywords, but as expected, learning gains were higher for the higher-difficulty (and thus initially less familiar) keywords (one-way ANOVA differences in

[2]Keywords were split into two groups of five keywords according to their age of acquisition (AoA) score in a standard psychometric database. If a keyword didn't have an AoA score, it was assumed to be maximum difficulty.

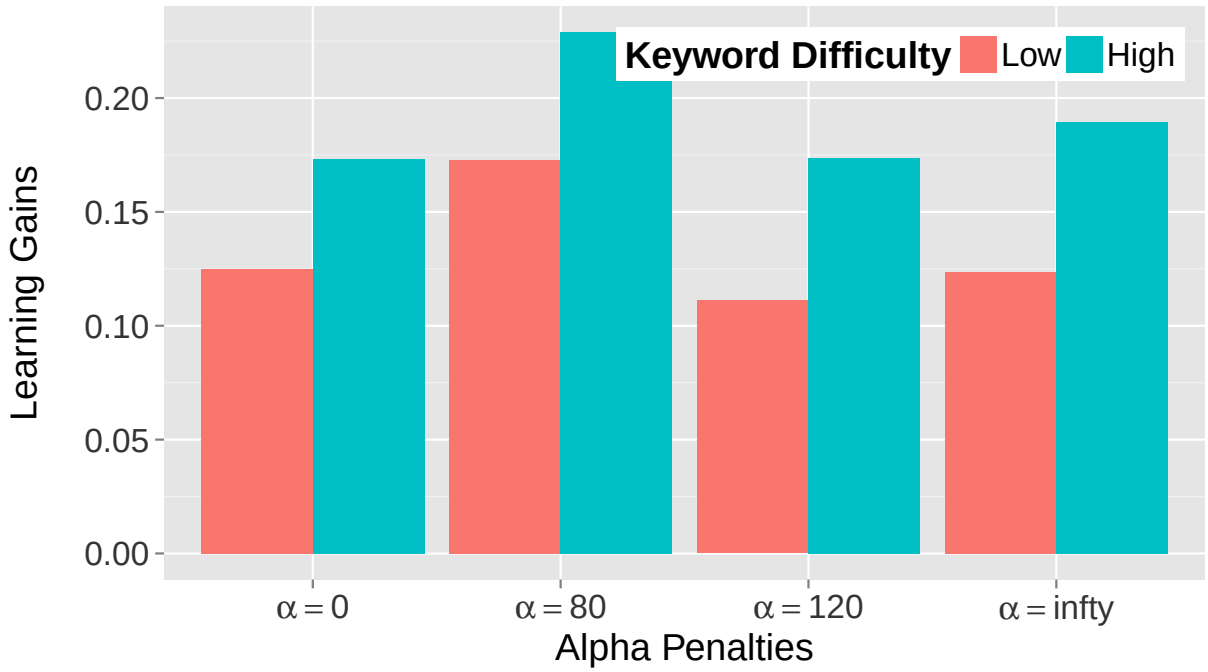

Figure 5.3: Learning gains were greater for keywords in the 'higher difficulty' category.

means between high and low difficulty words was statistically significant at the $p<.05$ level - tested for all four conditions).

Learning Gains per Word Read. In evaluating the Learning Gains normalized by total words read, we found a much more interesting result. The majority of the topics did show very strongly significant differences in means across the four α conditions and in three of the four topics that showed significant differences, the highest improvement was in the $\alpha = \infty$ condition. This much should have been expected in part because the $\alpha = \infty$ condition purely optimized for keyword density, a criteria that explicitly penalized documents that were lengthier. However, what is interesting in these results is that it appears that by selecting shorter documents, the participants' ability to learn the required content was not significantly impaired.

Topic	**α=0**	**α=80**	**α=120**	**α=∞**	**p-value**
Igneous rock	0.176	0.116	0.174	**0.316**	p=.001**
Tundra	0.093	0.203	0.138	**0.210**	p=.007**
DNA	0.234	0.203	0.206	**0.276**	p=.546
Cytoplasm	0.558	**0.811**	0.361	0.451	p=.006**
GSM	0.167	0.315	0.249	**0.614**	p<.001***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 5.2: ANOVA analysis for learning gains per 1000 words. Bold values are maximum across conditions (Syed and Collins-Thompson, 2017a).

We note that one topic, Cytoplasm, showed an opposite trend where higher alpha values mostly lead to worse normalized learning gains. We hypothesize that this may be because the total number of words used in each condition for Cytoplasm were significantly lower (almost half as many for $\alpha = 0$ and $\alpha = 80$) compared to the four other topics. It is thus possible that the positive impact of choosing higher α values is only effective after passing a certain threshold of minimum reading material.

5.2.5 Image Coverage vs. Keyword Density

To gain more insight into why pages with increased keyword density might contribute to more efficient learning, we investigated additional properties of the page content that might be correlated with keyword density. We found that while few result documents made use of multimedia such as animations, audio or video, a number did use images to supplement the text. Thus, the *picture superiority effect* (De Angeli, Coventry, Johnson, and Renaud, 2005), in which people tend to remember things better when they see pictures rather than words, could be relevant, since we were testing fact-based learning, which relies at least partially on recall. We thus examined whether there was a relationship between image coverage – defined as total images divided by total words – as a function of α. We determined the number of

relevant images manually for each page, excluding irrelevant images such as navigation icons and advertisements. We found that pages with higher keyword density did indeed tend to have increased image coverage, as shown in Figure 5.4. For three of the five topics, the highest image coverage is in the $\alpha = \infty$ condition.

We consider the possibility that a heavier coverage of images in teaching documents can improve learning outcomes regardless of condition. There is partial evidence of this in that ANOVA analysis of the topics "Igneous rock", "Tundra" and "DNA" showed no statistical significance in means (Table 5.1) and these three topics had the top three average image coverage (.0024, .0026 and .0034 respectively). On the other hand, the two topics that showed significant differences ("Cytoplasm" and "GSM") had the lowest coverage (.0015 and .0006 respectively). As such, it is possible that a higher image coverage can collectively improve or worsen learning gains regardless of conditions. Determining if the presence or absence of images actually has such an effect warrants further investigation.

We observe informally that pages using a higher density of keywords tend to be those that give an overview of topic for instructional purposes, and thus are more likely to be supplemented with images by the author. We intend to investigate this phenomenon and other content properties that may interact with learning in future work.

Because each condition lacked any variation in keyword density or image coverage (each condition produced only one distinct set of documents), we could not determine with this information alone if keyword density or image coverage was responsible for the learning gains improvement. However, we did conduct a follow-up study, that is currently unpublished (Syed and Collins-Thompson, 2017b), using the same framework but with some altered parameters where we tested three conditions, one of which was the $\alpha = \infty$ condition personalized relative to the participant's pre-reading scores (this simply means that the required s

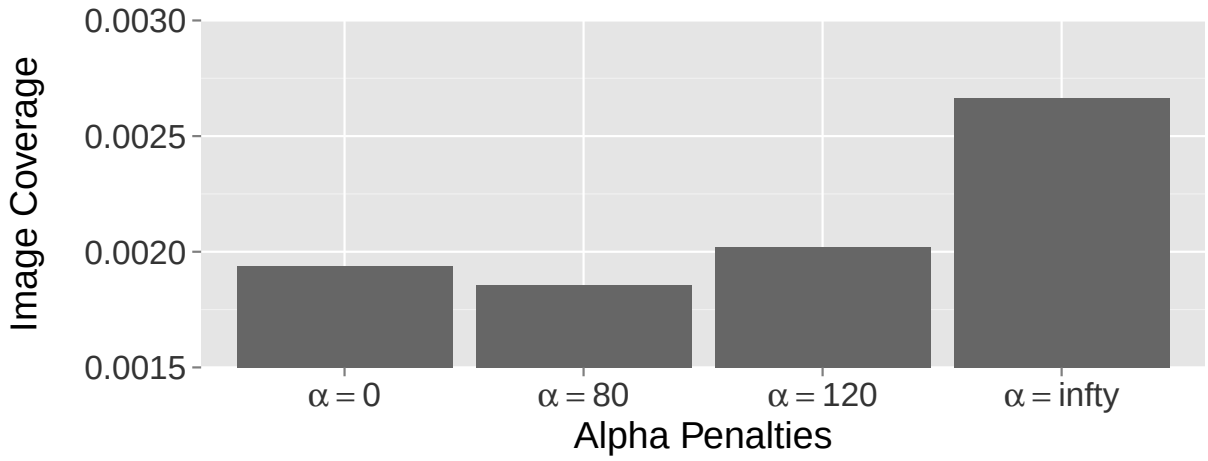

Figure 5.4: Higher α penalty generally results in documents with higher image coverage.

counts were modified to reflect what the participant already knew). This allowed for many data points of different keyword densities, image coverages and learning gains. We aggregated all participants in the personalized condition and created a two-by-two split of learning gains by median image coverage (lower (n=141) and higher (n=142) than median) and median keyword density (lower (n=141) and higher (n=142)) of the assigned document sets. We then conducted a two-way ANOVA with learning gains as the dependent variable to test for interactions between keyword density and image coverage. We found that there were no significant interactions (p=.36) and that image coverage did not yield significant differences in learning gain (p=.84). However, we did find that keyword density did yield significant differences (p=.01), suggesting that it was in fact changes in keyword density that yielded the learning gain improvements.

We also note that both image coverage and keyword density are measures that are normalized by total words in the document set. By removing this normalization, we repeated the above analysis with total images seen vs total keywords seen. We found that the interaction was still insignificant (p=.35) but that total keywords was now insignificant as well

($p=.27$) whereas total images was strongly significant ($p<.001$). This suggests that if we don't factor in the effort the participant has to spend in learning, simply looking at the total keywords they have read won't have any predictable effect on learning outcomes. However, this also shows that regardless of how much a user has to read, the more images they get to see, the better their learning outcomes will be. It might be worth noting that in the follow-up study - from where we're getting this data - the keyword density term additionally penalized documents that had higher vocabulary difficulty levels.

5.2.6 Conclusions

From these results, we can conclude that the choice of α certainly can affect the learning outcomes for one or two topics but does not appear to generalize to other topics and doesn't follow any evident trend where higher or lower α is consistently better or worse. This does, however, indicate that keyword density as an additional parameter does have the potential to improve learning outcomes for users who read resulting document sets. To this point, we further determined that normalizing the learning gains by the total words read showed a much more interesting picture where most participants showed strongly improved learning per word read, suggesting a potential retrieval formulation that could very likely reduce undesirable cognitive load on the users in a learning task while not impairing their learning effectiveness. In the next study we will discuss, we build on this very important conclusion and construct a further generalized framework that supports personalization and a more sophisticated method of determining the minimum necessary instances of a keyword to see.

Chapter 6

General Framework for Learning on the Web (Study 2a)

In this study, we built on results, primarily from the findings of Syed and Collins-Thompson (2017a), which found that optimizing the selection of documents to provide users by maximizing an effort-reducing function yielded strong improvements in learning gains per unit effort (assuming total words read as a measure of effort). This study expands the earlier study (Syed and Collins-Thompson, 2017a) by: (1) introducing a general framework for supporting search as learning; (2) incorporating the concept of personalization in the document selection procedure; (3) modifying the effort-reduction function to incorporate vocabulary difficulty of a document; (4) conducting a larger user study with twice as many topics; (5) evaluating the effects of personalization and establishing a commercial search engine's results as a baseline (Syed and Collins-Thompson, 2017b).

6.1 Overall Framework

In this study, we introduced a multi-stage high-level framework for how the system would operate, illustrated in Figure 6.1. Note that most of these stages have already been discussed in the preceding section with the exception of the Student model (Step 2). Other steps, like Steps 3 and 4 - the primary focus of this study - were introduced in the preceding section but were far more simplified. In this section, we detail the overall framework, its components and

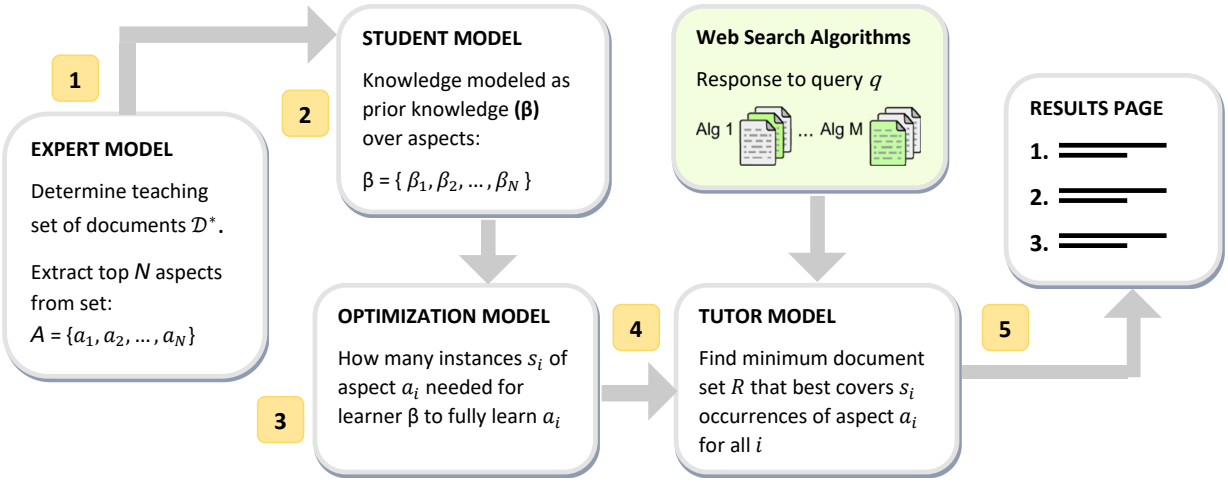

Figure 6.1: High-level learning-oriented optimization process.

how each stage of Figure 6.1 fits into the overall framework. We begin by describing what we consider as four fundamental components for our Intelligent Tutoring System (ITS): (1) an Expert model; (2) a Student model; (3) an Optimization model; (4) a Tutor model.

6.2 Expert model

The expert model, (Step 1 of Figure 6.1), is responsible for curating the set of documents $\mathcal{D}^*$ that best represents the knowledge aspects A_k of the subject. As in the preceding study (Syed and Collins-Thompson, 2017a), the aspects were represented as the N most representative unigrams for the topic. Also, as in the preceding section, we determined the representative document(s) manually. It is for a future work to construct an automated approach for detecting the most representative document.

To extract the N most representative keywords, we used the same approach as before (Equation 5.2) but we extended it with an additional variable. To avoid getting keywords that may be rare but not topically relevant, we weighted these values by their word2vec

Mikolov, Chen, Corrado, and Dean (2013) similarity to the first term in the base query q. Specifically, for each unique word u_i in the bag-of-words of $\mathcal{D}^*$ we determine the important words by the score:

$$Score(u_i, \mathcal{D}^*) = \frac{\text{TermFreq}(u_i, \mathcal{D}^*)}{\log\left(\text{TermFreq}(u_i, GC)\right)} \cdot word2vec(u_i, q)$$

In selecting the top N-scoring unigrams, we added an additional constraint where we skipped words that were semantically too similar to an earlier ranked word (word2vec similarity > 0.3) as they were likely the same word with a different tense/form (e.g. 'rock' and 'rocks').

From the extracted set of keywords K, we then generated the weight vector W as before, as the maximum likelihood estimation of these keywords' TermFreq values. Specifically, for $i = 1, \ldots, N$:

$$W_i = \text{TermFreq}_i \cdot \left(\sum_{j=1}^{N} \text{TermFreq}_j\right)^{-1}$$

For example, for the subject "igneous rocks" and $N = 5$, we get the distribution $W =$ {'igneous':0.302, 'magma':0.178, 'felsic':0.057, 'mafic':0.069, 'rocks':0.394}. Table 6.1 shows the top 5 keywords out of $N = 10$ for five different topics along with their corresponding weights.

6.3 Student model

The Student model (Step 2 of Figure 6.1), represents the knowledge state of the student who is learning about the topic given by query q. To find documents that teach the student, we can simply find the set of documents that minimally reaches the required set of counts S

Topic	**Keyword 1**	**Keyword 2**	**Keyword 3**	**Keyword 4**	**Keyword 5**
Igneous rock	rocks (.31)	igneous (.24)	magma (.14)	minerals (.08)	basalt (.06)
Tundra	tundra (.35)	arctic (.21)	plants (.13)	permafrost (.09)	soils (.08)
Phrenology	phrenology (.38)	brain (.16)	skull (.10)	science (.08)	perception (.07)
Pottery	pottery (.52)	clay (.15)	pots (.11)	potters (.06)	ceramic (.06)
Synapse	neurons (.39)	electrical (.17)	axon (.10)	synapse (.08)	membrane (.07)

Table 6.1: Top 5 (out of 10) selected keywords for five topics, sorted by descending keyword weights W_i. The keywords to be learned range from easy ('rock') to technical ('permafrost').

as we did in the earlier study (Syed and Collins-Thompson, 2017a) subject to some basic retrieval criteria. However, this approach ignores: (1) the fact that document length or keyword coverage is not necessarily indicative of topical relevance or quality and (2) different students may already know about certain aspects of q and their time would be better spent learning the aspects that they don't know.

We assume that we can measure a student's learning outcome in terms their performance on a test on the given subject, so that we can assess learning by measuring a learner's performance on a set of N test questions $T = \{T_n\}$ on those aspects (keywords). We code the learner's responses via the set Y of binomial variables Y_k such that:

$$Y_k = \begin{cases} 1 & \text{student answered } T_k \text{ correctly} \\ 0 & \text{otherwise} \end{cases}$$

This is similar to how we encoded user knowledge in the pre- and post-test in the user study of (Syed and Collins-Thompson, 2017a) (Chapter 5.2.4). We also make the assumptions that the student is a Bayesian learner and has no memory loss (post-reading knowledge is never less than pre-reading knowledge). We further assume that reading an instance of keyword K_i will monotonically increase the student's knowledge of that keyword (Step 3 of Figure

6.1). Let the student's prior knowledge β be a vector of how many instances of each keyword we expect them to have read before being provided $\mathcal{D}$. Then, we have:

$$\beta = \{\beta_1, \beta_2, \ldots, \beta_N\}$$

We assume the widely-used *item response theory* (IRT) model (Junker, 1999) as our cognitive learning model that defines the probability of a correct response Y_k on test T as a logistic function of user and item parameters:

$$P(Y_k = 1 \mid U, W_k, \beta_k, D_k, S_k) = \left(1 + e^{-f(U, W_k, \beta_k, D_k, S_k(\mathcal{D}))}\right)^{-1}$$

Here, the IRT model parameters are:

- U - The user's individual learning rate. This is defined such that the faster a student can learn, the less resources they will require to complete their understanding of q. In this study, we assumed a fixed U for all users.
- W_k - The weight given to term k where W is the weight multinomial defined in the Expert model. Terms with higher weight assigned are more important for the student to learn and hence are assigned higher number of s_k.
- β_k - The student's prior knowledge of keyword k, measured by the number of instances of k the student has already seen before being provided the document set $\mathcal{D}$.
- D_k - A parameter that quantifies the difficulty of learning for keyword k. Similar to the vector W, this is a multinomial.
- $S_k(\mathcal{D})$ - The target instances of keyword k the student sees in document set $\mathcal{D}$.

Now we define the function $f(\cdot)$ to be the log weighted sum of the total instances of the keyword the student has learned (prior knowledge + post-reading knowledge):

$$f(U, W_k, \beta_k, D_k, S_k(\mathcal{D})) = \log\left((\beta_k + S_k) \cdot (1 - D_k) \cdot U\right)$$

With these operational settings, we can then more specifically define the expected learning for the k^{th} term as:

$$P(Y_k = 1 \mid U, W_k, \beta_k, D_k, S_k) = \frac{1}{1 + \exp[-\log((\beta_k + S_k) \cdot (1 - D_k) \cdot U)]}$$

Observe that W doesn't appear in these formulations. This is because the W vector's importance only applies when considering all topic components together where different topics get different weights in estimating the user's aggregate knowledge of the topic. So when we consider the user's average knowledge as an aggregate of these probabilities, we get:

$$P(Y = 1 \mid U, W, \beta, D, S) = \sum_{k=1}^{N} \frac{W_k}{1 + \exp[-\log((\beta_k + S_k) \cdot (1 - D_k) \cdot U)]}$$

While this formulation supports an implementation that incorporates the keyword-specific parameters W and D as well as the user-specific parameter U, we kept these three parameters as constant for our study to avoid too many confounds. So by omitting these variables in our study, the user's average knowledge simplified to:

$$P(Y = 1 \mid \beta, S) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{1 + \exp[-\log(\beta_k + S_k)]}$$

We have thus far described the representation of the student's knowledge and what document content properties (keyword instances) can affect the expected knowledge. Recall how

in the previous study, we manually set the total number of instances T and correspondingly, the total number of each instances as $S_k = T \cdot W_k$ (Section 5.2). However, a more algorithmic approach to determining the S_k values as well as the total instances value T is more appropriate and more scalable. We now describe the optimization model we used to determine the optimal distribution of S_k values on an effort-reduction principle.

6.4 Optimization model.

In the Student model, we established a function of estimating the average user's knowledge in which the only document-dependent parameter was the vector S. Let the aggregate user's knowledge as a function of how much instances they have read thus far be $H(\beta, S)$ where:

$$H(\beta, S) = \sum_{k=1}^{N} \frac{1}{1 + \exp[-\log(\beta_k + S_k)]}$$

Observe that if we treat this as an optimization problem of trying to find the best value of S, the result would trivially be $S_i = \infty \quad \forall i$ as this is an unconstrained optimization problem. To fix this, we must add an effort constraint. While there could be many ways to formulate a function of effort, we chose to consider total number of instances of keywords as a measure of effort. Using this measure we can now directly constrain the optimization to force a finite vector S as the optimal solution. Specifically, we now have the following optimization problem to solve:

$$\arg\max_{S} \sum_{k=1}^{N} \frac{1}{1 + \exp[-\log(\beta_k + S_k)]} - \lambda \sum_{k=1}^{N} S_k \tag{6.1}$$

Figure 6.2 illustrates a simple instance of this optimization problem that shows the learning/effort tradeoff based on a topic with two keywords, using the sigmoid objective above.

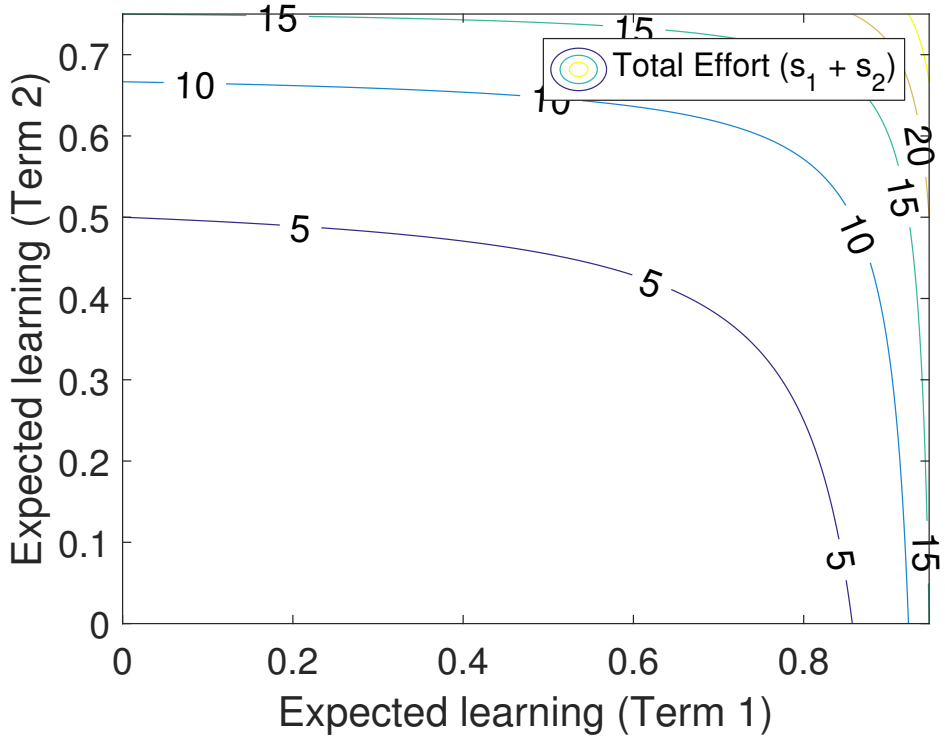

Figure 6.2: Possible tradeoffs in expected learning for each of two keywords (Term 1, Term 2) in a topic. Isolines show points of constant effort (total keyword instances read). Expected learning for each keyword is based on the logistic IRT definition above. (Ease of learning parameters for each keyword are set to $L_1 = 1.2$ and $L_2 = 0.2$ respectively.)

For a fixed total number of keywords to be read (shown by the isolines with total effort of 5, 10, 15, etc), there is a opportunity cost: for every additional unit assigned to one keyword, there also may be an expected potential loss in knowledge caused by *not* assigning the user's attention to the other keyword. In general these tradeoffs will also be affected by the ease of learning $L_k = 1 - D_k$ for the k^{th} keyword such that words that are easier to learn will result in a higher expected learning for the same total units assigned. For example, Fig. 6.2 shows that for a total effort of reading $S_1 + S_2 = 5$ keyword instances, to get the same expected learning for both keywords (i.e. the point that intersects the line $y = x$), we would assign $S_1 = \frac{1}{1.4}$ and $S_2 = \frac{6}{1.4}$ to get an expected learning outcome (probability of a correct test

result) of 0.461 for both keywords.

This optimization problem still required us to determine an optimal setting of λ manually and it is for future work to determine an algorithmic approach to choose the best λ penalty. In our study, $\lambda = 0.0060$ based on simulated resultant S_k values. While this optimization problem could be solved by standard SDP solvers, an implementation using the full set of parameters (including W, U and D) would form a more complex sum-of-sigmoids optimization problem. For such a case, more specific optimization solving methods would be appropriate such as the method proposed in recent work by Udell and Boyd (2013).

6.5 Tutor model

At this stage in the process, we have now constructed a representation of the keywords to teach the users and determined the set of minimum number of instances of each of those keywords in the vector S. We are now at the final stage (Step 4 of Figure 6.1) where we need to select the documents to provide the user.

The tutor model for this study followed largely the same algorithm described in the preceding model (i.e. Algorithm 1). The main difference was that in this study, because we added personalization into the optimization model, the S values were now user-dependent. The only structural difference was that we modified the ϵ_i keyword density variable to also penalize documents that used more difficult vocabulary. This was based on empirical results where we found that the keyword density variable itself was finding shorter pages but these sometimes included research article abstract pages which were dense in keywords but were also likely of little benefit to a novice to the domain. We used the Age-of-Acquisition model vocabulary scores r from the extended dataset by Kuperman, Stadthagen-Gonzalez, and

Brysbaert (2012) for this purpose. Specifically, if the vocabulary difficulty of unigram u_k in document d_i is given as r_k, then using the same notation we used for Equation 5.4, we have the *difficulty-weighted keyword density* as:

$$\epsilon_i = \left(\sum_{k=1}^{|d_i|} r_k \right)^{-1} \sum_{j=1}^{N} \begin{cases} C_{ij} & C_{ij} + C_{\mathcal{D}j} \leq S_j \\ \max(0, S_j - C_{\mathcal{D}j}) & \text{otherwise} \end{cases} \tag{6.2}$$

6.6 User Study Design

The Tutor model can now personalize the choice of documents to select to better meet the remaining gaps in knowledge of a specific user and do so in an efficient way using the ϵ variable. After we had established this new framework, we conducted a user study to evaluate the effectiveness of our model relative to a commercial Web baseline ('Google') and to further evaluate the value of personalization versus non-personalization. As in the previous study (Syed and Collins-Thompson, 2017a), we again tested on Crowdflower with the same quality control settings and same interface design.

For information needs, we developed a set of ten topics that were selected from top-level categories of the Open Directory Project to cover a range of areas, each having distinctive technical/expert vocabulary: Igneous rocks (geology), Tundra (environmental science), Cytoplasm (biology), Bioluminescence (biology), Phrenology (pseudo-science), Pottery (crafts), Cooking (food), Synapse (neuroscience), Refraction (optics) and Phenology (temporal phenomena). For each of these topics, we tested three conditions of document retrieval models:

1. Commercial Web search baseline ('Web'). The participant was simply provided the top Google Web search results for the topic using the Google Search API. Documents were only added until the stopping criteria in Algorithm 1 was met.

2. Non-personalized learning-optimized retrieval (α_N). The participant was provided a document set retrieved through the full Algorithm 1 with α parameter set to ∞. The $\alpha = \infty$ condition simply means that the difficulty-weighted keyword density ϵ_i term becomes the only factor in the ID retrieval objective. In this condition, we don't personalize results, so we assume all users had zero knowledge: $\beta = \{0\}$.

3. Personalized learning-optimized retrieval (α_P). The participant was provided a document set retrieved as defined above but with personalized S values calculated based on their prior knowledge β, computed from their pre-test scores.

In total, we had ten topics and three retrieval conditions, resulting in a total of 30 unique conditions. Participants were randomly assigned a condition through Crowdflower's proprietary random assignment. We gathered 40 unique contributions per condition, resulting in a total of 1200 total learning tasks completed by participants. After filtering out those who didn't pass the quality controls, we ended up with 863 participants, roughly evenly split across the three retrieval conditions ('Web': 290, 'α_N': 290, 'α_P':283).

6.7 Results - Learning outcomes

In this user study, we evaluated the following three research questions (Syed and Collins-Thompson, 2017b):

RQ1: Does learning-optimized retrieval framework offer higher learning effectiveness or efficiency compared to traditional retrieval results of a baseline commercial Web search engine?

Measure	**Web**	α_N	α_P	**p-value**
Absolute Learning Gains	1.721	1.831	**1.982**	p=.046*
Learning Gain Per 1000 Words	0.109	0.252	**0.347**	p<.001***
Realized Potential Learning	0.384	0.425	**0.471**	p=.008**
Time Per Word	12.007	29.176	**35.022**	p<.001***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 6.2: Aggregated averages of key learning-related measures. Bold values are maximum across conditions. (All tables use same significance codes and bold meaning.)

	Learning Gain			Learning Gain/Word		
Topic	**Web**	α_N	α_P	**Web**	α_N	α_P
Igneous rock	1.769	2.533	2.364	0.096	0.150	**0.311***
Tundra	2.115	1.655	2.231	0.145	0.280	**0.321***
Cytoplasm	1.567	1.577	1.758	0.057	0.083	**0.214****
Bioluminescence	1.929	1.808	1.567	0.127	0.319	**0.483*****
Phrenology	1.156	1.424	**2.097****	0.082	0.185	**0.430*****
Phenology	1.222	**2.036***	2.033	0.165	0.315	**0.356*****
Synapse	2.071	2.233	2.267	0.063	0.100	**0.175****
Pottery	2.156	1.710	1.600	0.121	0.481	**0.718*****
Cooking	1.407	1.471	1.957	0.057	**0.344*****	0.131
Refraction	1.824	1.957	2.107	0.170	0.240	**0.270**.

Table 6.3: Absolute learning gains (left) and learning gains normalized per 1000 words (right) averaged across different conditions and topics.

RQ2: Do personalized search results that account for a user's prior knowledge improve learning effectiveness or efficiency?

RQ3: How do learning effectiveness and efficiency vary across different topics (information needs) in different domains?

We now discuss the results of our findings using the same two measures of learning outcomes that we looked at in the earlier study (Chapter 5.2.4): (1) Learning Gains; (2) Learning Gains Per Word Read. We will then investigate other interesting results we found.

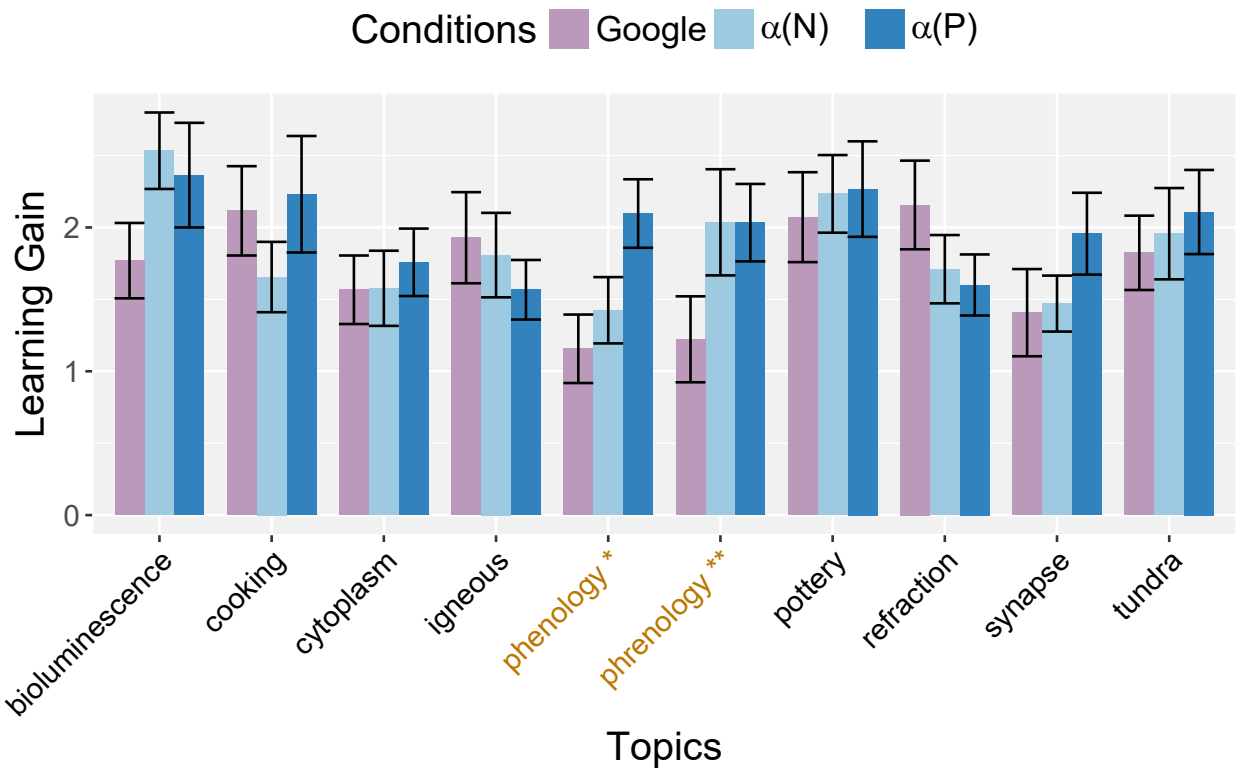

Figure 6.3: Breakdown of average learning gains by topic and condition. Error bars are standard errors.

Learning Gains. Recall that Learning Gains is simply the sum of instances where the user did not know the definition of a keyword in the pre-test ($Pre_k = 0$) and did know the definition in the post-test ($Post_k = 1$). In the earlier study, we found that the $\alpha = \infty$ condition was never the condition that yielded the optimal Learning Gains for any topic that showed significant differences. Overall, there actually were significant differences ($p<.05$) in Learning Gains when aggregating all topics (Table 6.2). The personalized condition α_P showed an approximately 15% improvement over the commercial baseline overall, suggesting that our model does show overall improvement over existing state-of-the-art. However, on closer inspection, we find that this result was not consistent across each topic, where two

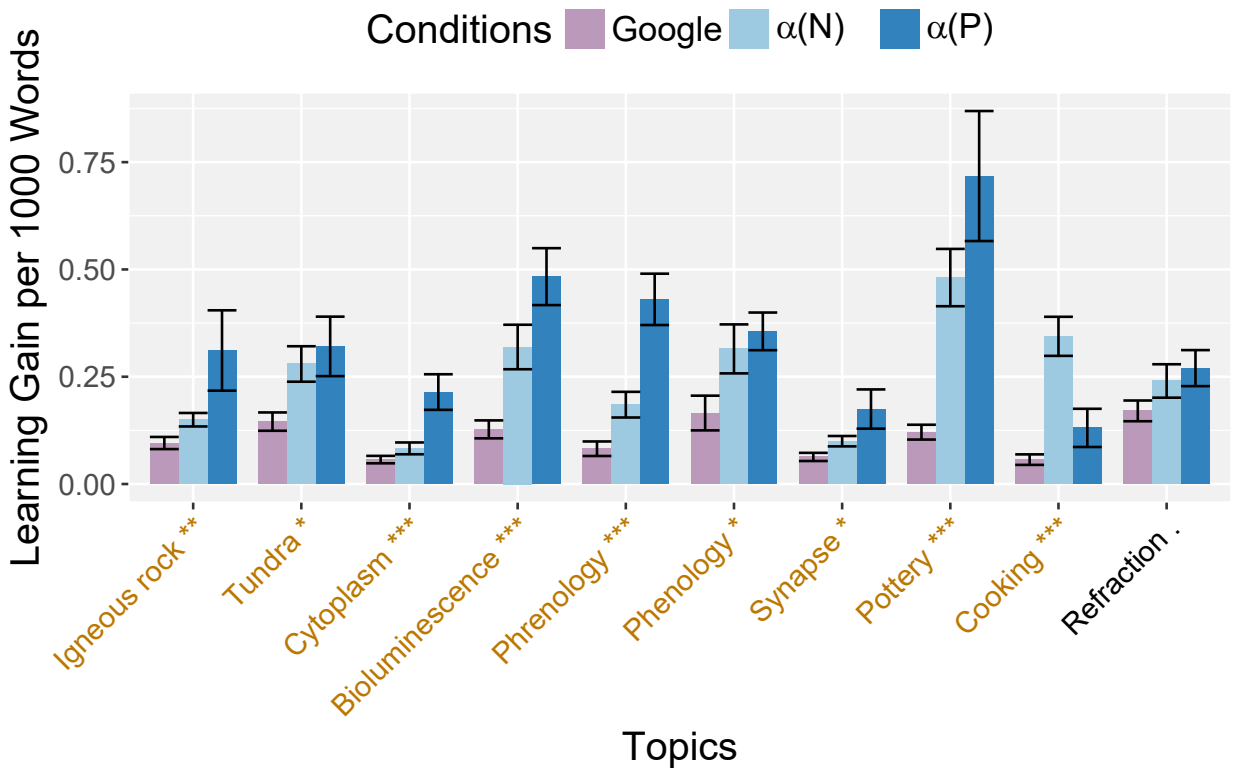

Figure 6.4: Breakdown of average learning gains per word read by topic and condition. Error bars are standard errors.

of the ten topics showed significant differences in means (Figure 6.3). It is possible that this result indicates that the keyword density optimization does do better, but possibly only for more obscure topics. Of all ten topics, the two that showed significant differences, "Phrenology" and "Phenology" were those that also happened to have the lowest Google search result count.

It is also possible that we simple required a larger sample size at the individual topic level to have detected significant differences. When comparing between the personalized condition α_P and the baseline Web, the Cohen's d effect size was strongest for the two topics "Phrenology" (d=.704) and "Phenology" (.537) with the next highest effect size being in the

topic "Igneous rock" (d=.384) and the average effect size being (d=.313). For the significance level p<.05, power $1-\beta = 0.80$, and using the Mann-Whitney U test, the number of samples needed in each of the two retrieval conditions to detect a difference would be n=133 (Faul, Erdfelder, Lang, and Buchner, 2007). This is substantially higher than the average number of participants we ended up with in each of the two conditions (n=29), suggesting that the study design was strongly underpowered for the effect size that was likely to show. This suggests that having a substantially larger sample size for each of the topics could have resulted in far more significant observed differences between retrieval conditions.

Learning Gains per Word Read. Recall that we defined Learning Gains per Word Read to simply be Learning Gains divided by the total word count in the document set the user was provided. In the earlier study, we found that the $\alpha = \infty$ condition significantly outperformed all other values of α in LGPW. In this study, we again found a very strong overall improvement in LGPW with the personalized condition outperforming the Web baseline by a factor of 3.18, suggesting that those in the α_P condition were able to accomplish the same learning improvement being provided less than $\frac{1}{3}$ the total content to learn from. Furthermore, unlike Learning Gains, we found this effect to be consistently strong across most topics (9 out of 10) and in almost all topics, α_P was the best performing condition (8 out of 9 significant topics) (Table 6.3 and Figure 6.4). We note that the values reported for Learning Gains per Word Read in Table (6.3) are slightly different from those reported in the original paper because these values are based on using the more sophisticated Python NLTK word tokenizer whereas in the original paper the values reported in this particular table were computed using simple whitespace separation.

6.8 Results - Time spent and Image coverage

In this section, we highlight some other interesting findings from the study that involved more nuanced analysis. Firstly, observe that though we have continuously considered effort to be defined as a measure of word count (keyword count in optimization problem 6.1 and overall word count otherwise), there are other ways of defining effort. One such measure is time spent. It should be noted that in both experiments, (Syed and Collins-Thompson, 2017a) and (Syed and Collins-Thompson, 2017b), we did not enforce any explicit time constraints nor did we tell participants anything about how much time they should or could spend. The four-minute minimum time quality control mentioned earlier was enforced in post-experiment analysis. As such, participants spent as much time as they chose, without any likely bias.

It is thus interesting to observe that though participants in the α_P condition got less than a third of the total word count as the Web condition participants, the total average time that participants spent in any of the three retrieval conditions showed no significant differences, suggesting either that participants in the α_P condition were willing to spend more time reading due to the lower content length or that participants in the Web condition were speed-reading and skipping over chunks of text. Either way, because the time spent was not significantly different across conditions, it was expected that Learning Gains per Time Spent should also not be significant and we found roughly the same trends of significance for Learning Gains per Time Spent. However, if we broke down Learning Gains per Time Spent $\frac{LG}{Time}$ by the following decomposition, we found an interesting result:

$$\frac{LG}{Time} = \frac{LG}{WordsTotal} \times \frac{WordsTotal}{Time} = \frac{LG}{WordsTotal} \Big/ \frac{Time}{WordsTotal}$$

The relationship of these two subfactors is visualized in Figure 6.5, with $\frac{Time}{WordsTotal}$ on

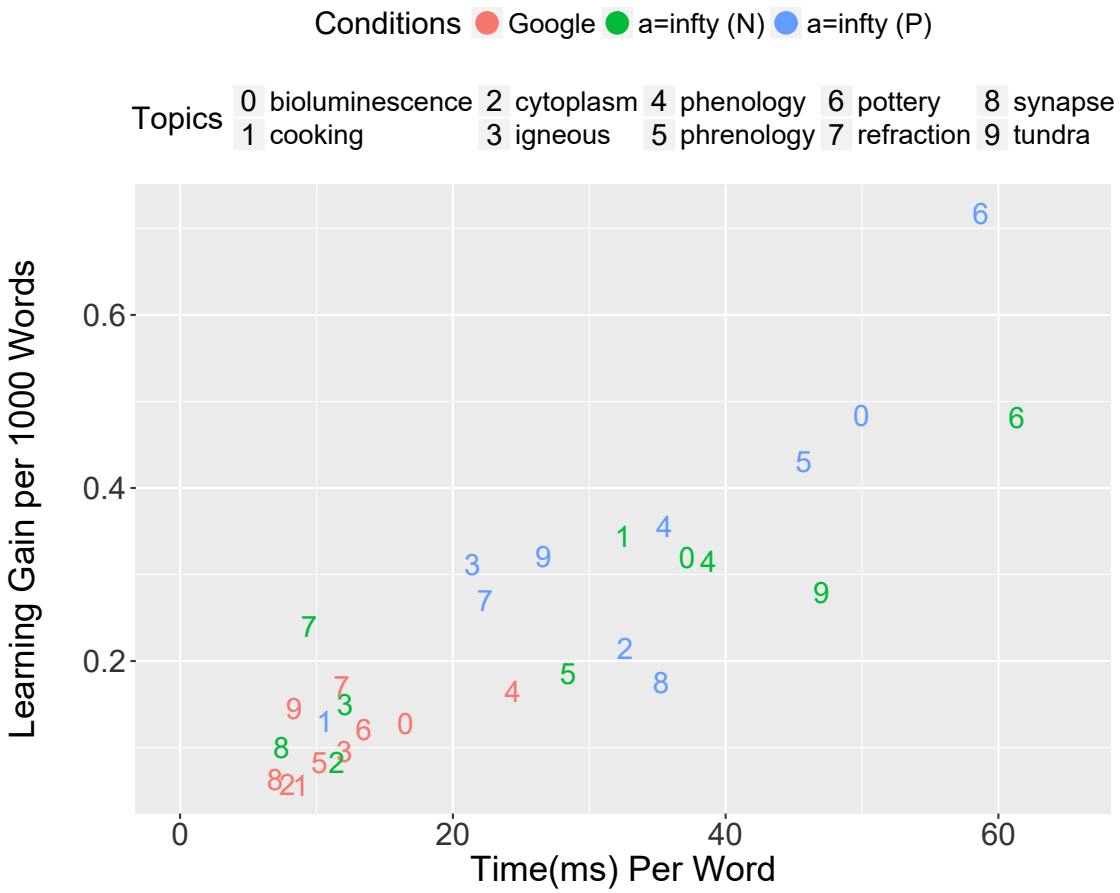

Figure 6.5: Learning gains per word generally increases with reading time per word. $\alpha = \infty$ (N) is the non-personalized condition and $\alpha = \infty$ (P) is the personalized condition.

the x-axis and $\frac{LG}{WordsTotal}$ on the y-axis. As the plot makes evident, there is a positive correlation (r=.374, p<.001) between these two subfactors. Moreover, while the slope of this approximately linear relationship (which is exactly $\frac{LG}{Time}$, learning per unit time), is relatively stable across conditions – as the initial analysis showed – there are in fact very different tradeoff regimes for user efficiency that lead to similar learning gains per unit time, across the three retrieval conditions. For example, the Web baseline is largely characterized by having the lowest average reading time per word as well as lowest learning gain per word

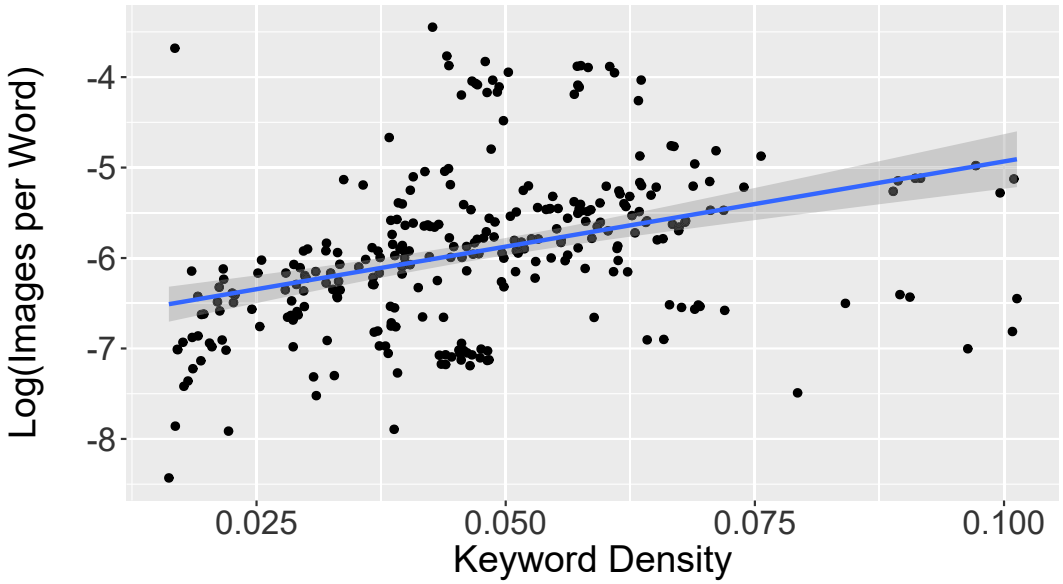

Figure 6.6: Image coverage increases with keyword density. Each data point represents a unique document set shown to a study participant.

(7/10 topics). In contrast, the personalized α_P condition is characterized by typically having the highest learning gain per 1000 words (8/10 topics).

Image Coverage. To gain more insight into why pages with increased keyword density might contribute to more efficient and effective learning, we investigated additional properties of the page content that might be correlated with keyword density. We found that while few result documents made use of multimedia such as animations, audio or video, some used images to supplement the text. Thus, the *picture superiority effect* (De Angeli et al., 2005), in which people tend to remember things better when they see pictures rather than words, could be relevant. We examined whether there was a relationship between image coverage – defined as total images divided by total words – and keyword density. We determined the number of relevant images manually for each page, excluding irrelevant images such as navigation icons and advertisements. We found that pages with higher keyword density did

indeed tend to have increased image coverage. On average, participants saw 1.5, 4.8 and 3.9 images per 1000 words, for the Web, α_N, and α_P conditions respectively. Thus, participants in either of the two $\alpha = \infty$ conditions saw almost three times as many images per word as those in the Web commercial search baseline. We observe informally that pages using a higher density of keywords tend to be those that give an overview of topic for instructional purposes, and thus are more likely to be supplemented with images by the author. The keyword density of the document set a participant read did indeed show a linear relationship to the log of the image coverage. Fig. 6.6 shows a linear correlation between these measures ($r=.37$, $p<.001$)[1]. It is for future work to investigate this phenomenon and how other content properties may interact with learning outcomes in search.

6.9 Results - Effect of Time Spent on Differences in Learning Gains

Thus far, we have found that the personalized condition typically yielded stronger learning gains, both in terms of absolute gains and in terms of realized potential gains. However, as we saw earlier, the magnitude of these improvements were not very strong (learning gains had an average improvement of about 15% in the personalized condition over the baseline condition and realized gains had an average improvement of about 22%). However, in this section we consider the possibility that maybe this is in part because some participants were simply not motivated and were just skimming through the task as fast as possible without really trying to learn. To test this hypothesis, we performed a median split on the full dataset ($n=432$) on the total time spent in the reading portion of the task. We repeated the analysis

[1]Documents with no images were omitted from the log calculation.

of learning gains and realized potential gains on both the set where participants spent less than median time and greater than median time.

Consistent with our hypothesis, we found that in the dataset where participants spent less than median time, the average improvement in learning gains relative to the baseline condition sharply dropped from 15% to 0.6% (insignificant differences in means across the three conditions). Conversely, when considering those who spent greater than median time, we found the average improvement *increased* sharply from 15% to 28% (p=.004). This suggests that for those who actively spend more time engaged in a learning task, reading documents that have higher keyword density can lead to even stronger learning gains.

6.10 Limitations

Our implementation of the framework makes several important assumptions. Firstly, we model user knowledge state as being binary (either the user does or does not know the meaning of the keyword). We also make the assumption that all of the keywords have single meanings - that is, we assume there is no polysemy in the keywords being taught. In future work we could explore the use of context to account for different word senses. Furthermore, our model makes the assumption that a document's readability is only a function of weighted keyword density. However, prior work has reliably found that a student's ability to learn in a reading setting is limited unless they have knowledge of at least 85% of the terms used in the content (Paul, 2003; Topping and Sanders, 2000). As such, future work should factor in not only the user's prior knowledge of the keywords but also their expected knowledge of *all* words used in a given document. We leave it to future work to investigate how, if at all, results change when these assumptions don't hold.

Chapter 7

Long-term Learning from a Web Search Retrieval Framework (Study 2b)

Thus far, we have considered the impact on learning outcomes from using three different retrieval models. However, in all the studies discussed so far, the focus has been on evaluating the short-term, immediate change in knowledge state. However, robust - or long-term - learning outcomes are arguably a more valuable measure to investigate as this may tell us how well a particular retrieval algorithm, measure of robust learning, or classification of keyword difficulty are characteristically different in the long-term. We now describe a crowdsourced longitudinal study (Syed and Collins-Thompson, 2018) of *long-term* retention (or robust learning), in which a subset of users who participated in the initial learning and assessment study (described above (Syed and Collins-Thompson, 2017b)) also completed a delayed post-test nine months later.

7.1 Study design

Our experiment used the same platform, Crowdflower, as the study by Syed and Collins-Thompson (2017b), as well as the original crowd response dataset used in the above analysis (Chapter 6). We altered the task design to include three pages of multiple-choice question tests for three topics out of the ten total that were originally tested. Afterwards, participants

completed a Likert-scale survey of the perceived importance of various "learning factors" (Abualsaud, 2017) on learning outcomes.

We limited this delayed post-reading assessment to only three topics to prevent participants from having to take too many tests and possibly having tiredness contaminate the results. We still added explicit quality control measures by adding gold standard test questions in each of the three tests that participants had to pass and we randomized the order in which the assessments appeared. Unfortunately, while the Crowdflower platform allows us to see the unique worker's ids after an experiment has terminated, they do not allow us to have this information during the experiment, nor do they allow us to specifically request certain workers. As such, we had to rely on chance that we would get repeat participants and further on chance that some of those repeat participants would have participated in one of the three selected topics. To maximize the number of data points we could get, we chose the three topics which had the lowest number of unique participants.

We accumulated a total of 600 judgments from unique crowd participants and of these, 36 were unique repeat participants (out of a maximum of 116 from the set of three topics we chose) and there were 83 unique (participant, topic) tuples that matched the original dataset. After filtering out those who did not answer all the gold standard test questions correctly, we ended up with 81 unique tuples. We perform the subsequent analysis on this dataset matched against the original dataset. For notation purposes, we consider "pre-test" to be the pre-reading test results from the original study, "post-test" to be the post-reading test results from that same study and "delayed-test" to be the test results from the current crowdsourced study.

We consider the following measures of robust, or long-term, learning outcomes: (1) robust learning gains; (2) robust retention of learning gains; (3) robust retention of post-test

knowledge and (4) robust change in post-test knowledge. We define these measures as follows:

1. **Robust Learning Gains.** Computed as the sum of keywords that a participant did not know in the pre-test and did know in the delayed-test.

2. **Retained Gains.** Computed as the sum of keywords a participant learned (as defined by Learning Gains) and that they still knew in the delayed-test.

3. **Retained Knowledge.** Computed as the sum of keywords that a participant did get correct in the post-test and still got correct in the delayed-test.

4. **Net Retained Knowledge.** Computed as signed sum of retentions in post-test knowledge (retention is positive if participant got the keyword correct in post-test and again in delayed-test; retention is negative if participant got the keyword correct in post-test and wrong in delayed-test).

7.2 Variation by Keyword Difficulty

We first analyze how the average robust measures compare when considering the averages of the lowest-difficulty keywords only versus the averages of the highest-difficulty keywords only. We split the set of ten keywords into sets of five by a median split on their Age-of-Acquisition scores (Kuperman et al., 2012). We then compute each of the robust measures as well as the pre-test scores on each of the sets and perform a Kruskal-Wallis test to test for significance. The results are shown in Table 7.1.

We find that of the four robust measures, Retained Gains and Net Retained Knowledge showed significant differences in means: (lower average = 0.457, upper average = 0.765,

Difficulty Split	**Lower Difficulty**	**Higher Difficulty**	**p-val**
Robust Gains (Long-term)	1.025	1.000	0.867
Retained Gains	0.457	0.765	0.002
Retained Knowledge	2.395	2.296	0.733
Net Retained Knowledge	1.815	1.160	0.067
Learning Prior	2.753	2.469	0.093
Learning Gains (Short-term)	0.679	1.296	$<.001$

Table 7.1: Averages for the two splits for each robust measure along with two short-term measures indicates better opportunity for gains in difficult terms.

p=.002) and (lower average = 1.815, upper average = 1.160, p=.067)[1] respectively. This suggests that in general, of the keywords participants were able to learn and remember, more of these were likely to be difficult ones. On the other hand, the opposite trend with Net Retained Knowledge suggests that overall participants were also more likely to *forget* the meanings of more difficult keywords. What does this mean?

Recall that Net Retained Knowledge expands the calculation of Retained Knowledge which itself expands the calculation of Retained Gains. We know from Table 7.1 that Retained Knowledge showed no significant differences, suggesting that the disparity must be driven from the negative sum in Net Retained Knowledge. This shows an interesting balance where participants who retained short-term learning gains tended to retain acquired knowledge of more difficult terms better. However, in cases where they forgot newly-learned terms, they tended to lose acquired knowledge more with difficult terms as well. In aggregate, there appears to be more forgetting than retaining with difficult terms, suggesting that participants with better post-test knowledge of easier terms will likely show a better net retention of that knowledge even after a considerable time delay.

Another interesting finding is that the Robust Gains split was unaffected by difficulty

[1]This significance was strengthened to $p<.05$ when normalizing by post-test knowledge

but the short-term learning gains were strongly improved by higher difficulty (almost twice as much). It is also interesting to observe that the averages of these measures suggest a negative relationship (i.e. lower short-term gains in easier terms led to better long-term gains of easier terms and vice versa for difficult terms). This may be explained by the fact that more difficult keywords are likely those that are more unfamiliar and novel to the learner and this novelty may facilitate better immediate recall but not long-term retention. Conversely, learning unknown but easier keywords may be less likely to cause learning gains as just a function of recall.

From the concept of *desirable difficulties* (Bjork, 994a), it is possible that the easier keywords that were unknown to the participant were those that were sufficiently difficult to learn but not so much that they inhibited long-term retention. This is further supported by the results of Net Retained Knowledge, suggesting that easier keywords showed substantially better net change in delayed-test knowledge. These results suggest that in personalizing document selection it is important to not just consider which words are known or unknown, as was done in (Syed and Collins-Thompson, 2017b), but also incorporate the difficulty of the known terms and choose documents with more unknown terms that are in an estimated zone of proximal development (Wertsch, 1984).

Measure	**Model 1**	**Model 2**	**Model 3**	**p-val**
Robust Gains	1.960	2.000	2.136	0.809
Retained Gains	1.280	1.059	1.409	0.856
Retained Knowledge	4.440	4.706	4.955	0.706
Net Retained Knowledge	2.520	2.941	3.545	0.439
Post-Test	6.360	6.471	6.364	0.966
Delayed-Test	5.560	6.118	6.091	0.764

Table 7.2: Averages of the median difficulty split applied to short and long-term knowledge states, broken down by retrieval models.

7.3 Variation by Retrieval algorithm

We now analyze whether there were differences in robust learning outcomes depending on the search condition a user was assigned in the original study. Recall that there were three possible conditions: (1) commercial search engine Web (Model 1); (2) non-personalized retrieval α_N (Model 2) and (3) personalized retrieval α_P (Model 3). In our long-term dataset, each condition had roughly similar, but small, sample sizes (n=25, n=34, n=22) respectively. The Model 2 and Model 3 algorithms exclusively considered a measure of difficulty-weighted keyword density as the document selection criteria, with Model 3 also incorporating information about the participants' prior knowledge and Model 2 assuming zero prior knowledge for all participants. Details on these algorithms were discussed in the preceding section (Chapter 6).

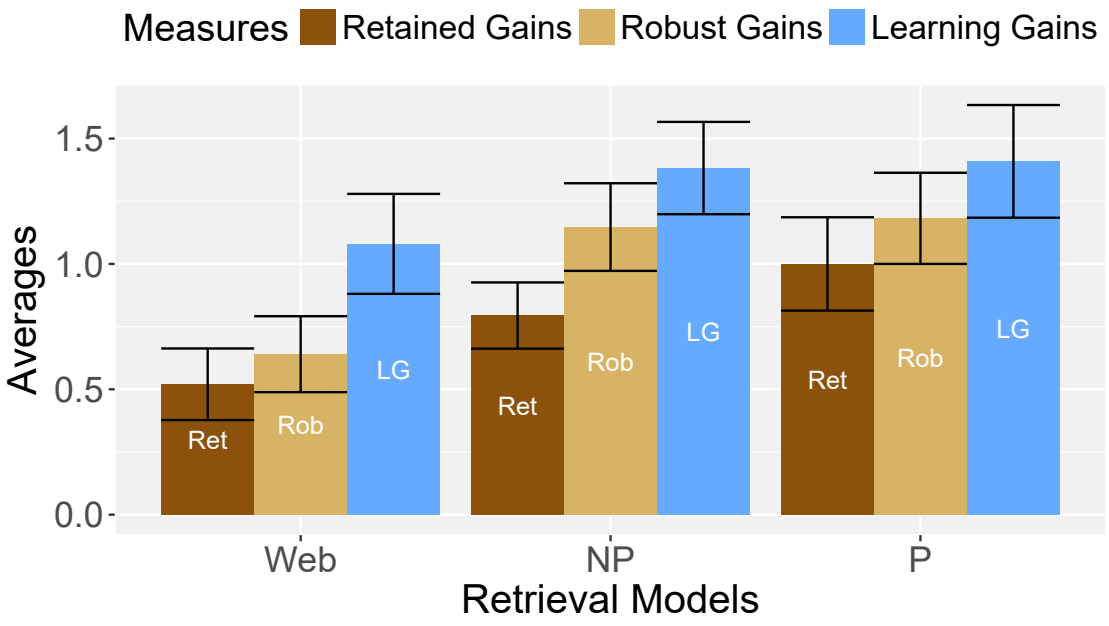

Figure 7.1: Average changes in knowledge state over three periods of assessment, for each retrieval model.

We found that omnibus Kruskal-Wallis tests between these three models showed no significant differences for each of the four robust measures (Table 7.2), suggesting that in aggregate the choice of retrieval model didn't have significant impact on robust learning outcomes. However, if we split these features again by difficulty, we find some significant differences. In particular, both Robust Gains and Retained Gains showed significant differences ($p<.05$) when comparing only Model 1 and Model 3 on higher difficulty keywords. In both cases, Model 3 outperformed Model 1 (by 85% and 92% respectively), suggesting that the personalized algorithm introduced in (Syed and Collins-Thompson, 2017b) produced significantly better long-term improvements in knowledge of more difficult terms, including better retention of short-term gains on such terms.

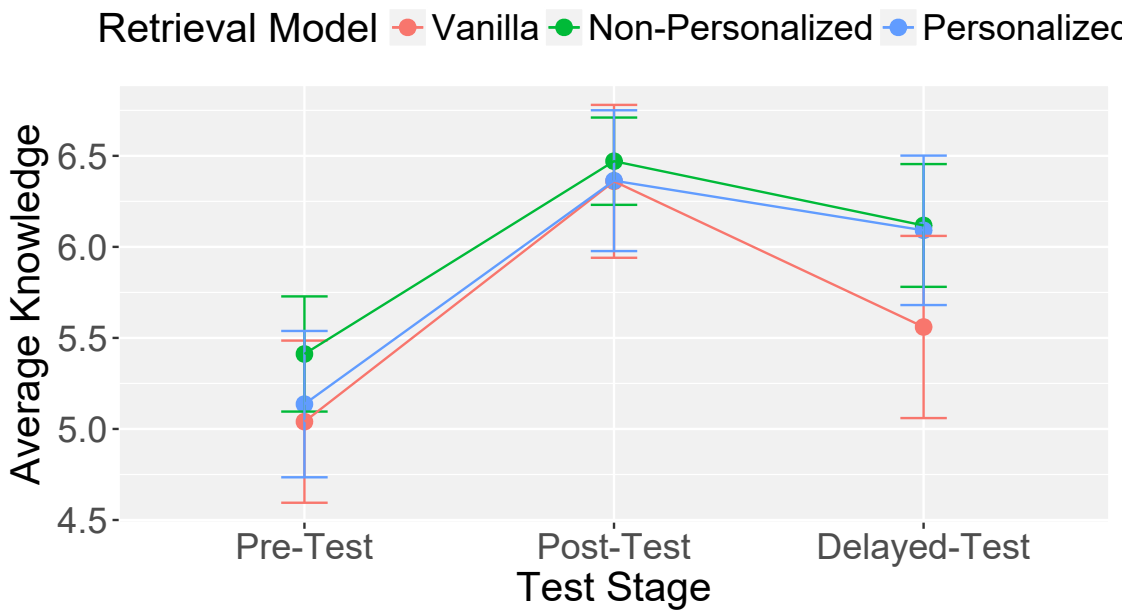

Figure 7.2: Average changes in knowledge state over three periods of assessment, for each retrieval model.

We also observe some interesting variations in measures of final knowledge state. In particular, observe in Table 7.2 that the post-test final knowledge state showed very small

differences across each of the models, suggesting that regardless of the retrieval model, the final knowledge state mostly ended up the same. However, in the delayed-test knowledge state, while there was consistent evidence of forgetting, this effect was distinctly stronger in Model 1, which was the commercial search baseline (Figure 7.2). This suggests that the other two models, proposed in (Syed and Collins-Thompson, 2017b) actually did demonstrate not just evidence of short-term improvements but very possibly evidence of long-term improvement as well.

Overall, we find that the personalized document retrieval model (Model 3) showed substantially better ability compared to a commercial Web search model (Model 1) to help participants achieve long-term understanding of more difficult keywords and retain short-term learning gains of such keywords as well. We further find that, though not significant, the commercial model produced relatively stronger overall forgetting from post-test to delayed-test.

Chapter 8

Predicting Learning Outcomes through Data-Driven Analysis (Study 2c)

In this chapter, we build on the results from the study in Chapter 6 and analyze what document features, at the individual document level, at the document set level and at the user interaction level, affected various types of learning outcomes. While the results from the earlier study demonstrated an improvement in terms of retrieval models, the study did not explore what variables of the documents themselves, besides keyword density, could be causing improvements and what factors should subsequently be encouraged for better educational Web page design.

Type	**Group**	**Feature**	**Description**
D	Effort	$WordCount$	Total number of unigrams in the document.
D	Effort	$KeyCount$	Total number of keywords in the document.
D	Effort	$DocumentCount$	Total number of documents in the set. This feature ranges from 1 to 10.
D	Effort	$WordsPerDocument$	Ratio of $WordCount$ to $DocumentCount$.
D	Effort	$DocumentAgeDifficulty$	85^{th} percentile Age-of-Acquisition score for the document. Uses the expanded set of scores from the study by Kuperman et al. (Kuperman et al., 2012).
D	Effort	$WeightedWordCount$	Each unigram is assigned its corresponding "age" from the Age-of-Acquisition dataset. These scores, for each occurrence of each unigram in the document, are added up.

D	Effort	*AverageParaLength*	Average length of each paragraph in the document. Computed as count of all unigrams in all HTML <p> tags divided by total instances of <p> tags.
D	Images	*ImageCountTag*	Total instances of the HTML <img> tag that appeared in the document. More images
D	Images	*ImageCountManual*	Total instances of non-advertising and non-navigational images that appeared in the document. Counted manually.
D	Images	*ImageToText*	Ratio of *ImageCountTag* to *WordCount*.
D	Links	*OutboundLinks*	The count of all outbound links.
D	Keywords	*KeywordDensity*	Computed as the count of occurrences of any of the N keywords $k_1, \ldots, k_N$ divided by the count of all words (i.e. *WordCount*).
D	Keywords	*WeightedDensity*	Same as *KeywordDensity* except the denominator is the *WeightedWordCount* feature.
U+D	Keywords	*IncorrectKeysRatio*	Total occurrences of keywords that the participant got wrong in their pre-test, divided by the total occurrences of any keyword in that document.
U+D	Keywords	*IncorrectSemanticRatio*	The *SemanticRelevance* score is computed as follows: first compute the relevance of each keyword in a document by computing the average Word2Vec similarity (Mikolov et al., 2013) of its five surrounding words (both ahead and behind). *IncorrectSemanticRatio* is the sum of all *SemanticRelevance* scores for keywords the participant got wrong on the pre-test, divided by the total sum of *SemanticRelevance* scores.
DS	Keywords	*LogWeightedDensity*	Same as *WeightedDensity* except that instead of simply summing the values over the set of documents, each successive document's value of *WeightedDensity* was reduced by a DCG discount factor of $\log_2(p+1)$ where p is the rank in the set of documents.
DS	Images	*Set_ImageToText*	Set-level calculation of *ImageToText*.
DS	Effort	*Set_AvgParaLength*	Set-level calculation of *AverageParaLength*.
DS	Keywords	*Set_KeyDensity*	Set-level calculation of *KeywordDensity*.
DS	Keywords	*Set_WeightDensity*	Set-level calculation of *WeightDensity*.
U+DS	Keywords	*Set_IncorrectRatio*	Set-level calculation of *IncorrectKeysRatio*.

U+DS	Keywords	$Set\_IncorrectSemsRatio$	Set-level calculation of $IncorrectSemanticRatio$.
U+DS	Keywords	$ExpectedKnowledge$	Expected knowledge computed as a personalized sigmoid function of keywords Syed and Collins-Thompson (2017b).
U		$PriorKnowledge$	Sum of initial correct answers to the vocabulary terms needed to be learned.

Table 8.1: Set of features that were considered. "U" are User features: those that involved prior data about the User's knowledge. "D" are Document features: required only individual document's raw data. "DS" are Document Set features: treated the set of documents as a single bag-of-words. In computing features in this dataset, their values were aggregated (by summation), since learning outcomes were measured against sets of documents.

8.1 Choice of Features

Overall, we considered a set of document features as candidate features for regression models that included features pertaining to image use, vocabulary difficulty, word count and content structure. A complete list, including user-dependent features, can be found in Table 8.1. We chose document and user features based on various concepts investigated in earlier studies. Broadly, the features we chose can be grouped as follows:

1. **Image content.** Some studies have found that providing plain-text filtered documents improves learning outcomes (Freund et al., 2016) over the original document, possibly suggesting a negative effect of image coverage and learning. However, other studies found positive association of image coverage and learning outcomes, when used appropriately (Mayer, 1997) and a positive association with the fraction of images in documents and the ability of users to find relevant content (Verma et al., 2016)

2. **Keywords content.** Prior work has found that optimizing document selection by difficulty-weighted keyword density improved multiple measures of learning outcomes

(Syed and Collins-Thompson, 2017b). We also investigate other keyword features like the coverage of keywords unknown to the user relative to all keywords.

3. **Effort.** Prior work has suggested that too much effort on the part of users can be overwhelming and, by Cognitive Load Theory, could hurt learning outcomes (DeStefano and LeFevre, 2007). On the other hand, having "desirable difficulties" has been found to improve learning outcomes. We consider effort as functions of document count, word count and reading-difficulty-weighted measures of content.

4. **Embedded links.** Several studies have found that embedded links in documents can disturb the linearity of the learning process (Zumbach and Mohraz, 2008) and can add extra cognitive load (DeStefano and LeFevre, 2007).

8.2 Measures of Learning Outcomes

Before fitting any of these features to models, we first determined a set of learning outcome measures of interest. In this section, we consider the following measures of learning outcomes, computed on the provided sets of $K = 10$ vocabulary questions, with Pre_k as prior knowledge of keyword k, $Post_k$ as corresponding post knowledge and r_k as vocabulary difficulty level of k:

1. **Learning Gains (LG).** As a simple measure of learning growth we compute the total instances where a participant did not know a keyword to be learned in the pre-reading test and did know the definition in the post-reading test.

$$LG = \sum_{k=1}^{K} \begin{Bmatrix} 1 & Pre_k = 0 \text{ and } Post_k{=}1 \\ 0 & \text{otherwise} \end{Bmatrix}$$

2. **Difficulty-Weighted Gains (DWG).** This meausure is essentially the same as Learning Gains but we weight the learning gains of each keyword by the vocabulary difficulty level associated with it. These difficulty scores are retrieved from the expanded dataset from work by Kuperman et al. (Kuperman et al., 2012). By weighting the learning gains by vocabulary difficulty, we can capture the intuition that learning more difficult words like 'luciferase' and 'eclogite' may require different features than those required for learning easier words like 'minerals' or 'soils'.

$$DWG = \sum_{k=1}^{K} r_k \begin{cases} 1 & Pre_k = 0 \text{ and } Post_k{=}1 \\ 0 & \text{otherwise} \end{cases}$$

3. **Realized Potential Gains (PG).** This is a measure of how much Learning Gain the participant got relative to how much they could have possibly gotten. Specifically, for a set of 10 vocabulary terms being tested, we have:

$$PG = \frac{LG}{10 - \sum_{k=1}^{10} Pre_k}$$

Participants who had perfect prior knowledge (10/10) were omitted from analysis as they could not have theoretically shown any improvement.

4. **Final Knowledge (FK).** This is a much simpler measure of learning outcome where we take the linear sum of the participant's final test scores, regardless of their prior performance. Specifically, we have:

$$FK = \sum_{k=1}^{K} Post_k$$

5. **Learning Hindrance (LH).** While previous measures of learning outcomes assessed positive learning outcomes, it is also important to understand features that may *hinder* learning. We consider Learning Hindrance to be the total keywords that a participant got wrong in the pre-test and got wrong again on the post-test, indicating that they were unable to learn the definition. Specifically, we have:

$$LH = \sum_{k=1}^{K} \begin{Bmatrix} 1 & Pre_k = 0 \text{ and } Post_k{=}0 \\ 0 & \text{otherwise} \end{Bmatrix}$$

6. **Total Reading Time (TR).** While this is not technically a measure of learning outcomes, it is an important measure to analyze as it can help determine what document and user features influence how much or how little time people are willing to spend when engaged in a learning task. This is measured as the total time (ms) a user spent reading the set of documents they were provided.

8.3 Analysis

We conducted our analysis on the personalized subset of participant records from the earlier study (Syed and Collins-Thompson, 2017b) that we discussed in Chapter 6. Following the quality controls filters we used in that study, we ended up with (n=283) records of personalized document sets per user, allowing us to analyze what properties of different collections of documents led to various learning outcomes.

8.4 Prediction without User Data

There are many scenarios in which for Web search it may be difficult or impossible to obtain an accurate assessment of a user's prior knowledge, especially for any arbitrary topic. Thus, here we investigate document features that are completely independent of the user ("D" and "DS" type properties only) and assess how well robust regression models trained on these features can predict learning outcomes. These models could facilitate learning-oriented retrieval for situations where a Web search framework has access to document data but not to a user's prior knowledge. We tabulate the trained models and cross-validated correlations in Table 8.2.

In selecting the features for each model, we considered two approaches: exhaustive search or stepwise algorithm using AIC criterion. While exhaustive search naturally produced the best model in the training phase, we found that the stepwise selection through the AIC criterion produced a lower average residual standard error (RSE) in 10-fold cross validation. For consistency, we used the stepwise AIC method for feature selection for all models discussed in this work. We also scaled both the predictors and the dependent variables in all models to the range $[0, 1]$. To avoid influential points affecting the model, we fit all the models with robust regression.

The results from Table 8.2 show that even without any features about the user, we can still get reasonably strong correlations between predicted learning outcomes and actual outcomes. For learning gains, the Difficulty-Weighted Gains tend to show substantially better improvement over the unweighted gains. On the other hand, the Final Knowledge state variable shows a much stronger correlation as does the Learning Hindrance variable. For a better understanding of what drives these correlations, we visualize the trained models

for Difficulty-Weighted Gains, Final Knowledge and Learning Hindrance in Figure 8.1.

For the selected features, all positive measures of learning showed positive weights for ImageCountManual and negative weights for ImageCountTag, suggesting that, in general, Web pages having more relevant images tend to help actual learning outcomes but irrelevant images (such as ads and navigational icons) may hurt the learning process, possibly by being distractive to the user. This is in accord with existing work in this area that has suggested that having images in learning material has been found to both help and harm learning outcomes, depending on the study(DeStefano and LeFevre, 2007). Note also that all measures of learning gains showed a negative relationship with the total number of links in the document, which is consistent again with what we would have expected from theory (Chapter 8.1). However, it is not entirely clear why Final Knowledge shows a positive relationship with total links. We also observe that both unweighted and weighted learning gains measures were positively affected by weighted keyword density, at the individual document level and negatively at the set-level. This is partially consistent with the results from (Syed and Collins-Thompson, 2017b) that found that document sets produced by optimizing for document-level weighted keyword density outperformed commercial baseline results in terms of learning gains. The disparity may be due to the document-level features being computed as sums across all documents in the set, thus making the DocumentCount feature an implicit feature in document-level weighted density. However, while DocumentCount was not a significant predictor for most models, it was significant for LearningGains where it had positive weight, suggesting that more documents actually helped, further suggesting that the weighted keyword density in general should be penalized rather than rewarded.

We find a similar tradeoff when it comes to average paragraph length. We note that at the set-level, the average paragraph length helped all measures of learning, as did total word

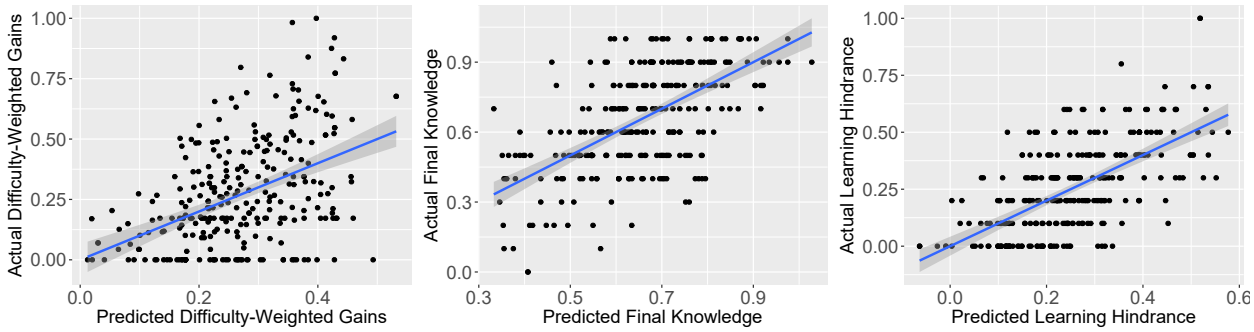

Figure 8.1: Expected and actual learning measures trained on non-user features.

count, possibly suggesting that more content and less segmentation of that content is beneficial to learners. However, we also note that the document-level average paragraph length showed the opposite trend for Potential Gains and Final Knowledge, possibly suggesting that average paragraph lengths should be longer but there should be fewer documents overall.

8.5 Predicting with User Data

We have so far seen that in the absence of any user-dependent features, we were able to train robust regression models on multiple measures of learning, resulting in observed trends that were commensurate with findings from existing literature. Now, we attempt to further augment the power of these results by modeling all the features from Table 8.1. Repeating the same feature selection and model fitting process as in the earlier section, we get the results in Table 8.3.

The first point that we note is that the cross-validated correlation for all measures of learning were improved, some quite substantially. This much was to be expected as we are adding new information signals to the model, signals which have a naturally strong correlation to most measures of learning already. For example, regardless of other properties, the

Feature	**LG**	**DWG**	**PG**	**FK**	**LH**	**TR**
WordCount		0.4379	3.6121		-0.5535	-2.8926
WeightedWordCount			-3.5873			2.3241
AverageParaLength			-0.2336	-0.2755	0.3486	
ImageCountManual	0.2904	0.3224	0.5441	0.2996	-0.2738	0.1544
OutboundLinks	-0.2394	-0.3990		0.2681	-0.1498	0.3157
KeywordDensity	-2.4830	-1.9237	-1.9809			
WeightedDensity	1.7599	1.8847	2.1748			
DocumentAgeDifficulty		0.3747	-0.2834	-0.2651	0.3308	
ImageToText						
ImageCountTag	-0.3068	-0.2283	-0.6259	-0.2909	0.2688	0.1498
KeyCount		-0.4071				
LogWeightedDensity	0.5371					
DocumentCount	0.4221					0.0864
WordsPerDoc					0.4481	
Set_AvgParaLength	0.1492	0.1832	0.2393	0.1142	-0.1181	0.2591
Set_ImageToText				-0.2189	0.1909	-0.2600
Set_KeyDensity	1.5808	2.0079	1.6829	-0.3255	0.2626	
Set_WeightDensity	-1.5624	-1.8801	-2.1677			
Performance	**LG**	**DWG**	**PG**	**FK**	**LH**	**TR**
Correlation (model prediction vs actual)	0.3296	0.3611	0.3436	0.5810	0.6117	0.2376

Table 8.2: Trained normalized features for different dependent variables. Values for corresponding features are learned coefficients in the robust regression model. LG = Learning Gains; DWG = Difficulty-Weighted Gains; PG = Potential Gains; FK = Final Knowledge; LH = Learning Hindrance; TR = Total Reading Time (ms).

user's prior knowledge could be expected to have a strong negative correlation with Learning Gains since users with higher prior knowledge naturally have less opportunities for improvement. Indeed, we trained the set of six learning measures against a robust model containing *only* PriorKnowledge as a predictor and found substantially strong correlations from that alone (last row of Table 8.3). However, training against the full set of features did show significant improvement in predicting Learning Gains, Difficulty-Weighted Gains, Potential Gains and especially Total Reading, which had almost no correlation with PriorKnowledge. As such, there are definitely advantages to incorporating both user features and document features for better results.

We note that we again see similar trends that we saw before: (1) all measures of learning outcomes had positive coefficients for the count of relevant images and those measures that had count of all images as a significant feature had negative weights as we also saw earlier; (2) weighted keyword density again shows a conflicting association with learning outcomes at the set-level and the sum of document-level; (3) we see a similar effect that we discussed earlier with average paragraph lengths as well as with total embedded links. However, we also notice some new effects and features that we didn't see earlier.

Firstly, note that the ImageToText ratio feature was in the original models as well but was not significant for most of the features. However, in this set of features, the set-level ImageToText feature is significant for *all* measures of learning and is consistently negative, suggesting that in general, while more images might be helpful, there needs to be an overall balance between how many images there are per unit of text. Secondly, we note the somewhat intuitive finding that the ratio of counts of unknown keywords to all keywords is a positive predictor of better learning outcomes at the document-level. However, it shows the opposite trend at the set-level, either suggesting that in aggregate a set of documents should *not* have

Feature	**LG**	**DWG**	**PG**	**FK**	**LH**	**TR**
WordCount						-2.5116
WeightedWordCount						1.8478
AverageParaLength	-0.1523				0.1066	
ImageCountManual	0.3077	0.3867	0.5178	0.2353	-0.2154	
OutboundLinks						
IncorrectSemanticRatio			0.7476			0.6536
KeywordDensity	-0.4643	-0.5915	-2.2101	-0.5441	0.3250	-0.2334
WeightedDensity			2.2856			
DocumentAgeDifficulty			-0.4410			
ImageToText						
IncorrectKeyRatio	0.3443	0.3578		0.3933	-0.2410	-0.5565
ImageCountTag	-0.1759	-0.2824	-0.5097	-0.1426	0.1231	0.2191
KeyCount						0.3497
LogWeightedDensity	0.3261	0.4702		0.3570	-0.2283	
DocumentCount						0.2834
WordsPerDoc						
ExpectedKnowledge	-0.1199			-0.1757	0.0839	-0.2341
Set_AvgParaLength	0.1466			0.1098	-0.1026	0.2404
Set_ImageToText	-0.2745	-0.1738	-0.3182	-0.2347	0.1921	-0.1901
Set_KeyDensity		1.4125	1.3909			
Set_WeightDensity		-1.4657	-1.7781			
Set_IncorrectRatio	-0.6198	-0.2546	-0.4914	-0.6803	0.4338	
Set_IncorrectSemsRatio	0.4063			0.4612	-0.2844	
PriorKnowledge	-0.3694	-0.3889	0.3289	0.7584	-0.6414	0.3565
Performance	**LG**	**DWG**	**PG**	**FK**	**LH**	**TR**
Correlation (model prediction vs actual)	0.4571	0.5091	0.3908	0.7156	0.7499	0.2650
Robust correlation with PriorKnowledge	0.3744	0.3397	0.2731	0.6657	0.7361	-0.0563

Table 8.3: Trained normalized features for different dependent variables (considering *all* possible features). Values for corresponding features are learned coefficients in the robust regression model.

stronger coverage of unknown keywords (that need to be learned). It is not entirely clear why this is true.

In aggregate, this enhanced set of features has given us trained models that do show expected improvements over the document-features-only models and much of the same observations remain valid in these new models as well. While the results from the study by Syed and Collins-Thompson (Syed and Collins-Thompson, 2017b) demonstrated strong improvements in learning efficiency (learning gains per unit of effort), the models introduced so far can give us a way to produce strong improvements in learning effectiveness (learning gains or final knowledge state) or strong reductions in learning hindrance.

8.6 Discussion

An interesting finding in the trained models was the often-conflicting weights between features aggregated at the set level versus those computed at the document level and then summed up. For ratio features, like AverageParaLength, KeywordDensity and WeightedDensity, the sign of their coefficients were almost always opposite when considering document-level vs set-level aggregation. This is an unexpected finding which conflates the interpretation of whether or not these features are good or bad for learning outcomes. It appears that document sets with more documents would be more affected by document-level features as these features are summations whereas the set-level features are ratios taken over the entire set. It is for future work to further tease out the effects of document- vs set-level features.

The choice of features we used in this study were chosen to be such that they could be reliably and efficiently reproduced at scale. All features were extracted algorithmically and efficiently with the exception being the ImageCountManual feature which required manual

effort. As such, re-training these models, excluding ImageCountManual, could be done with minimal human or computational effort in real-time, making this suitable for large-scale applications.

8.7 Conclusion

In this study, we performed deeper analysis into the causes of what user interaction variables and document properties, at the individual and set level, affect learning outcomes in Web search. We trained several regression models and demonstrating strong cross-validated ability for these models to predict such learning outcomes. We also demonstrated the ability for these models to perform very strong even in the absence of any user data and relying exclusively on document properties. The simple regression models allow for very easy and scalable integration into existing learning-to-rank frameworks to facilitate further work in search as learning. It should be noted that the study in this section and those in the previous two sections have focused on understanding and developing models for optimizing document selection for learning intents in search. Despite the fact that many search queries have been found in prior work to be of an educational or learning intent, this is certainly not always the case and we leave it for future work to investigate methods of *detecting* such queries (e.g. (Yu et al., 2018)) whereas here we focus on *improving* results for such queries.

Chapter 9

Towards Generalizable Models of Learning Gains (Study 3)

In the previous chapter, we demonstrated a strong ability to predict learning outcomes using only document features. While the results from the study indicated promising results for vocabulary learning on the open Web, there were a few limitations that needed to be addressed. In this study, we identify the areas that need to be investigated to provide a compelling argument in favor of generalizability.

1. **Type of learning**. The nature of the task in most of the prior chapters was focused on vocabulary learning which could be considered as the simplest type of learning as per the Bloom's taxonomy (Anderson, Krathwohl, Airiasian, Cruikshank, Mayer, Pintrich, et al., 2001). In this chapter, we consider other studies that focus on other types of learning as well.

2. **Replication study on the crowd**. Although we had promising results in our study, it was, to our knowledge, the first-ever crowdsourced study of how people learn in response to reading Web documents. As such, there is a need to understand how well our results could be replicated, especially when different types of questions are asked and different domains of interest are assessed.

3. **Replication study in the lab**. Even assuming our results can be replicated through

an independent crowdsourced study, it still raises the question of how well such results would generalize to the more controlled environment of a lab.

In this chapter, we describe how we addressed all of these fundamental questions. We further demonstrate that models trained on one dataset of search as learning was able to show strong generalization to two other datasets in the same space. In this process we evaluated generalizability along dimensions of learning type complexity, assessment platform, sample size and topic choice. To the best of our knowledge, this is the first study that investigates how a Web document model of learning generalizes to two independent studies in a similar space.

9.1 Datasets

In this study, we consider three independent datasets, all of which came from studies conducted within a span of two years from each other. The model training will be performed on **DS2**. Specifics of each dataset including sample size, topics assessed and platform type can be found in Table 9.1.

1. **DS1**. Participants were assessed on their learning gains for 10 topic-specific vocabulary words (Syed and Collins-Thompson, 2017b). They were initially given a pre-test, then provided a set of documents algorithmically selected for them, and finally a post-test consisting of the same questions. A total of 10 distinct topics were assessed with each task assessing exactly one topic.

2. **DS2**. This study also investigated learning from Web documents but gave users freedom in choosing which queries to enter and which documents to select (Yu et al.,

Dataset	Types of Learning	Assessment	Platform	Sample Size	Topics
DS1	Remember (Definitions)	Multiple-choice Questions	Crowd sourced	283	• Bioluminescence • Cooking • Cytoplasm • Igneous rock • Phenology • Phrenology • Pottery • Refraction • Synapse • Tundra
DS2	Remember (Facts)	Multiple-choice Questions	Crowd sourced	357	• Altitude Sickness • American Revolutionary War • Carpenter Bees • Evolution • HIV • NASA Interplanetary Missions • Orcas Island • Sangre de Cristo Mountains • Sun Tzu • Tornado • USS Cole Bombing
DS3	Remember Understand Apply Analyze Evaluate Create	Free Response	Lab	34	• Oil Spill • Open Data

Table 9.1: Comparison of multiple studies of learning in a Web document/search context.

2018). Similar to **DS1**, this study also used multiple-choice questions before and after the search session to assess changes in knowledge state. Unlike (Syed and Collins-Thompson, 2017b), this study focused on recall of facts rather than definitions.

3. **DS3**. This was a lab-based study that investigated what criteria influence how people choose documents in a learning task (Abualsaud, 2017). Unlike **DS1** and **DS2**, this study investigated all six dimensions of Bloom's taxonomy of cognitive complexities. The participants articulated their knowledge through free-response text forms and their responses were graded manually by two independent graders.

The third dataset we used, **DS3**, contained a total of 34 unique participant records and there were two distinct topics participants could study but they could not do both (Abualsaud, 2017). As a lab-based study, this study lasted between 60 and 90 minutes whereas the participants from **DS1** took an average of 3.55 minutes and those from **DS2** took an average of 13.35 minutes (Yu et al., 2018). Furthermore, in studies **DS1** and **DS2** the participants were only tasked with vocabulary or fact learning whereas in **DS3** the participants were taksed to learn in a more open-ended format and in more complex task types. In this study, we used our own dataset **DS1** as our training corpus and used the other two datasets as test corpora.

9.2 Preprocessing

Before training any models, we first performed some preprocessing. The full set of features we used is listed in Table 9.2. Most of these features are a subset of features used by Syed and Collins-Thompson in (Syed and Collins-Thompson, 2018). First, we had to compute the document features for each set of documents that users read through in all three datasets.

In this process, for **DS3** we determined there were a total of 135 unique links, 18 of which either returned 404 errors (page not found), could not be retrieved for other technical reasons or were not HTML documents. We excluded such pages from our analysis. In aggregate, since each participant used exactly 10 Web pages, there were 340 total links visited in the learning stage and of these, 35 could not be processed as explained above. For these instances of unusable documents in a set, we padded the features with their mean from the usable documents in the set. For **DS2**, we found a total of 279 unique links that had been clicked, after omitting links that could not be processed and those that were detected to be primarily non-English[1]. We further removed participant records where the total recorded time spent on documents was less than 1s and where the set-level features in Table 9.2 where infinite. All of these filters resulted in a total of 357 usable participant records. Finally, for **DS1**, we used the same data cleaning processes described in our earlier study (Syed and Collins-Thompson, 2018) which resulted in 283 usable participant records.

9.3 Measure of Learning Outcomes

The experiments that produced each dataset varied in implementation and in terms of how learning was quantified. For comparison across all three datasets, we needed a measure of learning that would consistently convey the same meaning. The simplest such measure was *percentage learning gain* (PLG) which we define as the difference between the sums of the pre- and post-task knowledge scores normalized by the max possible score. For **DS3**, because of how the pre- and post-test knowledge was assessed, we use the measure of *knowledge gain* as defined in (Abualsaud, 2017) also normalized similarly. As an example, an PLG of 0.35 would consistently mean the participant's knowledge state increased by 35% of the maximum

[1]Used Python's `langdetect` library for this

knowledge. Specifically, letting $MaxScore$ be the maximum possible assessed knowledge and having K test items, we have:

$$PLG = \frac{\sum_{k=1}^{K} Post_k - Pre_k}{MaxScore}$$

As **DS1** and **DS2** used multiple-choice questions, PLG was computed automatically whereas for **DS3**, two independent graders were used and the average of their graded pre- and post-test scores were calculated.

9.4 Model Fitting

In the training phase, we fit **DS1** to an L2-regularized linear regression model with PLG as the dependent variable. Prior to training all three datasets were independently standardized. Selection of the λ parameter (L2 penalty) was done via 10-fold cross-validation with the best value being $\lambda = 1.0$. The model weights are shown in Table 9.3. We evaluate the predictive power of this model on **DS2** and **DS3** in terms of the rank correlations between the model's predictions and the actual values in those datasets. For completeness we also include the results for **DS1**. These results are shown in Table 9.4.

9.5 Results

Our trained model showed significant predictive power in both the **DS2** and **DS3** datasets. The correlation in the **DS2** dataset were also aligned in the same direction as in the source **DS1** dataset. This was somewhat expected since **DS1** and **DS2** shared more in common than either dataset with **DS3** (the former two were both crowdsourced, both involved simpler

Type	Group	Feature	Description
D	Effort	$WordCount$	Total number of unigrams in the document.
D	Effort	$DocumentAgeDifficulty$	85^{th} percentile Age-of-Acquisition score for the document. Uses the expanded set of scores from the study by Kuperman et al. Kuperman et al. (2012).
D	Effort	$WeightedWordCount$	Each unigram is assigned its corresponding "age" from the Age-of-Acquisition dataset. These scores, for each occurrence of each unigram in the document, are summed.
D	Effort	$AverageParaLength$	Average length of each paragraph in the document. Computed as count of all unigrams in all HTML <p> tags divided by total instances of <p> tags.
D	Images	$ImageCountTag$	Total instances of the HTML <img> tag that appeared in the document.
D	Images	$ImageToText$	Ratio of $ImageCountTag$ to $WordCount$.
D	Links	$OutboundLinks$	The count of all outbound links.
D+	Effort	$Concreteness$	The concreteness score (Brysbaert et al., 2014) of the document measured at the sentence level and aggregated.
D+	Effort	$TermFamiliarity$	The familiarity of terms used in the document, measured by their global corpus frequency.
D+	Effort	$NumberCount$	The count of numbers used in the text of the document.
DS	Images	$Set\_ImageToText$	Set-level calculation of $ImageToText$.
DS	Effort	$Set\_AvgParaLength$	Set-level calculation of $AverageParaLength$.

Table 9.2: The table is a subset of features from the original study (Syed and Collins-Thompson, 2018). The "D" type features are computed treating each document as separate and applying a summation whereas the "DS" type features treat the set of documents as one bag-of-words. The "D+" features are denoted as $\{avg, total\}\_Feature\_\{avg, total\}$ signifying how the feature was aggregated (average or sum) at both the document set level and document level respectively.

Features	**Weight**
(Intercept)	0.11555
total_TermFamiliarity_total	0.05149
total_TermFamiliarity_avg	0.04224
avg_Concreteness_total	0.03991
Set_AvgParaLength	0.02952
total_NumberCount_avg	0.01921
total_Concreteness_avg	0.01775
AverageParaLength	0.01704
avg_Concreteness_avg	0.01302
avg_TermFamiliarity_total	0.01185
total_Concreteness_total	0.0104
ImageToText	0.00969
WordCount	0.00513
ImageCountManual	0.00165
DocumentAgeDifficulty	-0.00791
total_NumberCount_total	-0.02145
avg_TermFamiliarity_avg	-0.02197
OutboundLinks	-0.0262
Set_ImageToText	-0.0286
WeightedWordCount	-0.10807

Table 9.3: Features ordered in descending order of weights. Most positive features are metrics of ease of understanding - concreteness, paragraph length, familiar terms.

learning tasks and participants spent substantially less time). However, we found that the model's prediction on **DS3** was significant but in the *opposite* direction (Table 9.4).

We further analyze the fitted model's feature coefficients (Table 9.3). As all the features were standardized, the coefficients tell us a lot about how much impact each feature has on the model's output. Generally, features that showed the strongest positive association with learning outcomes were metrics of comprehensibility. For example features like term familiarity, term concreteness and paragraph length were some of the highest weighted positive features. Each of these features can be associated with some form of improving comprehension (higher term familiarity would indicate an easier ability to understand the language

DS1	**DS2**	**DS3**
$r_s = .336^{***}$	$r_s = .135^{*}$	$r_s = -.545^{**}$
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1		

Table 9.4: Spearman rank correlations r_s between predicted PLG and actual PLG using fitted model. Similar datasets like **DS1** and **DS2** showed positive and significant correlations. Dataset that was substantially different **DS3** had significant but *opposite* results.

being used; more concrete terms would indicate easier ability to visualize what is being discussed; and longer paragraphs may be associated with articles that provide more than just very basic information). By far the strongest *negative* indicator was weighted word count. This also ties in to the theme of comprehensibility: it follows that articles that are lengthy *and* that use more difficult terminology will be less suitable for a novice learner, especially for a vocabulary learning task.

Relationship of Features with Learning. Thus far we have considered how a model using only **DS1** could generalize to other independent datasets of search as learning. In this section, we instead consider all three datasets and investigate what individual Web document features have consistently strong association with PLG in all three datasets. For each feature in each dataset, we compute its Spearman rank correlation with PLG. We then rank each feature by their lowest cross-dataset p-value (i.e. compute maximum p-value for feature against PLG and sort in ascending order). We compile these results in Table 9.5.

The results generally indicate that word length and structure features (overall count and paragraph length) had the highest consistent correlation with learning outcomes. Other features had more varied results with some features showing strong association in particular datasets but not consistently with the other datasets. Of all the features only the paragraph length features had statistically significant associations with PLG across all three datasets.

Feature	**DS1**	**DS2**	**DS3**
Set_AvgParaLength	0.1657**	0.1873***	-0.5271**
AverageParaLength	0.2339***	0.1149*	-0.4244*
WeightedWordCount	0.2832***	0.1079*	-0.3408.
WordCount	0.2805***	0.0969.	-0.3799*
total_NumberCount_total	0.2459***	0.144**	-0.321.
total_Concreteness_total	0.284***	0.091.	-0.3181.
total_TermFamiliarity_total	0.2851***	0.0881.	-0.3426.
avg_Concreteness_total	0.2134***	0.0766	-0.3576*
avg_TermFamiliarity_total	0.195**	0.0733	-0.4041*
total_NumberCount_avg	0.1328*	0.1305*	-0.2241
avg_TermFamiliarity_avg	0.0678	0.1071*	-0.5229**
total_TermFamiliarity_avg	0.2581***	0.0444	-0.4244*
avg_Concreteness_avg	-0.0445	0.1002.	-0.2835
total_Concreteness_avg	0.2137***	0.0344	-0.368*
ImageCountManual	0.1455*	0.0332	-0.3056.
OutboundLinks	0.216***	0.0849	-0.1124
ImageToText	0.0685	-0.0258	-0.1058
DocumentAgeDifficulty	0.2016***	0.0135	0.1062
Set_ImageToText	-0.1161.	-0.0645	-0.0447
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1			

Table 9.5: Features ordered in descending order of cross-dataset correlation with PLG. Results suggests paragraph length is a strong cross-dataset predictor of PLG. Compared to **DS1** and **DS2**, **DS3** seems to have an opposite relationship with PLG across nearly all features.

9.6 Discussion

One of the main findings from this study was that a model of learning trained using our dataset could show generalizability to an independent dataset that had comparable, though still substantially different, task and experiment settings. We further found evidence that certain features like average paragraph length had significant association with learning outcomes in all three datasets, possibly suggesting a more robust argument for using this as a guiding design feature.

However, one of the unusual results was that the trained model had a significant but *opposite* relationship with learning for the **DS3** dataset. While we don't know for certain why this happened, we note some observations regarding the datasets and their associated experiments. Firstly, **DS1** and **DS2** were quite similar in their experiment design and task settings compared to **DS3**. Both **DS1** and **DS2** were crowdsourced experiments, involved relatively simple forms of learning as per the Bloom's taxonomy (Anderson et al., 2001) and participants generally spent much less time in their tasks. By contrast, **DS3** participants spent between 60 to 90 minutes in their experiments and were given more types of tasks as well as more complex and open-ended tasks. **DS3** participants were not only tasked with learning about their assigned topic but they were also asked to select documents that would help others learn as well. These factors may have meant participants in **DS3** had less priority on efficiency of task completion and more priority on content quality which may have affected what types of documents they were better able to learn from.

9.7 Conclusion

In this work, we investigated how well a model of learning outcomes trained in one study could predict learning outcomes in two completely independent studies. We found early evidence that the model trained on a large, crowdsourced dataset was in fact able to show significantly strong correlation against actual learning scores from both lab-based and crowdsourced datasets. While test set results were statistically significant, we found conflicting results between the crowdsourced and lab-based datasets which may need to be investigated further. The consistency of the results between the training dataset and the holdout crowdsourced dataset was promising considering that these datasets came from studies that involved very different platforms, complexities of learning tasks, topics that were assessed, sample sizes and average time spent by participants. Furthermore, we focused only on document properties that could be automatically and efficiently computed at scale, allowing easy integration with other models as well as in existing large-scale Web search engines. The results suggest that the regression model was able to capture document properties that indicate learning gains in potentially generalizable settings, allowing for future studies to reliably benefit from the regression model and learned weights we provided in this work.

Chapter 10

Investigating Scalable Use of Adjunct Questions to Support Learning (Study 4)

In the previous chapters, we have detailed several major studies conducted to investigate and support learning-oriented objectives in Web search. In this process we used both theory-driven approaches that did not rely on raw data as well as data-driven approaches that leveraged data on how well people learn given the content they read. However, one of the limitations of these approaches is their reliance on passive learning by the user. Here we define *passive learning* as learning by the user when only being provided a static learning resource (e.g. a static Web document). By contrast, numerous studies have found advantages to *active learning* which involve a more dynamic learning process of engagement (e.g. interactive content, feedback mechanisms).

In this chapter, we investigate a particular form of active learning that has long been studied in text materials - the adjunct questions effect (Peverly and Wood, 2001). Adjunct questions are questions inserted into text to draw attention to important textual material (Dornisch, 2012). Many prior studies on the use of adjunct questions effect have found promising results (Callender and McDaniel, 2007; Dornisch and Sperling, 2006; Peverly and Wood, 2001); however, a limitation to these studies is their reliance on manually generated questions. To support scalable benefits of the adjunct questions effect, we investigate the use of an automatic question generation (AQG) API for the purpose of generating adjunct

questions. We further investigate the use of gaze-tracking to personalize when and which questions to ask a particular user based on what they have read so far. We demonstrate that AQG-generated questions showed comparable and sometimes even better learning outcomes compared to human-generated questions. We further show the value of gaze-tracking signals as a metric for predicting both short- and long-term learning outcomes. These results suggest a promising direction for gaze-tracking and adjunct questions effect at scale.

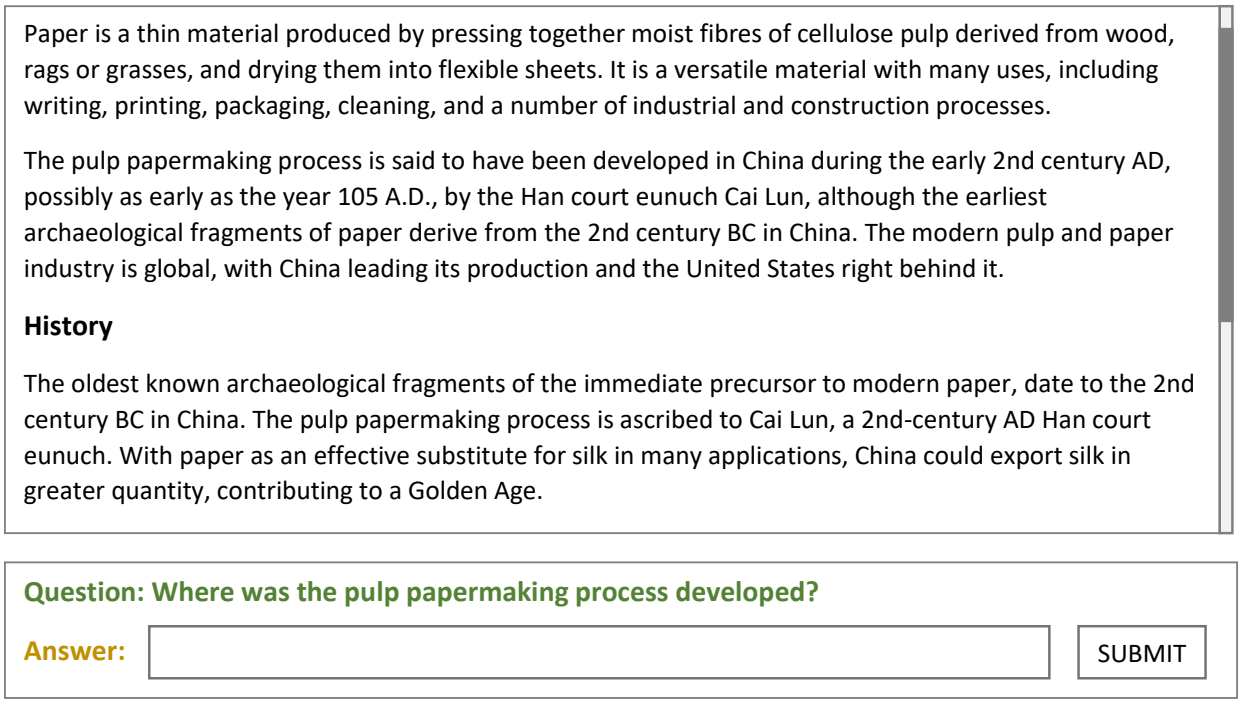

Figure 10.1: Example of Adjunct Questions in an expository text piece.

10.1 Related Work: Adjunct Questions Effect

A core objective of this study was on investigating a scalable form of active learning in text material. Specifically we focused on the *adjunct questions effect* (Peverly and Wood,

2001) which refers to the knowledge improvement found when interrupting the passive flow of reading with questions the learner must address before continuing. Such questions are considered *adjunct questions* (see Figure 10.1 for an example). We build on this form of active learning as it has consistently shown positive knowledge gains in prior work that has investigated it.

In earlier work by Peverly and Wood (2001), the authors reported that augmenting reading material with questions in text led to improved learning. Others also found the use of questions as part of the reading process produced benefits in learning outcomes. Work by Callender and McDaniel (2007) found significantly better learning outcomes among participants who had embedded questions as part of the learning text material. There is also evidence that adjunct questions presented alongside the reading material resulted in both short-term and long-term learning gains Dornisch and Sperling (2006).

However, to the best of our knowledge, all prior studies in this space have used manually constructed adjunct questions. This has significant limitations in applying the adjunct questions effect at scale for arbitrary text documents. In our study, we investigate the use of an automatic question generation (AQG) API as a means of generating content-based adjunct questions at scale. We further investigate the use of personalization to calibrate when and what type of questions should be asked during the reading process.

10.2 Related Work: Eye Tracking and Learning

An early motivation for the use of eye-tracking for observing information processing behaviors was the Eye-mind hypothesis proposed by Just and Carpenter (1980) which stated that in a reading task, "the eye remains fixated on a word as long as the word is being processed." In

other words, the hypothesis claims that there is a direct causal link between where people's visual focus is and what cognitive processing is happening at that location. Later work by Underwood and Everatt (1992) cautioned that this hypothesis may be unrealistically strong, noting simple examples where a reader might stare at the end of the text while taking a moment to reflect on what they read. In such a case, that location of fixation in and of itself doesn't tell us anything about the cognitive processing being invoked (Underwood and Everatt, 1992). Nevertheless, even considering the limitations to the hypothesis, many later studies were able to find great success in using eye-tracking apparatus to better understand user behavior. Before we go deeper into this literature, it may be useful to familiarize the reader with the terminology of some common measures evaluated in eye-tracking (Poole and Ball, 2005):

1. **Saccades.** These are events where the eyes are moving focus rapidly from one point to another. This typically indicates the user is shifting their focus from one fixation point to another. Regressive saccades are a special case where there is evidence of "backwards" movement (e.g. re-finding behavior, comprehension difficulty or oculomotor correction (Eskenazi and Folk, 2017)).

2. **Fixation.** This is an event where the participant's eye is mostly focused on one area or object for a relatively stable time, especially as compared to the rapid movement time characteristic of a saccade.

10.2.1 Defining Fixation Time.

Early work by Inhoff and Radach (1998) showed how there isn't consensus on the best cutoff time to consider a stable eye position as a "fixation". The amount of time that qualifies

as a fixation varies based on various studies but an average minimum is around 100ms. Eskenazi and Folk (2017) suggest it is appropriate to remove fixations that were less than 80ms or greater than 1000ms and Ozcelik, Arslan-Ari, and Cagiltay (2010) uses a minimum of 100ms. Other work by Copeland, Gedeon, and Caldwell (2014a) suggests fixations qualify for durations between 60 and 500ms with a suggested average of 250ms. Work by Joachims et al. (2005) suggest fixations are between 200-300ms while it has also been suggested that an average of 113ms should be considered (Cole, Gwizdka, Liu, Belkin, and Zhang, 2013) and for general reading of English words, a comprehensive study found that an average of 200-300ms should be expected (Rayner, 1998) although it is pointed out that the time for fixations can vary substantially depending on the task (e.g. silent reading had an average time of 225ms whereas typing had an average fixation time of 400ms) (Rayner, 1998).

10.2.2 Why not use Mouse Movements instead?

The reader might instinctively question why we can't simply use mouse movement data instead of eye-tracking data, considering that both involve continuous user actions that can be captured at a very granular scale. Unfortunately, mouse movement data can be indicative of a diverse set of behavior intents, some of which may be completely unrelated to gaze location (e.g. if the user is reading the page in horizontal scanpaths but only uses the mouse to scroll vertically). Prior work has attempted to investigate patterns of mouse movement data with eye tracking data and distinguished between two general pattern differences: *incidental* and *active* mouse usage (Rodden, Fu, Aula, and Spiro, 2008). The cases of *active* mouse usage are those where the user is moving the mouse along with their general gaze as they are processing information on the page and deciding on their action. The cases of *incidental* mouse usage are those where the user may be browsing the page, processing the information, but

only moves the mouse to perform some action such as a click. Rodden et al. (2008) further found that differences in type of mouse usage emerged when considering lookup-oriented search tasks versus more exploratory and open-ended search tasks. While the authors found preliminary evidence that active mouse usage patterns did follow a template of gaze tracking patterns, this still leaves open the question of how to deal with incidental mouse usage. Even if there is active mouse usage, how do we automatically classify between the two types? A later study by Guo and Agichtein (2010) conducted a similar study where they confirmed some of the findings from (Rodden et al., 2008). Their analysis showed that the deviation between the eye gaze and the mouse position showed roughly a roughly normal distribution along both the x- and y-axes and the average Euclidean eye-mouse distance was about 200 pixels. Furthermore, the study by Guo and Agichtein (2010) took initial steps towards building a classifier to predict whether or not the eye-mouse distance was above or below a certain threshold. This can be useful in identifying whether or not the mouse data should be trusted as representing gaze location for any given Cartesian location and point in time. However, the precision and accuracy of their classifier is far from optimal (precision never passed 75% for three separate thresholds that were assessed).

Later work by Huang, White, and Buscher (2012) addressed the concern that not all mouse movement data will align well with gaze location and instead focused on identifying what types of tasks and search situations the alignment *does* show strong correlation. Their study also confirmed an earlier result by Guo and Agichtein (2010) that the eye-mouse distance in both the x- and y-axes are roughly normal with the y-axis having sharper spike around 0. The study also found other interesting patterns of gaze location and mouse movement. They found results supporting the hypothesis that evidence of user interest first manifests in gaze location and is then followed by cursor movement. They found a

700ms time lag between gaze and cursor movement minimized the RMSE error of the eye-mouse distance. More recent work by Papoutsaki, Gokaslan, Tompkin, He, and Huang (2018) similarly investigated the distance between eye tracking and mouse movements and specifically found significant differences in the distances when faceting by touch-typist and non touch-typist users.

While earlier work by Rodden et al. (2008) classified mouse actions into three broad groups: (1) incidental; (2) following; (3) bookmarking, the work by Huang et al. (2012) considered a four-way classification of: (1) inactive; (2) reading; (3) action; (4) examining. The study by Huang et al. (2012) found that if they only considered the *reading* type of behavior, based on heuristics they developed, the average eye-mouse distance went down to 150px (compared to the overall average of 200px in the study by Rodden et al. (2008)).

Overall, we find that there has definitely been progress towards approximating gaze location from mouse movement raw data as well as mouse movement patterns. However, it is also clear that the deviations between mouse movements and gaze locations can vary substantially and may also vary differently depending on task type and nature of mouse movement pattern. While there is promising potential for using proxy signals like mouse movements as approximations, we will be using more accurate gaze-tracking using a commercial gaze-tracking device for our study.

10.2.3 Eye movements and Search Behavior

Prior studies have investigated how eye tracking data can inform better understanding of how users engage with a search engine in an information seeking task. Early work by Salojärvi, Kojo, Simola, and Kaski (2003) investigated whether eye tracking signals could be used for prediction in information retrieval. In their study, participants were shown a task assignment

and then presented with a list of titles, some of which contained answers to the assignment and others did not. The authors found that eye-tracking signals could discern between relevant and non-relevant titles with clear differences in a two-dimensional PCA projection. A limitation of their work was the use of only 3 participants which a later study by Joachims et al. (2005) improved on. The study by Joachims et al. (2005) found that eye-tracking data could provide a more exact understanding of how much time users spend analyzing different possible links in the SERP page during an information-seeking task. While the eye-tracking data in their study indicated the intuitive finding that users will typically reach lower-ranked documents after more total fixations, it also showed how differences may emerge in the amount of fixations spent at each rank. The authors further used eye tracking signals to look deeper into whether or not a document click could be considered as a guarantee of relevance judgment. Other work by Pan, Hembrooke, Gay, Granka, Feusner, and Newman (2004) focused on investigating how eye-tracking signals differed based on the contents of the actual sites being visited. They found significant differences in mean fixation duration based on which websites were viewed (content domain), the order in which they were viewed and the gender of the participants.

10.2.4 Eye movements and Knowledge

Recent work by Bhattacharya and Gwizdka (2018) investigated how eye movement behavior differed between those who showed low and high changes in knowledge (KG) during a Web-based learning task. Their study found that those who showed higher KG also tended to have less fixations in sequences of fixations and also tended to spend less time per fixation. There was also evidence that the low KG group tended to show more and longer backwards regressions. All of this seems to suggest that those who showed greater knowledge change

actually appeared to be spending *less* effort and time than those who achieved worse results. However, it is possible that the high-KG users were actually more *efficient* and smart searchers than the low-KG group and thus were able to find what they needed quickly and effectively.

Earlier work by Cole et al. (2013) also investigated links between eye movement behavior and user prior knowledge in the medical domain and found that features like perceptual span and reading time were strongly predictive of prior knowledge. Later work by Mao et al. (2018) further applied the main eye movement variables from (Cole et al., 2013) to also investigate the link between domain-specific knowledge and eye movement behaviors and reached similar conclusions. In a similar direction, later work by Copeland, Gedeon, and Mendis (2014b) investigated the use of a neural network to predict learning outcomes using raw eye movement behaviors and work by Copeland and Gedeon (2013) similarly used a feedforward neural network to predict learning outcomes from eye movement behavior. In addition to understanding how eye movements predict learning, it is also useful to understand general eye behaviors in the context of learning. Work by Copeland and Gedeon (2014) found that in a learning task, most visual attention is paid to the first and last paragraphs of text content. As people have limited effort budget they will expend, this might suggest content producers should try to maintain user interest in those paragraphs more so than in intermediate paragraphs. Their study further established early evidence that people will show more fixations on paragraphs containing the answers to questions even before the questions are presented in the case of cloze test assessments. This may suggest that even without knowing what specifically will be assessed, people tend to be good judges of what content likely will be important for them to learn and what would not. It remains an open question of how people's eye movements within the correct paragraph change in the process

of locating the fact to learn.

Several studies of eye tracking have also investigated how eye movement behaviors change when the user is aware of an explicit learning task to accomplish versus not being aware. Work by Copeland and Gedeon (2013) found that users who spent more time reading text material prior to knowing what explicit learning tasks they had to accomplish spent *less* time reading that same material when they had the chance to go back to it. This is a fairly intuitive finding and as the authors point out, the time spent and the number of fixations can be influenced by many factors including changes in topic understanding, motivation, re-finding actions and so on.

10.2.5 Applications of Eye Tracking for Learning

Other studies have investigated possibilities of how eye tracking could be used as applications for supporting learning. A study by Copeland et al. (2014a) presented a framework for providing adaptive difficulty in text content as a function of estimated user comprehension level which would be determined through eye tracking. Earlier work by Sibert, Gokturk, and Lavine (2000) introduced The Reading Assistant, which was an adaptive tutoring system that helps users struggling with understanding a particular word by detecting their eye movements and taking appropriate personalized action in the form of auditory feedback.

10.3 Study Design

We are interested in exploring how adjunct questions presented *during reading* impact learning outcomes. Questions are selected from parts of the text which the user had just read, determined using gaze input from an eye tracker. We compare questions that are auto-

matically created from a question generation system to questions that are manually created by humans. *Learning outcomes* are measured by how well participants are able to answer questions about the content (different from the adjunct question asked during reading) after they had finished reading.

We now describe our research questions as well as the design and data preparation for our study.

10.3.1 Types of Questions and Method of Assessment

There are different ways of classifying types of *questions*. We consider two complementary types of questions: (1) factoid/low-level; and (2) synthesis/high-level. In Bloom's taxonomy (Anderson et al., 2001), factoid questions are questions that address the "Remember" level of cognitive complexity, whereas synthesis questions address the "Analyze" level of complexity. Factoid questions may be those that ask about specific facts, locations, numbers, times, etc. that can often be found directly in the text. Synthesis or high-level questions require the participant to search through multiple paragraphs, combining information from these to form a correct answer. In principle, synthesis questions would require more integration of different facts and thus more effort to answer correctly.

While there are many ways of assessing learning, we measured learning outcomes by asking participants to write short free-form answers to the above question types about the content. Although our study asked both factoid and synthesis questions, most of our analysis will focus on participants' answers to factoid questions, since the presentation of synthesis questions was specific to a single condition. Factoid questions are fairly straightforward to grade, typically having objectively correct answers which makes grading easier. Our evaluation of the correctness of participants' free-form answers was done via careful crowdsourcing;

further details are given in the Grading section below.

To produce the automatically generated questions used in our study, we trained a generative model similar to Wang, Yuan, and Trischler (2017). We later refer to this service as AQG API for automatic query generation API.

10.3.2 Research Questions

We aim to answer the following research questions:

RQ1: Do participants show any difference in post-reading learning scores using attention-based, dynamically-presented questions during reading, compared to a non-interactive condition?

RQ2: Do participants show any difference in post-reading learning scores when asked questions from a human-curated source versus an automatically generated source?

RQ3: Do participants show different outcomes or behaviors when given only factoid questions, versus being given factoid questions plus an additional synthesis question?

RQ4: Do participants show any difference in learning outcomes when the system incorporates participants' gaze focus history to select questions?

RQ5: Are there characteristics of participant gaze data that are potentially indicative of lower vs. higher learning?

RQ6: For all the above questions, how do results compare between short-term learning (assessed immediately after reading) versus long-term retention (assessed after a one week delay)?

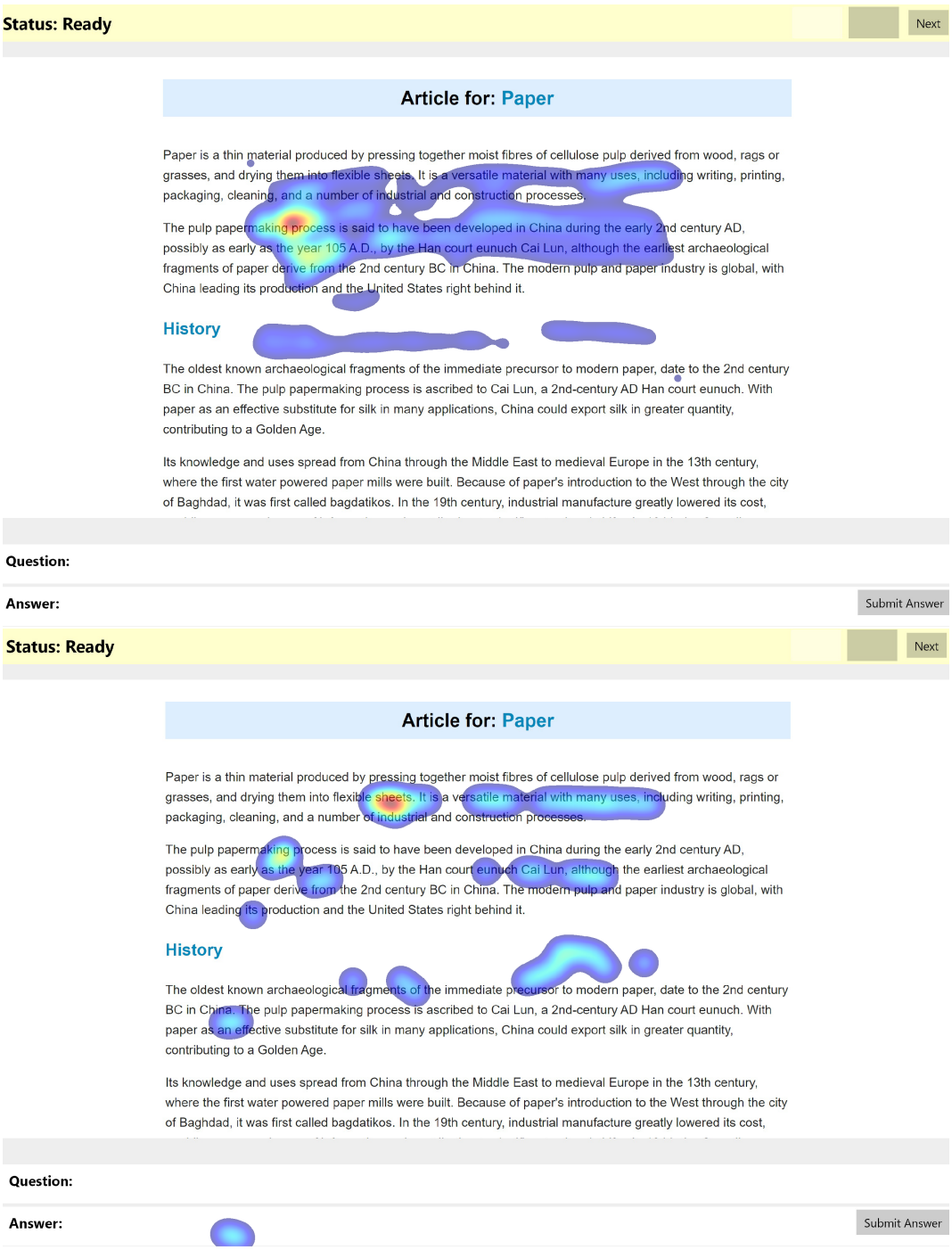

Figure 10.2: Gaze fixation heatmap on article page for a participant on topic 'paper'. Question/response area is below the content area. **Top:** Fixation heatmap before a question was asked. **Bottom:** Fixation heatmap after a question was asked: "What is a common use for paper?".

To answer these questions, we designed a study where participants took the role of learners and read Wikipedia articles while their gaze behavior was tracked. Gaze fixations were used to determine what parts of the article the participants had read and how. Adjunct questions were generated from the text that the participants had shown gaze fixations on based on the conditions listed below (with implementation details provided in Chapter 10.4).

10.3.3 Reading Material

Participants read reconstructed Wikipedia articles as a principal learning resource. As mentioned above, by using Wikipedia articles covered in the `SQuAD` question-answering dataset (Rajpurkar, Jia, and Liang, 2018), we had access to many curated question and answer (q, a) pairs for every paragraph in the article.[1] Furthermore, as one of the most visited websites, participants were likely to be familiar with the design and content structure of Wikipedia articles. Because the content and structure of the articles may have evolved since they were used in the creation of the `SQuAD` dataset, we recreated the original article by concatenating the set of paragraphs from the `SQuAD` dataset in sequential order. We verified that each of the reconstituted articles maintained coherent reading flow from start to finish. The result was a set of useful articles for which we had an exact mapping for each question to the passage containing the answer.

We chose a set of four articles for our study that covered diverse topics ('Economy of Greece', 'Norfolk Island', 'Pain' and 'Paper'). Once we had reconstituted these articles, we also produced a new set of questions, one for each paragraph, that was automatically generated using our AQG API on the same set of paragraphs, with the intention of comparing

[1] The original `SQuAD` questions were crowdsourced in a task where crowdworkers were provided a paragraph and instructed to ask 3-5 questions about the content. They were especially encouraged to ask difficult questions (Rajpurkar et al., 2018)

auto-generated questions with crowdsourced questions in a learning task.

10.3.4 Determining Reading Attention State

We aggregated gaze data at the paragraph level within a document. To determine whether a participant was "skimming" versus more deeply "focus-reading" a paragraph, we employed a common approach using a statistic called Normalized Number of Fixations (NNF) (Copeland and Gedeon, 2014). We defined NNF for a paragraph as the total fixation events focused on that paragraph normalized by the total word count of that paragraph. For a given participant, we denoted "Skim-Reading" questions as those whose answer was in a paragraph that the participant was determined to have skimmed based on the NNF for that paragraph being too small ($0<\text{NNF}<0.7$). We chose the threshold of 0.70 based on prior work (Copeland and Gedeon, 2014)). We considered a paragraph for generatingv "Focus-Reading" questions if its NNF was at or above this threshold ($\text{NNF} \geq 0.70$).

10.3.5 Adjunct Questions

We implemented four conditions reflecting how the questions were presented in our study.

In an adaptive condition ($\mathbf{Q_{Auto}}$), our system used an Automatic Question Generation system to generate questions based on the paragraphs where a learner's visual attention had been, as indicated by a dynamic gaze tracking model while reading in real time.

To contrast automatically generated question presentation with human-curated questions, we included a condition ($\mathbf{Q_{Human}}$) where the system also adaptively presented questions, but used ones taken directly from the `SQuAD` question-answering dataset. We chose this dataset for three reasons: (1) the questions are manually curated and associated with a small passage rather than the whole document; (2) the questions are based on Wikipedia

articles, which are a commonly-used source for learning on the web; (3) SQuAD has been used extensively in the deep learning literature as a benchmark, and state-of-the-art models are available to automatically generate questions similar to SQuAD- style questions based solely on input passage text from a reading source. Thus, using SQuAD enabled us to compare manually curated questions with automatically generated questions that are meant to emulate the same style.

We added another condition ($\mathbf{Q^*_{Human}}$) that was identical to using the manually curated questions from SQuAD but which also included a high-level synthesis question. This condition enabled us to create a common approach seen in learning settings directed by a teacher, where the majority of questions focus on simple factoid questions to encourage basic learning, and a synthesis question is used to encourage higher-level thinking. Our design also enabled us to evaluate potential benefits to asking high-level questions in this setting.

Finally, as a control condition ($\mathbf{Q_{None}}$), we presented a non-interactive system that asked no questions during reading: only pre- and post-test questions were presented. This provides a condition where a learner does undirected learning by reading.

10.3.6 Measuring Learning Outcomes

We measured how well participants had learned the content by asking questions based on the text immediately after they finished reading (*post-test*) and after a week (*delayed*). *Delayed* questions allowed us to distinguish between short-term memorization learning and more permanent retention effects. To measure prior knowledge, we also asked questions on the content prior to reading the article (*pre-test*).

To reduce question priming effects, we designed our pre-test, post-test, and delayed test questions so that there was no overlap with adjunct questions shown *during* reading in any

of the conditions. The post-test questions were designed to be a superset of the pre-test questions, so that we could separately measure knowledge gain for those questions where we measured the learner's prior knowledge before reading the article. We refer to the set of questions given in the pre-test (and repeated in the post- and delay-test) as the *Base* questions (Q_{Base}). Because the pre-test introduced the possibility of a priming effect where learners are implicitly primed to look for the answers to pre-test questions, the post- and delay-test also contained questions not shown during the pre-test. We refer to this set of questions not shown during the pre-test and only shown during the post- and delay-test as the *New* questions (Q_{New}). These *New* questions enable us to measure learners' knowledge gain on a set of questions that had no possibility of a priming effect.

All questions were designed as requiring short, free-response answers, to avoid allowing learners to simply guess the answers and to provide a richer source of response data to analyze in the future for learning effects. In post-hoc analysis, questions were graded through crowdsourced judgments.

10.4 Methodology

The main user interface across all conditions consisted of an article viewing window that rendered the Wikipedia article. As participants read the article, our gaze tracking package indicated for each paragraph if the reader likely skimmed (S) the paragraph or performed focused reading (F). In all of the conditions which asked questions, we alternated, when possible, between these two types of paragraph when selecting questions, in order to average out any potential impact across conditions.

The adjunct questions were presented in a question prompt panel fixed at the bottom of the window (see Figure 10.2). The condition assigned at any given point determined

which questions (if any) would be asked during the reading phase. If questions were asked, users would be required to submit an answer to each before being allowed to continue to the post-reading test. In this study, users would not get any feedback as to whether their question was right or wrong. We chose to not give feedback as it would introduce another confound as well as due to the difficulty in automatically (real-time) assessing the validity of a free-response question. The experiment design involved four conditions (described below) in a within-subjects design.

1. $\mathbf{Q_{Auto}}$. In this condition, the bottom panel displayed a new question approximately every $K = 3$ paragraphs that the participant skimmed or focus-read (measured by gaze tracking). The system alternated the type of paragraphs from which questions were drawn in the order S, F, S, F (S questions are from paragraphs the participant skimmed over; F are from paragraphs the learner showed focused-reading over). Each participant answered *exactly* four questions based on what paragraphs they had gaze fixations over. Questions were automatically generated from paragraphs using the AQG API source.

2. $\mathbf{Q_{Human}}$. (`SQuAD`). Same in design as the $\mathbf{Q_{Auto}}$ condition but all the questions were selected from the `SQuAD` source.

3. $\mathbf{Q^*_{Human}}$. Same design as the $\mathbf{Q_{Human}}$ condition but the system asked a high-level synthesis question in addition to the four factoid questions.

4. $\mathbf{Q_{None}}$. No Questions. In this condition, the bottom panel remained blank throughout the reading phase for a particular topic.

Each participant in the study completed four learning tasks.[2] There was one task per

[2]In a pilot study we chose six topics. Participants reported the experiment took too long and individual

condition, with the ordering of conditions randomized – where each learning task consisted of a pre-test, reading phase, and post-test. The four topics were randomly ordered across the tasks in order to help ensure ordering effects were balanced on average across participants with respect to topic and condition.

10.4.1 Participants

To determine the number of participants needed, we conducted a statistical power analysis with significance level of $\alpha = 0.05$ and power of $1 - \beta = 0.80$ and a medium expected effect size by Cohen's d ($d = 0.50$). This gave a base requirement of $n = 51$ participants; to accommodate an attrition rate of 20% required $n = 64$ participants. In the actual experiment we ended up recruiting $n = 80$ participants, well beyond the required number.

The experiment was conducted in a lab setting and subjects were recruited through a recruitment email sent to the UMSI Experiment Server at the University of Michigan where we gave an overview of the experiment and what would be expected of participants in terms of time and nature of the task. There were 21 male and 58 female participants with 1 reporting other gender. Ages ranged from 18 to 50 with a median of 21 and all participants had at least a high-school level of education.

During the experiment some participants had faulty experiences with the eye-tracker that resulted in requiring a manual override. We removed the specific (participant, topic) pairs where this occurred from analysis. There were also two participants who reported not being aware that there was more to read for one of the topics and had clicked ahead without getting a chance to read the full article. We have omitted these (participant, topic) pairs as well. Furthermore, there were a small number of participants who simply did not complete

articles were too long. We reduced to four topics and reduced content length by 25% for the full study.

the four topics in the allotted two hours time. In these cases, we still include the data for topics that they did complete. In total there were 18 (participant, topic) pairs that were removed from analysis.

For the post-test session, 72 of the 80 participants completed the delayed test.

We compensated participants in the form of a base amount of USD 12 for taking part in the study along with an additional compensation of USD 13 contingent on how many answers in the during-reading and post-tests they answered correctly. In total there were 57 such questions, with the USD 13 evenly split across each correct answer. Thus each participant could earn a maximum total of USD 25 in the first part of the study. The same participants would then return for the second part of the study where they would earn a lump sum of USD 5 for participating, resulting in a final maximum of USD 30 per participant.

10.4.2 Procedure

We structured the experiment procedure into the following phases:

1. **Gaze Tracking Check.** Before beginning the experiment, all participants completed a personalized gaze calibration using commercial software. In addition to this, before proceeding, a second-stage gaze-tracking check was performed using the main application we developed for this study.

2. **Instructions.** Participants read through the instructions of what the task entails and what was expected of them. Following this screen, the participant started the main experiment.

3. **Pre-test.** This comprised a set of five (5) free-response questions about the topic (covering an initial subset of all the questions we eventually wanted to assess).

4. **Reading phase.** Participants were provided a Wikipedia article corresponding to the topic. This phase was where we implemented the four different conditions described above – in particular that varied whether any questions were presented during reading and if so, what the source of the questions was.

5. **Post-test.** Another test was administered that was also free-response and which included all of the questions asked in the pre-test but also included five (5) unseen questions, for a total of ten (10) questions.

6. **Repeat.** The participant repeated steps 3-5 for each of the remaining topics.

7. **Demographics/Survey.** Participants completed a demographics survey which also included questions regarding their use of search engines and Web documents for learning.

8. **Delayed Post-test session.** Following a one-week period, all participants were provided a follow-up assessment that comprised exactly the same questions used earlier in the immediate post-tests for each of the four topics. The order of the topics and of the questions was re-randomized for each participant in the delayed test.

10.4.3 Grading

Since the answers were free-response answers, we had to manually grade them. To do this, we crowdsourced graded judgments on the correctness of the question responses using the Figure Eight platform.[3] We restricted the worker pool to those who: (1) had the highest quality rating on the platform (level 3); (2) were from either the US or Canada and (3)

[3]Formerly Crowdflower

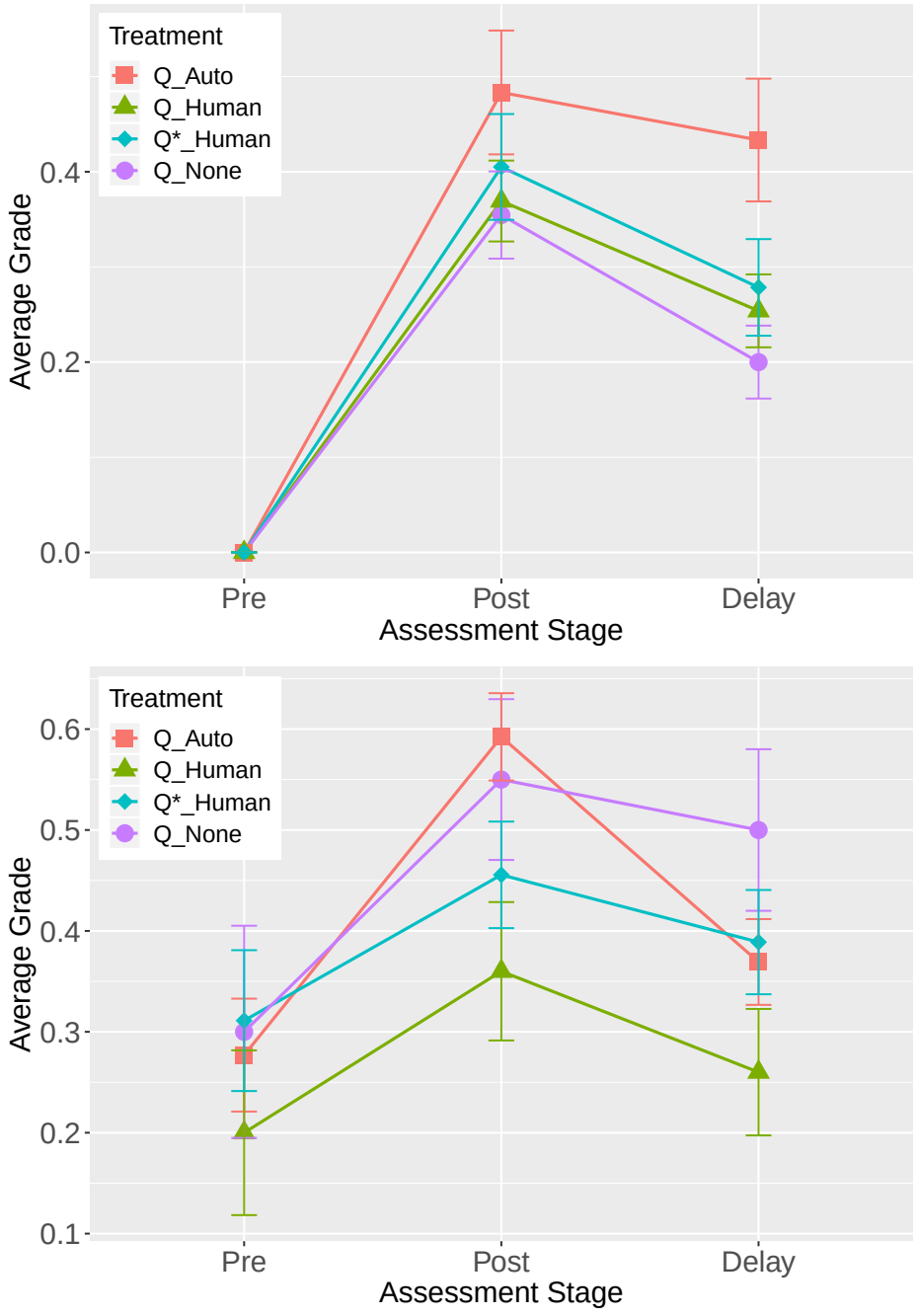

Figure 10.3: Breakdown of average test item scores at each stage, showing that in general both short-term and long-term learning is happening for all conditions. **Top:** Low-knowledge (LK) learners. **Bottom:** High-knowledge (HK) learners. Error bars are standard errors.

who were able to correctly grade several gold standard exemplar responses. For each unique (paragraph, question, answer) tuple we crowdsourced three (3) graded judgments and took the majority class response as the adjudicated answer.

10.4.4 Data Preparation and Filters

Due to the experiment setup and based on participant feedback, there were clear signs of fatigue/boredom that impacted behavior and performance after the first topic/condition in a session. For this reason, in the present paper we simplify our analysis to only the first topic/condition that a participant completed and a between-subjects analysis. We leave the remaining data for future analysis. This filter reduces our dataset sample size by about 75% from $n = 2718$ to $n = 689$ for post- and delay-test results and from $n = 1360$ to $n = 345$ for pre-test results.[4]

The amount of knowledge a learner has before reading about a topic may impact both performance and the ideal experience. To control for this and deal with chance differences across topics/conditions, we stratify the analysis based on knowledge demonstrated in pre-test. We consider a participant to be *low-knowledge* (LK) for a particular topic if they got *all* pre-test answers for that topic incorrect. Otherwise, if they answered at least one question correctly for topic, they were considered *high-knowledge* (HK) learners. Nearly 47% participants were classified as low-knowledge through this approach. After this stratification, our data was split in a 4x2 design (conditions x learner knowledge). There were no significant differences in pre-test scores by condition when split by learner knowledge.

[4]The pre-test results have half the number of data points because the pre-test has half as many questions as post- and delayed post-test.

10.5 Results - Learning Outcomes

We present an analysis of learning outcomes here, and an analysis of real-time reading behavior patterns in Chapter 10.6.

10.5.1 Overall Learning Trends

We first present, as a sanity check, the overall trends in learning gains from the pre-test, to the immediate and delayed post-tests in Figure 10.3. Participants achieved both short- and long-term learning gains in all conditions. Long-term (delayed post-test) learning as measured by overall grades dropped somewhat compared to short-term (immediate post-test) grades but was still significantly higher than the initial pre-test baseline for every condition on average after reading the topical material. LK participants generally showed stronger improvements as they were starting from zero prior knowledge while HK learners showed more variation. These results help validate our experimental setup and that participants on average are indeed learning.

Table 10.1 presents an overall summary of learning outcomes and time patterns, stratified by LK and HK participants as well as the four different conditions.[5] Our significance computations for the grade performance comparisons compared each of the interactive question conditions solely to the $\mathbf{Q_{None}}$ condition (using the Chi-Squared test), since our main focus is on first replicating the adjunct question effect in this dynamic setting. For task time comparisons, we seek to understand the tradeoffs across all conditions and used an omnibus Kruskal-Wallis test.

[5]Note that pre-test sample sizes are half of post-test size because there are half as many questions in the pre-test.

Measure	$\mathbf{Q_{None}}$	$\mathbf{Q_{Auto}}$	$\mathbf{Q_{Human}}$	$\mathbf{Q^*_{Human}}$
Low-Knowledge Learners				
Sample Size	110	60	130	79
Pre-score Base	0.00	0.00	0.00	0.00
Post-score				
All	0.35	0.48	0.37	0.41
Base	0.44	0.60	0.43	0.52
New	0.27	0.37	0.31	0.28
Delay-score				
All	0.20	0.43**	0.25	0.28
Base	0.22	0.50*	0.26	0.18
New	0.18	0.37	0.25	0.38.
High-Knowledge Learners				
Sample Size	40	130	50	90
Pre-score Base	0.30	0.28	0.20	0.31
Post-score				
All	0.55	0.59	0.36	0.46
Base	0.75	0.71	0.56	0.62
New	0.35	0.48	0.16	0.29
Delay-score				
All	0.50	0.37	0.26*	0.39
Base	0.65	0.48	0.32.	0.51
New	0.35	0.26	0.20	0.27
Time Patterns				
Task Time (sec)***	519.0	1025.	850.3	1200.
Task Time (sec) (No_QA)	519.0	764.9	648.1	772.6
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1				

Table 10.1: Average values for different learning measures by condition. Marked values indicate significant differences b/w that condition and $\mathbf{Q_{None}}$. Also shown is breakdown by question type: Base (seen in pre-test), New (post-test only), and All (Base+New).

We observe that $\mathbf{Q_{None}}$ generally exhibited the worst long-term results for LK participants but showed the best results for HK participants. We refine this analysis further in the following sections, presenting results for each of our research questions.

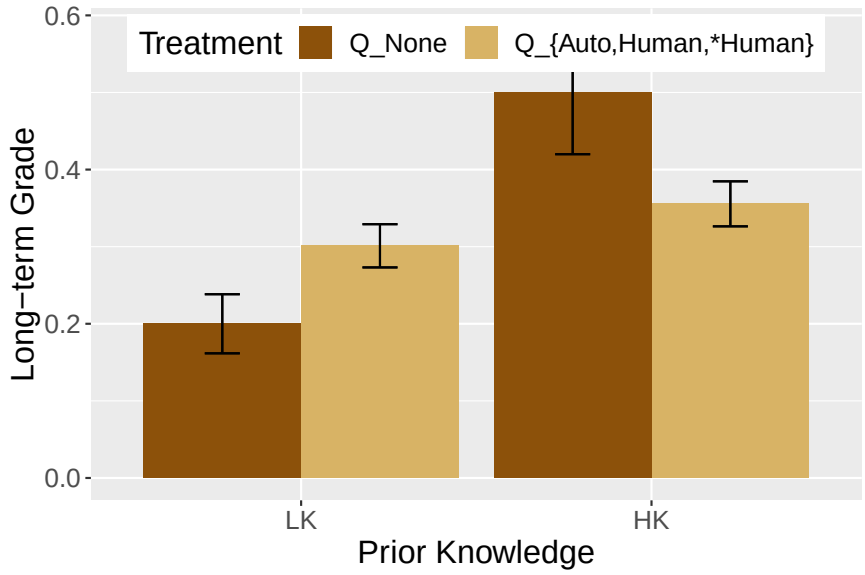

Figure 10.4: Breakdown of long-term grades by condition and knowledge level. LK participants particularly benefit from interactive conditions.

10.5.2 Effects of Adjunct Questions on Learning

In **RQ1**, we asked if participants show any difference in post-reading learning scores using adjunct questions that are dynamically presented while reading based on their gaze, compared to when no questions are presented. We found that learners who received adjunct questions while reading had significantly higher grades in the delayed post questions $\mathbf{Q_{None}}$ participants (M=.30 vs M=.20, p=.04). For HK learners there was a slight decline in long-term grades, but this difference was not statistically significant (M=.36 vs M=.50, p=.08). Neither LK nor HK showed significantly different short-term grades. These results are shown in Figure 10.4. This suggests that adjunct questions has a positive effect on long term retention of content for those who have no prior knowledge on the topic; however, the adjunct questions may not be as beneficial for those who already have some knowledge of the topic, and perhaps impede their natural reading flow.

10.5.3 Effects of Adjunct Question Source on Learning

In **RQ2**, we asked how learning outcomes measured through post-reading learning scores compared across the auto-generated and human curated questions. For fair comparison, we omit $\mathbf{Q^*_{Human}}$ from this section's analysis.

In general, we found that participants in the automatically generated questions condition ($\mathbf{Q_{Auto}}$) showed better results in the short- and long-term for both LK and HK learners. LK learners showed significantly better results in the long-term (M=.43 vs M=.25, p=.01) whereas HK learners showed significantly better results in the short-term (M=.59 vs M=.36, p=.005).

We explored what may have been driving these improvements relative to the $\mathbf{Q_{Human}}$ condition. In terms of differences in questions, we found that $\mathbf{Q_{Auto}}$ questions were about 11% longer (by token count) than $\mathbf{Q_{Human}}$ questions (M=12.72 vs M=11.43, p=.003) possibly indicating more detailed questions may have encouraged more fine-grained reading behaviors. When we examined the reading behavior data, we found that participants in the $\mathbf{Q_{Auto}}$ condition had significantly more normalized regression fixations (M=.060 vs M=.043, p=.01). Prior work has linked reading regression fixations to concentrated reading behavior (e.g. re-reading, confusion clarification), and this evidence helps support our hypothesis that these detailed questions gave rise to more focused reading and the difference in performance. Interestingly, AQG questions may often appear too detailed and simplistic (as simple textual rewrites of input passages) at first glance, but in a learning scenario these exact properties may help readers quickly find the right passage in the document and then require focused reading which results in greater learning.

10.5.4 Effects of the Synthesis Question on Learning

RQ3 asked if learning outcomes were different when synthesis questions were asked in addition to factoid questions, compared to just asking factoid questions.

We saw no significant gains relative to the other question conditions when adding a synthesis question. For LK learners, we did see higher long-term grades compared to $\mathbf{Q_{None}}$ for New questions (Table 10.1). This may be in part due to the extra time on task (see Chapter 10.6.1) that is spent when a synthesis question is asked.

10.5.5 Skim- vs Focus-Reading Adjunct Questions

In **RQ4** we asked if learning outcomes varied based on questions that were selected based on gaze focus patterns. More specifically, we wanted to see if differences existed in the outcomes when participants had skimmed over content, versus focused reading, which we could determine through our gaze tracker.

Recall that in our experiment design, for all conditions except $\mathbf{Q_{None}}$, we asked each participant four factoid questions. These questions could be generated from paragraphs that were skimmed ('S'), or those that were read with deeper, focused reading ('F'). Our system attempted to interleave these two different question focus types in the order (S, F, S, F). Because some participants showed focused reading throughout, the system never got to ask them skim-reading questions. In this section, we analyze if those participants who got at least one skim-reading question showed different learning outcomes than those who didn't. We denote this binary variable as **GotSkim** and denote those who got at least one skim-reading question as $\mathbf{GotSkim}_{YES}$ and those who did not get any skim-reading questions as $\mathbf{GotSkim}_{NO}$.

We start this analysis by initially excluding $\mathbf{Q_{None}}$, as participants in this condition *could not* possibly get any adjunct questions. We found that both LK and HK learners showed significantly better long-term grades when they got at least one skim-reading question ($\mathbf{GotSkim}_{YES}$). In particular, among LK learners, $\mathbf{GotSkim}_{YES}$ participants strongly outperformed $\mathbf{GotSkim}_{NO}$ participants (M=.40 vs M=.27, p=.04). This gain was also evident for HK learners (M=.48 vs M=.32, p=.02). For short-term grades, LK learners had nominally worse grades but this difference was not statistically significant in $\mathbf{GotSkim}_{YES}$ (M=.31 vs M=.44, p=.07). HK learners also showed no significant differences in the short-term. These results suggest that those participants getting questions based on skimmed reading may have been motivated to reread more carefully to answer the question - which resulted in better long term retention. These results indicate the potential importance in having dynamically-chosen, focus-based, adjunct questions for better long-term results.

10.6 Results - Reading/Time Patterns

In Chapter 10.5, we analyzed learning outcomes across the four experiment conditions, faceted by different types of questions. Here we analyze participant reading behavior patterns detected via gaze tracking over time and how they relate to learning outcomes, addressing **RQ5**. We first analyze time patterns, and then specifically analyze reading fixation patterns.

10.6.1 Variation in Time Across Conditions

We analyzed how the total time spent reading each article varied depending on the assigned condition, where total time spent is defined as the timestamp difference between the first and last gaze event on the article. As expected, $\mathbf{Q_{None}}$ had the lowest average time, $\mathbf{Q_{Auto}}$

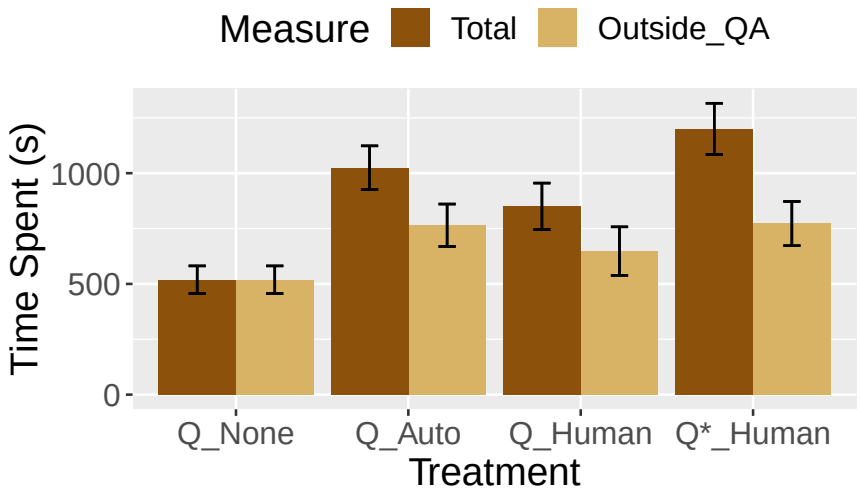

Figure 10.5: Breakdown of average reading time by treatment. **Outside_QA** is the reading time not spent answering questions. Results suggest that being given questions encourages participants to spend more time reading excluding time needed to answer questions.

and $\mathbf{Q_{Human}}$ had comparable averages, and $\mathbf{Q^*_{Human}}$ had the highest average time: this matches the approximate activity level these conditions required from the participants. See Table 10.1 for details.

To examine how the additional requirement of answering questions affected participants' time on task, we subtracted the time participants spent actually answering questions from the total time they spent on the topic.[6] After this subtraction, the significance of the above total time differences across the conditions drops sharply, suggesting that participants may have been spending limited additional time outside of the task requirements. The three interactive conditions generally had higher averages of time spent outside of question-answering compared to $\mathbf{Q_{None}}$ though these differences did not reach statistical significance.

[6]We compute the time spent answering a question as the time elapsed from being asked a question to submitting an answer for it.

10.6.2 Change in Reading Behavior when Asked Questions

We hypothesized that when questions were generated, participants would direct their attention to the paragraph containing the answer. To test this, we first computed the total fixation count for all three types (Skimming, Reading and Regression) at two times: (1) before a question was generated and (2) in the time between question generation and user answer submission. To account for differences in the duration of these ranges, we normalized these fixation counts by the total fixations on the article in those time spans, producing a *fixation ratio* measure.

Overall, we found strong evidence for our hypothesis: participants did indeed allocate more attention (fixations) to reading target paragraphs when asked a question, compared to before being asked (M=0.48 vs M=0.16, $p<.001$).

10.6.3 Relationship between Read Time and Post-Test Grades

We investigated the relationship between how much time participants spent attending to an article, and their immediate and delayed post-test grades for questions on that article. We define Article Read Time as the elapsed time between the first and last gaze event triggered on the entire article. We found Article Read Time was positively correlated with both post-test grades (ρ=.19, p=.12, n=69) and delay-test grades (ρ=.27, p=.02, n=69), according to Spearman correlation, although the correlations were not significant in either the LK or HK breakdown (likely due to the small sample size).

10.6.4 Relationship between Reading Fixation Behavior and Learning Outcomes

To explore the research question:

> **RQ5**: Are there characteristics of participant gaze data that are potentially indicative of lower vs. higher learning?

we investigated the relationship between normalized number of fixations (NNF) and post-test scores. We define NNF as the total number of reading fixation events on a paragraph divided by the word count of that paragraph. Our gaze reading tracker fired separate fixation events for different expected reading states: (1) Reading; (2) Skimming; and (3) Regression Reading. All of these fixation types were accumulated into an overall NNF score (NNF All), as well as individual NNF scores for each fixation type. Table 10.2 shows a comparison between NNF scores for correct vs. incorrect answers on a paragraph, expressed as a percentage change, including a break-down by fixation type.

We found that when users correctly answered post-test questions, their corresponding overall NNF scores tended to be higher, with strong significance (M=1.558 vs M=1.335, p=.0017**). It should be noted that the NNF scores observed were almost 1.5 times as high as the average found in prior studies (Copeland and Gedeon, 2014). However, we also used significantly longer articles by word count and a number of participants reported in feedback that the articles were difficult. A greater number of fixations per passage is expected in such a case, as demonstrated by Rayner, Chace, Slattery, and Ashby (2006). We found no statistically significant difference in overall NNF scores for long-term learning outcomes (M=1.464 vs M=1.420, p=.6878). However, upon further analysis, we did find significant differences when considering specific *types* of fixations (like skimming and reading

Measure	LK Learners	HK Learners
Post-test results		
NNF (All)	29.36%***	3.742%
NNF Skimming	34.17%***	7.249%
NNF Reading	20.75%*	-1.53%
NNF Regression	**92.83**%***	**22.45**%
Delayed post-test results		
NNF (All)	14.76%	-7.90%
NNF Skimming	23.82%*	-2.73%
NNF Reading	3.564%	-14.3%.
NNF Regression	**37.88**%.	**-1.24**%
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1		

Table 10.2: Percentage increase in NNF scores for correct vs. incorrect answers on a paragraph, overall and by fixation type. LK learners exhibited relatively more active regression reading (large Regression NNF scores) for correct answers.

regressions).

Broken down by fixation type, we found that in the case of low-knowledge learners, the Skimming and Regression NNF scores were significantly higher for correct answers both for immediate and delayed post-tests. Across fixation types, NNFs were significantly different for LK learners but with no conclusive differences for HK learners. This may suggest that the use of NNFs as a method of estimating a learner's short- and long-term knowledge could be particularly precise in identifying low-knowledge users.

10.7 Survey Analysis

In the demographics/search usage survey, we collected demographics information including: (1) age; (2) gender; (3) level of education. All 80 participants completed this survey. We also gathered information regarding their search usage, asking the following questions which had either multiple-choice answers (MC) or free-form answers (F):

Result	**Short-term learning**	**Long-term learning**
Adjunct Questions improved grades better than $\mathbf{Q_{None}}$ (Chapter 10.5.2)	No	Yes (for LK learners)
$\mathbf{Q_{Auto}}$ performed comparable to $\mathbf{Q_{Human}}$ (Chapter 10.5.3)	Yes	Yes
Synthesis question affected grades (Chapter 10.5.4)	No	No
Focus-based question selection improved grades (Chapter 10.5.5)	No	Yes
Gaze behavior was different for those who would answer questions correctly (Chapter 10.6.4)	Yes (for LK learners)	Yes (for LK learners)

Table 10.3: Major conclusions regarding learning outcomes and reading behaviors/treatments.

1. How often do you use Web search engines (e.g. Google, Bing)? **MC**
2. How often do you use Web search engines (e.g. Google, Bing) for learning purposes? **MC**
3. If you use search engines for learning, how useful do you find the experience? **MC**
4. If you could request a feature to make search as learning a better experience, what would you ask for? **F**

Current Usage of Search Engines. The results indicate overwhelming use of Web search engines in general with only 2/80 participants reporting less than daily frequency of usage. Furthermore, 65% of participants reported usage on an hourly or every few hours basis (exact breakdown in Figure 10.6). This is consistent with past trends of increasing general search engine adoption (Purcell et al., 2018) as well as specifically strong adoption and use by students (Niu et al., 2018; Salehi et al., 2018). Unlike some prior studies that have

surveyed participants about search engine usage, our findings also show the finer granularity of frequency of usage up to the hourly level. These findings indicate very frequent use of Web search engines in general.

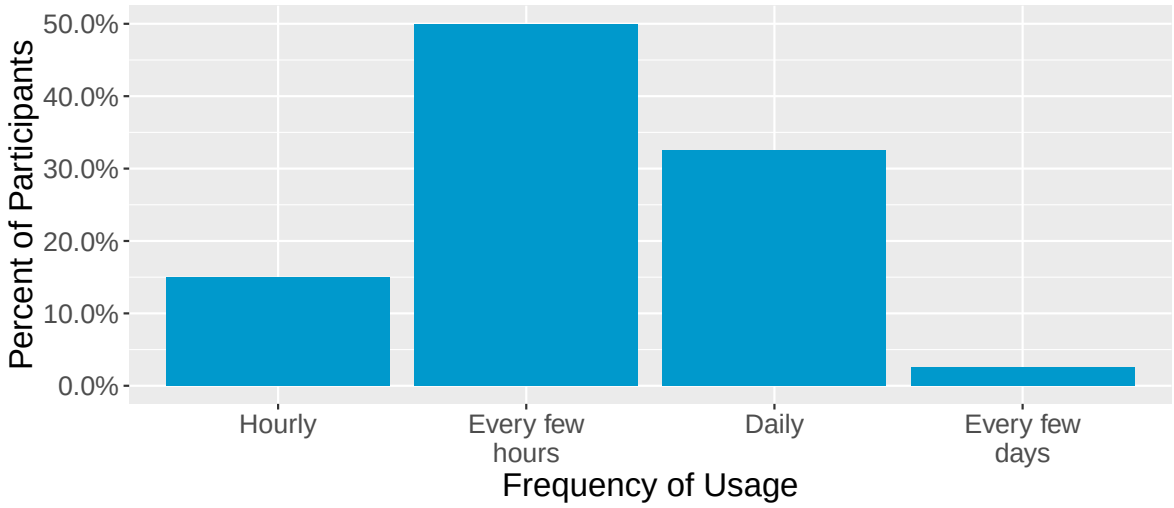

Figure 10.6: General search engine usage frequency. Almost everyone uses search engines on at least daily basis.

Use of Search Engines for Learning. We further investigated how often participants specifically use Web search engines for learning purposes. An overwhelming number of participants (85%) reported using search for learning at least on a daily basis with about 34% reporting usage on a hourly or every few hours basis (exact breakdown in Figure 10.7). This, too is consistent with past studies investigating student participants' use of search engines for learning (Abualsaud, 2017; Niu et al., 2018; Salehi et al., 2018).

Usefulness of Search as Learning. Finally, we investigated how useful participants reported search as learning has been for them. There were largely positive experiences in using search engines for learning with 91% of participants reporting search results were either good enough to use again for learning (56%) or search results almost perfectly helped them learn (35%). Overall, these results strongly indicate that existing search engines already

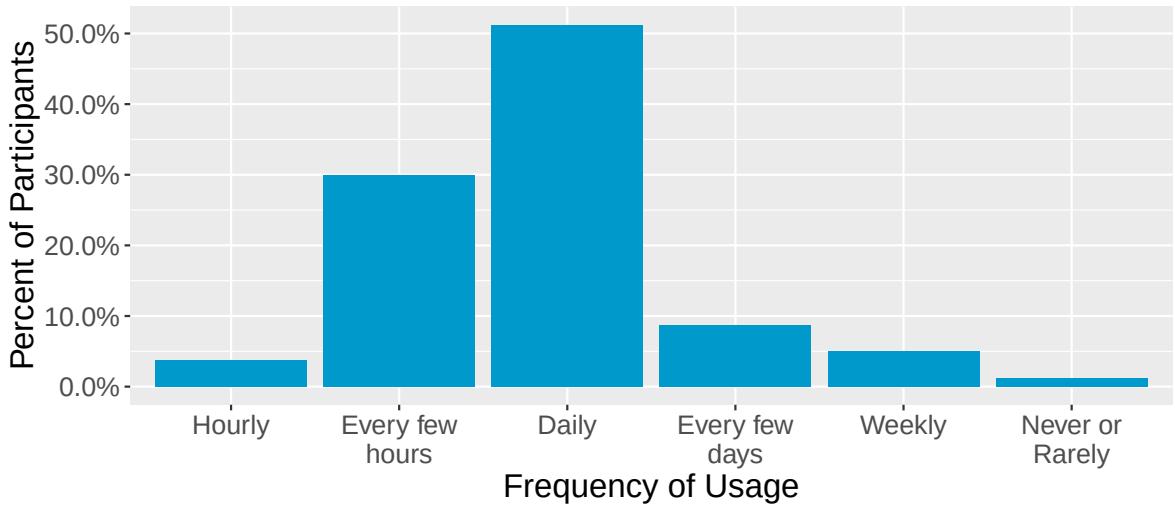

Figure 10.7: Frequency of using search engines for learning purposes. Overwhelming majority use search engines for learning on at minimum a daily basis.

provide good support for learning intents. At the same time, there is plenty of room for improvement as nearly 65% of participants did not rate their experience at the highest rating of "Very Useful: Search results almost perfectly help me learn." (exact breakdown in Figure 10.8). This finding further highlights the importance of developing models or interventions that improve the search as learning experience.

10.8 Discussion

A summary of our study findings is shown in Table 10.3. In addressing **RQ1**, we did find evidence that the interactive conditions yielded superior long-term grades for low-knowledge participants. In this analysis we also found that the beneficial value of adjunct questions is quite sensitive to the user's prior knowledge. In particular, high-knowledge participants found the *opposite* results: worse long-term results when using interactive conditions. This suggests that there is a value to using adjunct questions but the target audience should be

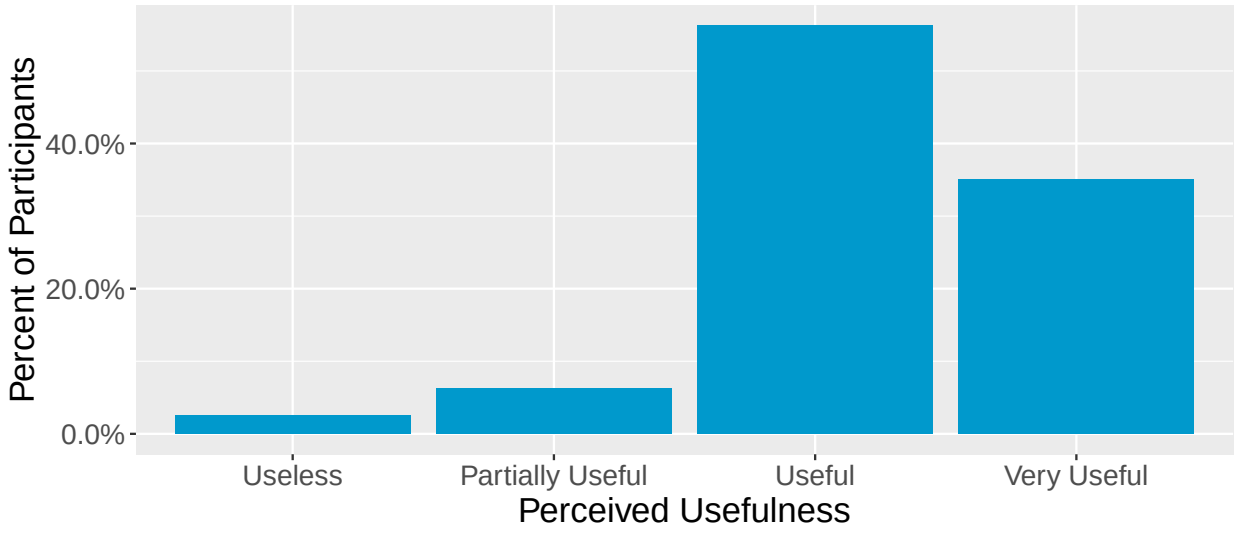

Figure 10.8: Perceived usefulness of search engine results when searching for learning purposes. Participants expressed strongly positive perceived usefulness though 65% did not rate quality at highest level.

relatively new to the subject. It is possible that high-knowledge participants were familiar enough with the topic that the adjunct questions were less of a learning opportunity and more of a distraction.

In addressing **RQ2**, we found that $\mathbf{Q_{Auto}}$ performed comparably (and to some extent even better) to $\mathbf{Q_{Human}}$ suggesting a promising potential use of auto-generated questions for applying the adjunct questions effect at scale. It remains an area of future work to investigate the quality of questions generated using our AQG system in different article contexts.

In addressing **RQ3**, we found $\mathbf{Q^*_{Human}}$ yielded significantly better long-term grades for New questions compared to $\mathbf{Q_{None}}$. However, it is unclear if this was due to the use of interactive and synthesis questions or due to the fact that participants in $\mathbf{Q^*_{Human}}$ spent substantially more time on the task than $\mathbf{Q_{None}}$ participants.

In addressing **RQ4**, we found evidence that participants did show significantly better long-term results when asked at least one focus-reading question as opposed to those who

got only skim-reading questions. This finding highlights the importance of asking questions personalized to content participants did and did not pay attention to, something we achieved through real-time gaze tracking.

In addressing **RQ5**, we found strong evidence that a measure of gaze fixations *normalized number of fixations* was substantially higher when participants answered post and delayed-test questions correctly. This was particularly true for the reading regressions and skimming types of fixations. However, this was largely limited to low-knowledge learners. High-knowledge learners showed almost no significant differences in almost any type of NNF both in short- and long-term. It is possible that HK learners were able to engage in more complex learning patterns that were not adequately captured by the three reading states that we investigated. This has important implications for using gaze behavior as an indicator of how much people are actually learning and can be useful as an estimate of long-term knowledge.

In our experiment implementation, there was a potential concern that the gaze tracking software's calibration may have needed re-calibration, especially after the half-time five-minute break. There were a few participants who had technical difficulties where the gaze tracking was not properly working and these data points were removed from analysis. Nevertheless, to isolate potentially erroneous results, we restricted the analysis in this paper to only the first topic a participant saw, which was presented almost immediately after the two rounds of initial calibration succeeded.

Overall, for high knowledge learners, there is limited benefit to introducing adjunct questions and in some cases detrimental effects. Thus we suggest not using adjunct questions for high-knowledge participants. Participant knowledge can be estimated through a pre-reading test or implicitly (e.g. using vocabulary used for a search query to estimate a user's

knowledge of topic).

For low knowledge learners, higher learning performance is seen in both short-term and long-term. In the long term these effects are significant and extend to both the Base questions (primed questions) and generalization (new questions never seen during pre-test or reading) and is maintained over time. Thus, we recommend the use of adjunct questions for low-knowledge learners.

10.9 Limitations/Future Work

In this work, we investigated multiple aspects of how the adjunct questions effect could be applied at scale. That being said, there are several limitations to the present study that could be addressed in future studies. Firstly, our study uses an automatic question generation model that was trained on the same corpus as the human-curated questions (SQuAD). It is possible there may be some confounds introduced here which may affect the applicability of our results to a more general setting. For example, if our model did comparably well to human-curated questions, it might be influenced by the fact that both sources are the same and so the AQG model is just mimicking the human-curated ones on this source. However, it is possible that if the same pre-trained AQG model were to be applied on a non-Wikipedia text, it might render worse quality questions or questions that might not be helpful for learning. It is for future work to investigate this.

Regardless of this concern, it is also an open question as to why the AQG model outperformed human-curated questions from the same source. While we briefly analyzed this earlier in the results, this warrants deeper investigation. We may be able to qualitatively tease out the reasoning through user studies where users annotate each question based on

factors such as detail, specificity and difficulty. If there are linguistic aspects such as additional details that influence their effectiveness, we could also apply heuristic linguistic filters on the question's parse tree to control the amount of detail and determine how this influences users' learning outcomes.

We also note that in this study we only looked at measures of fixations when analyzing gaze data and patterns. However, there are other gaze signals that could have been used such as saccades (Poole and Ball, 2005), average LADE and perceptual span (Mao et al., 2018). While not covered in this study, the use of such signals in conjunction with measures of fixation would likely yield a richer representation of user learning modeling.

There is also an interesting question regarding how user learning outcomes and reading behaviors differ when given a factoid vs synthesis question when faceted by prior knowledge. Specifically, it would be interesting to investigate what quantitative and qualitative differences emerge in LK vs HK learners when they are given a factoid vs synthesis question. While not addressed in this study, we expect there may be certain differences in reading strategies especially for synthesis questions between those who already have some knowledge of the topic versus those who are novices.

Finally, in this paper we analyzed how NNFs could separate between low and high learning outcomes but we didn't investigate how it could be used to classify or predict *prior* knowledge state. Such modeling would be of significant value if deployed in a scalable setting where either generating or grading pre-reading tests is infeasible or impractical.

10.10 Contributions

In this study we investigated the adjunct questions effect in two novel scenarios: (1) where the questions are determined in real-time based on live gaze-tracking; (2) where the questions are

generated through an automatic question generation (AQG) API versus the more traditional manual methods. We found evidence supporting earlier findings on the learning benefits of adjunct questions, though limited to novice learners. We further found evidence that AQG performed comparably - and in some cases, better - to human-curated questions. These results have very promising potential for applying the benefits of adjunct questions effect to large-scale applications such as embedding questions directly into arbitrary Web pages, encyclopedia entries or digital textbooks. We also showed that gaze tracking signals like NNF can be predictive of both short- and long-term learning outcomes suggesting a promising use of gaze tracking for estimating how much a learner will remember even after a one-week time delay.

Chapter 11

Future Work

In this dissertation, I have described studies I have worked on towards understanding how people learn with Web resources and developing models that support such goals. The work presented in this dissertation lays the foundation for multiple directions of potential future work. There are also other directions towards the general goal of supporting learning in search that are open to future work. In this chapter, I describe additional potential studies that would further support the overarching goal of supporting scalable search as learning.

11.1 High-level Future Directions

Towards supporting idealized learning objectives in a Web search context, there are multiple additional areas of research that would be important. I will elaborate on some specific directions that would be valuable for a production environment deployment.

11.1.1 Modeling Prerequisites Dependencies

The prerequisites dependencies of a subtopic S are the set of other subtopics that a person should have sufficient knowledge of to be able to learn S (Vuong, Nixon, and Towle, 2011). For example, Algebra 1 could be considered a prerequisite for Algebra 2. Our models currently do not factor in prerequisites dependencies when selecting a set of documents.

However, especially for learning goals, this is very important: If the first few documents we provide cover content that assumes certain prerequisite knowledge that has not been covered, it is highly unlikely the learner will benefit from those documents. Conversely, if the documents are ordered from those that have the least prerequisite dependencies to those that have the most, this could give the learner a better chance of acquiring more knowledge.

11.1.2 Detailed Personalization

In our studies we incorporated personalization in both our search retrieval framework (in the form of prior knowledge) and in our gaze tracking model (in the form of user-specific real-time gaze history). However, there are multiple other dimensions of the search as learning experience that could also benefit from personalization. For example, in our search framework, our difficulty-weighted keyword density objective assumes all readers will benefit from easier language. However, we know from prior work (Collins-Thompson et al., 2011; Tang et al., 2015) that this isn't necessarily true and that different readers have different readability comfort levels. Furthermore, user history can indicate user preferences for other features like content length, text-to-image ratio, preferred language, etc. all of which could be personalized.

11.1.3 Modeling Learning in Multi-Query Sessions

Our models currently demonstrate strong performance on single-query use cases but do not explicitly account for the additional concerns of multi-query or multi-session use cases which may be more probable in organic search as learning settings. For example, multi-session contexts over time may introduce forgetting effects (Murre and Dros, 2015) where content the user learned earlier may need to be reinforced based on factors like time lapse and

density of additional content exposed to the user in between. Furthermore, even within a single session, multiple queries and the sequence of those queries can provide rich signals as to what subtopics the learner needs help with and what queries seem to be leading to non-useful results (Hassan, White, Dumais, and Wang, 2014; Raman, Bennett, and Collins-Thompson, 2014). These signal can also help support more accurate knowledge tracing.

11.1.4 Feedback Mechanisms

Incorporating feedback in the system would introduce multiple positive benefits. In our gaze tracking study we had learners answer questions but didn't give any feedback. Simply getting feedback of whether the user answered questions correctly or not could help them better understand their own knowledge state and how much they have understood the topic. Furthermore, by providing corrections when the user answers a question incorrectly this could help resolve confusions or misunderstandings early on. However, providing accurate assessment feedback at scale for questions that may be open-ended is a non-trivial task and an ongoing direction of research. We leave it to future work to investigate automatic answer grading models that can provide a reasonably strong level of grading accuracy.

11.1.5 Detailed Gaze Tracking Analysis and Modeling

In our study, we used the Normalized Number of Fixations (NNF) measure to model reading behavior on different paragraphs. However, we used a plain text document which may not be representative of arbitrary documents on the Web. It would be valuable to not only model reading behavior on text but also model the value of supplementary materials such as images, videos, animations based on gaze patterns over these. Furthermore, there would be value in using gaze behaviors to model affective states during learning such as boredom

and curiosity (Jaques, Conati, Harley, and Azevedo, 2014). Such emotion tracking could be useful in knowing what types of documents, content style, etc. likely cause beneficial affective states and how this translates to better to learning outcomes.

11.1.6 Identifying Patterns - Collaborative Filtering

Our studies have focused on individual users using Web documents to learn. In a production system that has a sufficiently large number of users, there would be additional benefit in collaborative filtering (Resnick, Iacovou, Suchak, Bergstrom, and Riedl, 1994). This could be especially beneficial for "cold start" situations (new users to the system for whom there is minimal prior history). In such cases, we could use patterns and preferences observed for similar users and apply this to the new user. Furthermore, such an approach might help identify different clusters of learners. In scenarios where it is challenging to develop pedagogical tools tailored to every individual learner, clustering learners could allow a feasible approach to develop appropriate tools for certain types of learners as opposed to a one size fits all approach.

11.1.7 Query Intent Classifier

Thus far, our studies have largely operated on the assumption that users are indeed searching and reading Web documents for learning goals. For the context of our studies that was a valid assumption but this does not necessarily hold in an organic Web search context. While the intent of our model was to select documents that help with learning, it is possible that such a selection might also be beneficial from the standpoint of user satisfaction, content relevance or other metrics as well. It is for future work to evaluate the potential usefulness of our approach for other metrics of success. If our selection criteria is mostly useful for

learning-oriented objectives, we could selectively apply our model in a search engine using query intent classification. That is, the search results would be selected using our proposed approaches when the user's intent is likely learning-oriented and the system would default to its existing selection criteria otherwise.

11.1.8 Modeling other Types of Learning

Most of the studies we conducted focused on the lowest-complexity form of learning - the 'Remember' level which only requires being able to remember certain facts (in our case, definitions of technical terms). Our earliest study (Collins-Thompson et al., 2016) did go deeper in addressing all six types of learning complexities though that study didn't specifically develop an algorithm to support the different learning needs of each level of complexity. Some prior studies have investigated multiple complexities of learning tasks in Web search (Jansen et al., 2009; Kalyani and Gadiraju, 2019; Wu et al., 2012) as we discussed in Chapter 3.4. However, these studies primarily focus on *understanding* how tasks of varying cognitive complexity affect search behaviors, patterns and task difficulty whereas our focus is on *optimizing* selection of documents to maximize learning outcomes. I believe that having a better understanding of the nuances in search activity based on learning task complexity combined with the work we have done in optimizing learning outcomes lays a strong foundation for future work to expand towards optimal models for multiple complexities of learning tasks.

11.1.9 Investigating other Facets of Learning

A central focus in this dissertation has been on achieving measurable improvements in learning outcomes in the direct form of topic assessments. However, there are other aspects of the learning process that are important as well that warrant further investigation. One such

direction is user *perceptions* of learning usefulness of a document. If a user decides early on that the document is not likely to help them learn, it is unlikely they will achieve much learning benefit from that document. Inital impressions of learning usefulness can be very critical as humans tend to make very quick judgments about a website's general quality (Lindgaard, Fernandes, Dudek, and Brown, 2006).

Specifically for learning, there is strong evidence that impressions of aesthetics, usability and content structure can influence the learning experience (Rieh, Kim, and Markey, 2012; Zhang and Quintana, 2012). While these studies did investigate how various dimensions of learning outcomes were affected by differences in interface, navigational difficulty and content readability (Ng and Gunstone, 2002), there is still no clear analysis of what precise features of Web documents influence how a user makes an initial judgment of its learning potential. In particular, I believe an important direction of future work is to develop trained models using document features that classify the likely perceived usefulness of a website for learning. Such a model could integrate well with our existing model as a filtering step to avoid documents that are not likely to be perceived positively by a user.

11.1.10 Model-based vs Model-free Algorithms

In the studies presented in this thesis, we focused heavily on model-based assumptions of how people learn, particularly Item Response Theory (IRT). However, there are other methods that have been investigated for supporting learning that don't make as strong assumptions of how people learn and are instead general frameworks that can be adapted to learning. Prior work by Clément (2018) used a multi-armed bandit approach to model the sequence of activities a student will encounter as part of an intelligent tutoring system. While our application is somewhat different, it is also possible to model learning on the Web as a multi-

armed bandit problem involving a set of possible documents to select and the context-specific expected rewards from each. An advantage of a model like IRT is that it is easy to interpret and use though this isn't always the desired goal. In cases where it is more important for the student to accomplish their learning goals and have some idea of what did and didn't help them learn, using a model-free algorithm - if it can produce better results - may be a more promising direction. It is also an interesting and open question as to whether some sort of hybrid of the two approaches (model-based and model-free) could be developed that may address the shortcomings of each.

Furthermore, the model-based approach we used, Item Response Theory, is only one such model and has its own limitations. Another popular model is Bayesian Knowledge Tracing (BKT) (Corbett and Anderson, 1994) which explicitly factors in probabilities of students guessing correct answers as well as forgetting previously learned answers. Other models like the Half-Life Regression (HLR) model have specifically focused on vocabulary acquisition and incorporate aspects such as recency of assessment, expected degree of forgetting and user-specific memory capabilities (Settles and Meeder, 2016). As such, it is for future work to investigate how different model-based or model-free algorithms might improve on results we have already seen using only IRT.

11.2 Example Use Case

To illustrate how the above directions of future work could integrate into a holistic experience, I give an example use case of how such a system might work. Let's say a student is tasked to learn about the topic "Igneous rocks". The student begins by entering a search query such as "What are igneous rocks?". The system classifies this as a learning intent and the above modules now activate. The system will first look at historical signals for this user in terms

of other learning-intent search queries they have issued and the types of documents they visited, estimates of their satisfaction, boredom, etc. on those websites and their feedback assessment scores during those search sessions. This will give a good understanding of what types of documents this particular user is more likely to engage with and learn from.

Next, using our trained prediction models as well as modules mentioned above, the system will select a large set of topic-specific candidate documents and rank them. This ranking will include our own regression model scores as well as a separate ranker for prerequisites dependencies. Documents whose perceived learning usefulness is classified as very weak will be removed from the rankings. As the student reads the documents, automatic question generation and gaze tracking will be applied to generate questions for the learner to answer. These questions will be factoid questions to enable easier auto-assessment for giving real-time feedback. Based on the feedback results, the system can perform a re-ranking of the remaining documents when returning to the SERP to help resolve potential confusions or misunderstandings. For example, if the student failed to answer a question of "What differentiates Igneous rocks from Metamorphic and Sedimentary rocks?", the SERP's next document could be one that specifically focuses on these differences.

Gaze tracking signals in this whole process can also give an indication of what content the student has paid closer attention to and what they have skimmed. This can allow the system to also put more emphasis on re-ranking future documents to potentially put emphasis on content the student has been skimming.

Chapter 12

Conclusion

In this dissertation, I have discussed several studies and proposed a new framework towards accomplishing an overarching goal: the development, application and evaluation of scalable learning-oriented information retrieval models. The primary focus of the studies I have completed thus far (Chapter 4) was on developing retrieval models that could support learning-oriented information retrieval. I showed that not only was this accomplished through a topic modeling approach in the vocabulary domain but that a data-driven modeling approach could also be used for predicting multiple measures of learning outcomes. The results for data-driven modeling were able to show strong generalization in two other independent studies, paving the way for future models and search systems to use and learn from the results presented in these works to support learning intents in search.

Core Research Questions. At the start of the dissertation, I described the following high-level research questions I would address with the studies presented here. In this chapter, I will describe how these particular questions were addressed:

RQ1: Can we apply a model of domain-specific user knowledge state that updates based on what Web documents they read? Does such a model improve learning outcomes?

Results: We used the sigmoidal function from Item Reponse Theory (IRT) to model how people learn. We implemented this in a vocabulary learning context

where we made assumptions of how people learn as a function of how many keywords they are exposed to. These counts could be directly computed from any Web document and could thus model an estimate of expected learning. For our study, we kept several useful parameters of our cognitive model fixed, including individual learning rate, subtopic difficulty, and subtopic importance. Incorporating and tweaking these components of the model could potentially provide an even more personalized and effective learning experience for the user (Section 5.2 and Chapter 6).

RQ2: Can we develop an information retrieval framework that explicitly uses estimated user knowledge gain as its optimization objective? Can such a model outperform a commercial baseline?

Results: Building on the model described above, we developed a novel retrieval framework that incorporated a cognitive model in optimizing the retrieval objective. Specifically, our framework determined the optimal number of topic aspects (in our case, vocabulary keywords) the user needs to be exposed to. This step enforced upper bounds on how many documents would be necessary to retrieve. The retrieval criteria was a novel metric of difficulty-weighted keyword density which rewarded concise, readable and keyword-dense documents (Chapter 6).

RQ3: Are there document, user or document set features that are good predictors of knowledge state and knowledge gain in a Web documents context?

Results: We fit regression models to our user study data and found a variety of signals that were good indicators of multiple measures of learning outcomes. This included features that had independently been investigated in prior work (like use

of relevant images, decrease use of links, etc.). We also found an interesting result where set-level features (micro-averaging) often showed opposite sign coefficients to their document-level (macro-averaging) counterparts. As such, learning outcomes may be affected by set-level features which suggests an importance in performing set-level optimization for learning applications (Chapters 8 and 9).

RQ4: Can automatic question generation be used to scale the adjunct questions effect to support scalable active learning in Web documents? (In this dissertation, we refer to *active learning* in the pedagogical context not the machine learning context.)

Results: We conducted an experiment to compare how well people learn when using human-curated vs auto-generated questions (AQG) on the same content corpus and topic. We found strong evidence that AQG questions provide not only comparable but sometimes superior learning outcomes in the long-term. This has significant implications for the potential of facilitating active learning at scale for arbitrary Web text documents (Chapter 10).

RQ5: Are there differences in learning outcomes in the Web context when considering short- vs long-term assessment?

Results: In two of the studies we discussed in this thesis, we showed that long-term learning outcomes show significantly different results than what we find in the short-term. In Chapter 7 we showed that in the long-term, the benefit of personalization for easier terms mostly vanishes while the benefit of harder terms stays strong. In Chapter 10 we found showed that the benefits of the adjunct questions effect only showed significant differences in the long-term with no significant differences in the short-term. These results highlight the importance

of considering both short- and long-term learning outcomes when modeling and evaluating learning-oriented algorithms and frameworks (Chapters 7 and 10).

Main contributions. In this dissertation I demonstrate the importance of choosing a cognition-aware user representation when selecting Web documents for learning goals. Prior research in Web search optimization has explored many directions of optimizing towards different success metrics (e.g. relevance, user satisfaction, comprehensibility). However, the work in this dissertation introduces for the first time a Web search framework that explicitly incorporates cognitive models into the retrieval objective to optimize a metric of expected knowledge gain.

In the domain of applied algorithms for education, we further demonstrate the importance of not only evaluating short-term outcomes but also long-term outcomes. In this dissertation, we evaluated short- and long-term results in both our search framework study as well as our gaze tracking study. In both cases we observed how different metrics of learning varied substantially when considering the short- vs long-term. This suggests a crucial importance in evaluating both types of assessment periods when evaluating the usefulness and value of any novel pedagogical tool, even beyond the types investigated in this work.

Implications and Future Work. The studies conducted in this dissertation provide a solid foundation for understanding how multiple forms of learning can be supported in a Web search context. We know from prior work that a significant fraction of information seeking tasks start with or at some point involve the use of Web search engines. Implementation of the models introduced in this dissertation in large-scale Web search systems could yield substantial benefits in facilitating self-paced and self-directed learning at scale. We introduced models in this work that were designed for scalable deployment by using features that could be computed automatically and efficiently at scale. We further showed that this model could

show strong predictive power on independent datasets of learning through Web documents, suggesting stronger generalizability. While this work has thus far only been tested at relatively small scale (order of hundreds of participants), it remains for future work to investigate the effect on learning outcomes when deployed in a large-scale organic search environment. To facilitate this, such an implementation may be paired with a query intent classifier to only provide the proposed re-ranked results when users issue queries of educational intent.

We further demonstrated a promising direction for supporting a form of active learning at scale. The results from this study show promising potential for applied automatic question generation for creating adjunct questions for arbitrary text articles as opposed to the previous methods of manually constructing such questions. This could have significant implications for how interactive learning benefits could be scaled to arbitrary expository documents.

The work presented in this dissertation collectively investigated multiple aspects of learning outcomes, short- and long-term impacts, passive vs interactive experiences and transferability of learned models to other datasets. Some of these studies resulted in trained models and classifiers that form a solid foundation for future work to build on. These studies and the associated results offer valuable insight and tools for practitioners to enhance the quality of self-paced and self-directed search as learning tasks which, if past findings remain consistent, remains on a strong and rising trend.

References

Mustafa Abualsaud. 2017. *Learning Factors and Determining Document-level Satisfaction In Search-as-Learning*. Master's thesis. University of Waterloo.

Deanne M Adams, Bruce M McLaren, Richard E Mayer, George Goguadze, and Seiji Isotani. 2013. Erroneous Examples as Desirable Difficulty. In *Artificial Intelligence in Education*. Springer Berlin Heidelberg, 803–806.

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 5–14.

Kouichi Akamatsu, Adam Jatowt, and Katsumi Tanaka. 2015. Towards Solving Comprehensibility-Relevance Trade-off in Information Retrieval. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 1. IEEE, 1–8.

LW Anderson, DR Krathwohl, W Airiasian, KA Cruikshank, RE Mayer, PR Pintrich, et al. 2001. A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational outcomes: Complete edition. *NY: Longman* (2001).

Peter Bailey, Liwei Chen, Scott Grosenick, Li Jiang, Yan Li, Paul Reinholdtsen, Charles Salada, Haidong Wang, and Sandy Wong. 2012. User task understanding: a web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Kanagawa, Japan*.

Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.

Nilavra Bhattacharya and Jacek Gwizdka. 2018. Relating eye-tracking measures with changes in knowledge on search tasks. *arXiv preprint arXiv:1805.02399* (2018).

Dania Bilal. 2000. Children's use of the Yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for information Science* 51, 7 (2000), 646–665.

Jacob Lowell Bishop and Matthew A Verleger. 2013. The flipped classroom: A survey of the research. In *ASEE National Conference Proceedings, Atlanta, GA*.

Elizabeth L Bjork and Robert A Bjork. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society* (2011), 56–64.

Elizabeth Ligon Bjork, Jeri L Little, and Benjamin C Storm. 2014. Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition* 3, 3 (2014), 165–170.

Robert A Bjork. 1994. Institutional impediments to effective training. *Learning, remembering, believing: Enhancing human performance* (1994), 295–306.

Robert A Bjork. 1994a. Memory and metamemory considerations in the training of human beings. *Metacognition: Knowing about knowing* (1994a), 185–205.

Benjamin S Bloom. 1956. *Taxonomy of educational objectives: The classification of educational goals*. New York, Longmans, Green.

Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher* 13, 6 (1984), 4–16.

Tina Penick Brock and Scott R Smith. 2007. Using digital videos displayed on personal digital assistants (PDAs) to enhance patient education in clinical settings. *International journal of medical informatics* 76, 11 (2007), 829–835.

Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10.

Bertram C Brookes. 1980. The foundations of information science Part I. Philosophical aspects. *Journal of Information Science* 2, 3-4 (1980), 125–133.

Jan Brophy and David Bawden. 2005. Is Google enough? Comparison of an internet search engine with academic library resources. In *Aslib Proceedings*, Vol. 57. Emerald Group Publishing Limited, 498–512.

Peter Brusilovsky, Steven Ritter, and Elmar Schwarz. 1997. Distributed intelligent tutoring on the Web. *Artificial Intelligence in Education: Knowledge and Media in Learning Systems. IOS, Amsterdam* 482 (1997), 489.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46, 3 (2014), 904–911.

Sahan Bulathwela, Emine Yilmaz, and John Shawe-Taylor. 2019. Towards Automatic, Scalable Quality Assurance in Open Education. (2019).

Aimee A Callender and Mark A McDaniel. 2007. The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology* 99, 2 (2007), 339.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 335–336.

Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information

retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 659–666.

Charles LA Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. *An effectiveness measure for ambiguous and underspecified queries*. Springer Berlin Heidelberg. 188–199 pages.

Benjamin Clément. 2018. *Adaptive Personalization of Pedagogical Sequences using Machine Learning*. Ph.D. Dissertation. Bordeaux.

Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. 2013. Inferring user knowledge level from eye movement patterns. *Information Processing & Management* 49, 5 (2013), 1075–1091.

Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. 2011. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 403–412.

Kevyn Collins-Thompson and Jamie Callan. 2004. Information Retrieval for Language Tutoring: An Overview of the REAP Project. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. ACM, New York, NY, USA, 544–545.

Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as learning (dagstuhl seminar 17092). In *Dagstuhl reports*, Vol. 7. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. 2016. Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. ACM, New York, NY, USA, 163–172.

Leana Copeland and Tom Gedeon. 2013. Measuring reading comprehension using eye movements. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*. IEEE, 791–796.

Leana Copeland and Tom Gedeon. 2014. What are you reading most: attention in eLearning. *Procedia Computer Science* 39 (2014), 67–74.

Leana Copeland, Tom Gedeon, and Sabrina Caldwell. 2014a. Framework for Dynamic Text Presentation in eLearning. *Procedia Computer Science* 39 (2014), 150–153.

Leana Copeland, Tom Gedeon, and Sumudu Mendis. 2014b. Fuzzy Output Error as the Performance Function for Training Artificial Neural Networks to Predict Reading Comprehension from Eye Gaze. In *International Conference on Neural Information Processing*. Springer, 586–593.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

Diana I Cordova and Mark R Lepper. 1996. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of*

educational psychology 88, 4 (1996), 715.

Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media* 29, 3 (2004), 241–250.

Van Dang and Bruce W Croft. 2013. Term level search result diversification. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 603–612.

Beth Davey and Susan McBride. 1986. Effects of question-generation training on reading comprehension. *Journal of Educational Psychology* 78, 4 (1986), 256.

Antonella De Angeli, Lynne Coventry, Graham Johnson, and Karen Renaud. 2005. Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies* 63, 1 (2005), 128–152.

Cathy De Rosa. 2006. *College students' perceptions of libraries and information resources: A report to the OCLC membership*. OCLC.

Brenda Dervin. 1983. An overview of sense-making research: Concepts, methods, and results to date. In *International Communication Association Annual Meeting, Dallas, TX*. `http://faculty.washington.edu/wpratt/MEBI598/Methods/An%20Overview%20of%20Sense-Making%20Research%201983a.htm`

Diana DeStefano and Jo-Anne LeFevre. 2007. Cognitive load in hypertext reading: A review. *Computers in Human Behavior* 23, 3 (2007), 1616–1641.

John L Dobson. 2011. Effect of selected "desirable difficulty" learning strategies on the retention of physiology information. *Advances in physiology education* 35, 4 (2011), 378–383.

Michele M Dornisch. 2012. Adjunct questions: Effects on learning. *Encyclopedia of the sciences of learning* (2012), 128–129.

Michele M Dornisch and Rayne A Sperling. 2006. Facilitating learning from technology-enhanced text: Effects of prompted elaborative interrogation. *The Journal of Educational Research* 99, 3 (2006), 156–166.

Allison Druin, Elizabeth Foss, Leshell Hatley, Evan Golub, Mona Leigh Guha, Jerry Fails, and Hilary Hutchinson. 2009. How children search the internet with keyword interfaces. In *Proceedings of the 8th International conference on interaction design and children*. ACM, 89–96.

Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. 2010. An Analysis of Queries Intended to Search Information for Children. In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10)*. ACM, New York, NY, USA, 235–244.

Geoffrey B Duggan and Stephen J Payne. 2008. Knowledge in the head and on the web: Using topic expertise to aid search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 39–48.

William H Dutton and Ellen Johanna Helsper. 2007. Oxford internet survey 2007 report:

The internet in Britain. *Available at SSRN 1327033* (2007).

Yuka Egusa, Hitomi Saito, Masao Takaku, Hitoshi Terai, Makiko Miwa, and Noriko Kando. 2010. Using a concept map to evaluate exploratory search. In *Proceedings of the third symposium on Information interaction in context*. ACM, 175–184.

Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 223–232.

Michael A Eskenazi and Jocelyn R Folk. 2017. Regressions during reading: The cost depends on the cause. *Psychonomic bulletin & review* 24, 4 (2017), 1211–1216.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel 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.

Nigel Ford, David Miller, and Nicola Moss. 2003. Web search strategies and approaches to studying. *Journal of the American Society for Information Science and Technology* 54, 6 (2003), 473–489.

Susannah Fox. 2014. The social life of health information. *Pew research center, Washington, DC, Pew Internet & American Life Project* (2014). `http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/`

Susannah Fox and Sydney Jones. 2009. The social life of health information. *Pew research center, Washington, DC, Pew Internet & American Life Project* (2009). `http://www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx`

Marita Franzke, Eileen Kintsch, Donna Caccamise, Nina Johnson, and Scott Dooley. 2005. Summary Street®: Computer support for comprehension and writing. *Journal of Educational Computing Research* 33, 1 (2005), 53–80.

Luanne Freund, Rick Kopak, and Heather O'Brien. 2016. The effects of textual environment on reading comprehension: Implications for searching as learning. *Journal of Information Science* 42, 1 (2016), 79–93.

Gwen A Frishkoff, Kevyn Collins-Thompson, Leslie Hodges, and Scott Crossley. 2016. Accuracy feedback improves word learning from context: evidence from a meaning-generation task. *Reading and Writing* 29, 4 (2016), 609–632.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48, 4 (2005), 612–618.

Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 478–479.

Jillian R Griffiths and Peter Brophy. 2005. Student searching behavior and the web: use of academic resources and Google. (2005).

Philip J Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 41–50.

Qi Guo and Eugene Agichtein. 2010. Towards Predicting Web Searcher Gaze Position from Mouse Movements. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 3601–3606. https://doi.org/10.1145/1753846.1754025

Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. 2014. Struggling or exploring?: disambiguating long search sessions. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 53–62.

Jannica Heinström. 2006. Broad exploration or precise specificity: Two basic information seeking patterns among students. *Journal of the American Society for Information Science and Technology* 57, 11 (2006), 1440–1450.

Michael Henderson, Neil Selwyn, Glenn Finger, and Rachel Aston. 2015. Students' everyday engagement with digital technology in university: exploring patterns of use and 'usefulness'. *Journal of Higher Education Policy and Management* 37, 3 (2015), 308–319.

Christoph Hölscher and Gerhard Strube. 2000. Web search behavior of Internet experts and newbies. *Computer networks* 33, 1 (2000), 337–346.

Jeff Huang, Ryen White, and Georg Buscher. 2012. User see, user point: gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1341–1350.

Yun Huang, Michael Yudelson, Shuguang Han, Daqing He, and Peter Brusilovsky. 2016. A Framework for Dynamic Knowledge Modeling in Textbook-Based Learning. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. ACM, 141–150.

Scott B Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction?. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 567–574.

Albrecht Werner Inhoff and Ralph Radach. 1998. Definition and computation of oculomotor measures in the study of cognitive processes. In *Eye guidance in reading and scene perception*. Elsevier, 29–53.

Bernard J Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management* 45, 6 (2009), 643–663.

Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2007. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 1149–1150.

Natasha Jaques, Cristina Conati, Jason M Harley, and Roger Azevedo. 2014. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International conference on intelligent tutoring systems*. Springer, 29–38.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 57–66.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data As Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, New York, NY, USA, 154–161. https://doi.org/10.1145/1076034.1076063

Terry Judd and Gregor Kennedy. 2010. A five-year study of on-campus Internet use by undergraduate biomedical students. *Computers & Education* 55, 4 (2010), 1564–1571.

Brian W Junker. 1999. Some statistical models and computational methods that may be useful for cognitively-relevant assessment. *Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved April* 2 (1999), 81.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

Rishita Kalyani and Ujwal Gadiraju. 2019. Understanding User Search Behavior Across Varying Cognitive Levels. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. ACM, 123–132.

Sabbir Ahmed Kazi. 2005. VocaTest: An intelligent tutoring system for vocabulary learning using the" mLearning" approach. (2005).

Aytürk Keleş, Rahim Ocak, Ali Keleş, and Aslan Gülcü. 2009. ZOSMAT: Web-based intelligent tutoring system for teaching–learning process. *Expert Systems with Applications* 36, 2 (2009), 1229–1239.

Jin Young Kim, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. 2012. Characterizing Web Content, User Interests, and Search Behavior by Reading Level and Topic. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 213–222.

Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. 1997. Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education* 8 (1997), 30–43.

Nate Kornell and Robert A Bjork. 2008. Learning concepts and categories is spacing the "enemy of induction". *Psychological science* 19, 6 (2008), 585–592.

David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into Practice* 41, 4 (2002), 212–218.

Carol C Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science* 42, 5 (1991), 361–371.

Carol C Kuhlthau, Jannica Heinström, and Ross J Todd. 2008. The 'information search

process' revisited: Is the model still useful. *Information Research* 13, 4 (2008), 13–4.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44, 4 (2012), 978–990.

Isabelle Lamoureux, Jamshid Beheshti, Charles Cole, Dhary Abuhimed, and Mohammed J AlGhamdi. 2013. Gender differences in inquiry-based learning at the middle school level. *Proceedings of the American Society for Information Science and Technology* 50, 1 (2013), 1–5.

Krittaya Leelawong and Gautam Biswas. 2008. Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education* 18, 3 (2008), 181–208.

Xin Li, Yiqun Liu, Rongjie Cai, and Shaoping Ma. 2017. Investigation of user search behavior while facing heterogeneous search services. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 161–170.

Gitte Lindgaard, Gary Fernandes, Cathy Dudek, and Judith Brown. 2006. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & information technology* 25, 2 (2006), 115–126.

Jeri L Little and Elizabeth Ligon Bjork. 2012. Pretesting with multiple-choice questions facilitates learning. In *Proceedings of the annual meeting of the cognitive science society*. 294–299.

Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. "Satisfaction with Failure" or "Unsatisfied Success": Investigating the Relationship between Search Success and User Satisfaction. In *Proceedings of the 2018 World Wide Web Conference (WWW'18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland*. 1533–1542.

Jakub Lokoč, Adam Blažek, and Tomáš Skopal. 2014. On Effective Known Item Video Search Using Feature Signatures. In *Proceedings of International Conference on Multimedia Retrieval (ICMR '14)*. ACM, New York, NY, USA, Article 524, 3 pages. `https://doi.org/10.1145/2578726.2582617`

Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How Does Domain Expertise Affect User's Search Interaction and Outcome in Exploratory Search? 36 (07 2018), 1–30.

Gary Marchionini. 1997. *Information seeking in electronic environments*. Number 9. Cambridge university press.

Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

Carolyn B Marks, Marleen J Doctorow, and Merlin C Wittrock. 1974. Word frequency and reading comprehensiony1. *The Journal of Educational Research* 67, 6 (1974), 259–262.

Richard E Mayer. 1997. Multimedia learning: Are we asking the right questions? *Educational Psychologist* 32, 1 (1997), 1–19.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of

word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

Jaap MJ Murre and Joeri Dros. 2015. Replication and analysis of Ebbinghaus' forgetting curve. *PloS one* 10, 7 (2015), e0120644.

NetDay. 2004. Voices and views of today's tech-savvy students: National report on NetDay speak up day for students 2003. (2004).

Delia Neuman. 2011. *Learning in information-rich environments: I-LEARN and the construction of knowledge in the 21st century.* Springer Science & Business Media.

Wan Ng and Richard Gunstone. 2002. Students' perceptions of the effectiveness of the World Wide Web as a research and teaching tool in science learning. *Research in Science Education* 32, 4 (2002), 489–510.

Xi Niu, Fakhri Abbas, Mary Lou Maher, and Kazjon Grace. 2018. Surprise Me If You Can: Serendipity in Health Information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 23, 12 pages.

Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. 2015. Struggling and success in web search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1551–1560.

Stellan Ohlsson. 1996. Learning from Performance Errors. *Psychological Review* 103, 2 (1996), 241–262.

Erol Ozcelik, Ismahan Arslan-Ari, and Kursat Cagiltay. 2010. Why does signaling enhance multimedia learning? Evidence from eye movements. *Computers in human behavior* 26, 1 (2010), 110–117.

Joao Palotti, Allan Hanbury, and Henning Müller. 2014. Exploiting health related features to infer user expertise in the medical domain. In *Web Search Click Data workshop at WSCM, New York City, NY, USA*.

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* 12, 3 (2007), 801–823.

Bing Pan, Helene A Hembrooke, Geri K Gay, Laura A Granka, Matthew K Feusner, and Jill K Newman. 2004. The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*. ACM, 147–154.

Alexandra Papoutsaki, Aaron Gokaslan, James Tompkin, Yuze He, and Jeff Huang. 2018. The eye of the typer: a benchmark and analysis of gaze behavior during typing. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ACM, 16.

TD Paul. 2003. Guided independent reading: An examination of the reading practice database and the scientific research supporting guided independent reading as implemented in Reading Renaissance. *Madison, WI: Renaissance Learning. Retrieved September* 25 (2003), 2007.

Stephen T Peverly and Rhea Wood. 2001. The effects of adjunct questions and feedback on improving the reading comprehension skills of learning-disabled adolescents. *Contemporary Educational Psychology* 26, 1 (2001), 25–43.

Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

Peter Pirolli and Sanjay Kairam. 2013. A knowledge-tracing model of learning from a social tagging system. *User Modeling and User-Adapted Interaction* 23, 2-3 (2013), 139–168.

Alex Poole and Linden J Ball. 2005. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. (2005).

Kristen Purcell, Joanna Brenner, and Lee Rainie. 2012 (accessed May 22, 2018). Search Engine Use 2012. *Pew research center, Washington, D.C.* (2012 (accessed May 22, 2018)). `http://www.pewinternet.org/2012/03/09/search-engine-use-2012/`

Filip Radlinski and Susan Dumais. 2006. Improving Personalized Web Search Using Result Diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 691–692.

Lee Rainie and Paul Hitlin. 2009. The Internet at School. *Pew research center, Washington, DC, Pew Internet & American Life Project* (2009). `http://www.pewinternet.org/2005/08/02/the-internet-at-school/`

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv preprint arXiv:1806.03822* (2018).

Karthik Raman, Paul N. Bennett, and Kevyn Collins-Thompson. 2013. Toward Whole-session Relevance: Exploring Intrinsic Diversity in Web Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 463–472.

Karthik Raman, Paul N Bennett, and Kevyn Collins-Thompson. 2014. Understanding intrinsic diversity in web search: Improving whole-session relevance. *ACM Transactions on Information Systems (TOIS)* 32, 4 (2014), 20.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading* 10, 3 (2006), 241–255.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 175–186.

Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.

Soo Young Rieh, Yong-Mi Kim, and Karen Markey. 2012. Amount of invested mental effort

(AIME) in online searching. *Information Processing & Management* 48, 6 (2012), 1136–1150.

Stephen E Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* 33, 4 (1977), 294–304.

Kerry Rodden, Xin Fu, Anne Aula, and Ian Spiro. 2008. Eye-mouse Coordination Patterns on Web Search Results Pages. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems (CHI EA '08)*. ACM, New York, NY, USA, 2997–3002. `https://doi.org/10.1145/1358628.1358797`

Doug Rohrer, Robert F Dedrick, and Sandra Stershic. 2015. Interleaved practice improves mathematics learning. *Journal of Educational Psychology* 107, 3 (2015), 900–908.

Daniel E Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. ACM, New York, NY, USA, 13–19.

Sara Salehi, Jia Tina Du, and Helen Ashman. 2018. Use of Web search engines and personalisation in information searching for educational purposes. *Information Research: An International Electronic Journal* 23, 2 (2018), n2.

Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. 2003. Can relevance be inferred from eye movements in information retrieval. In *Proceedings of WSOM*, Vol. 3. 261–266.

Mark Sanderson and W Bruce Croft. 2012. The history of information retrieval research. *Proc. IEEE* 100, Special Centennial Issue (2012), 1444–1451.

Richard A Schmidt and Robert A Bjork. 1992. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science* 3, 4 (1992), 207–217.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1848–1858.

B Shneiderman and G Marchionini. 1988. Finding facts vs. Browsing Knowledge in Hypertext systems. *Proceedings of Computer* (1988), 70–80.

John L Sibert, Mehmet Gokturk, and Robert A Lavine. 2000. The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*. ACM, 101–107.

Norman J Slamecka and Peter Graf. 1978. The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory* 4, 6 (1978), 592–604.

Mark D. Smucker and Charles L.A. Clarke. 2012. Time-based Calibration of Effectiveness Measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 95–104.

Amanda Spink, Bernard J Jansen, Dietmar Wolfram, and Tefko Saracevic. 2002. From e-sex

to e-commerce: Web search changes. *Computer* 35, 3 (2002), 107–109.

Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P Lorence, Seda Ozmutlu, and H Cenk Ozmutlu. 2004. A study of medical and health queries to web search engines. *Health Information & Libraries Journal* 21, 1 (2004), 44–51.

Rohail Syed and Kevyn Collins-Thompson. 2016. Optimizing Search Results for Educational Goals: Incorporating Keyword Density as a Retrieval Objective. In *Second International Workshop on Search as Learning (SaL 2016)*. ACM. `http://ceur-ws.org/Vol-1647/SAL2016_paper_21.pdf`

Rohail Syed and Kevyn Collins-Thompson. 2017a. Optimizing search results for human learning goals. *Information Retrieval Journal* 20, 5 (2017), 506–523.

Rohail Syed and Kevyn Collins-Thompson. 2017b. Retrieval algorithms optimized for human learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 555–564.

Rohail Syed and Kevyn Collins-Thompson. 2018. Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 191–200. `https://doi.org/10.1145/3176349.3176397`

Rohail Syed, Kevyn Collins-Thompson, Paul N Bennett, Mengqiu Teng, Shane Williams, Wendy Tay, and Shamsi Iqbal. 2020. Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation. In *The World Wide Web Conference*. ACM. In press.

Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. 2012. To each his own: personalized content selection based on text comprehensibility. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 233–242.

Steven Tang, Hannah Gogel Elizabeth McBride, and Zachary A Pardos. 2015. Desirable Difficulty and Other Predictors of Effective Item Orderings. In *International Conference on Educational Data Mining*. International Educational Data Mining Society, 416–419.

Robert S Taylor. 1968. Question-negotiation and information seeking in libraries. *College & research libraries* 29, 3 (1968), 178–194.

KJ Topping and WL Sanders. 2000. Teacher effectiveness and computer assessment of reading relating value added and learning information system data. *School Effectiveness and School Improvement* 11, 3 (2000), 305–337.

Madeleine Udell and Stephen Boyd. 2013. Maximizing a sum of sigmoids. *Optimization and Engineering* (2013).

Geoffrey Underwood and John Everatt. 1992. The role of eye movements in reading: some limitations of the eye-mind assumption. *Advances in psychology* 88 (1992), 111–169.

Pertti Vakkari and Saila Huuskonen. 2012. Search effort degrades search output but improves task outcome. *Journal of the American Society for Information Science and Technology* 63, 4 (2012), 657–670.

Manisha Verma, Emine Yilmaz, and Nick Craswell. 2016. On Obtaining Effort Based Judgements for Information Retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 277–286.

Annalies Vuong, Tristan Nixon, and Brendon Towle. 2011. A Method for Finding Prerequisites Within a Curriculum.. In *EDM*. 211–216.

Tong Wang, Xingdi Yuan, and Adam Trischler. 2017. A Joint Model for Question Answering and Question Generation. *arXiv preprint arXiv:1706.01450* (2017).

James V Wertsch. 1984. The zone of proximal development: Some conceptual issues. *New Directions for Child and Adolescent Development* 1984, 23 (1984), 7–18.

Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 132–141.

Ryen W White and Eric Horvitz. 2009. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)* 27, 4 (2009), 23.

Barbara M Wildemuth. 2004. The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology* 55, 3 (2004), 246–258.

Kevin H Wilson, Yan Karklin, Bojian Han, and Chaitanya Ekanadham. 2016. Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. *arXiv preprint arXiv:1604.02336* (2016).

Mathew J Wilson and Max L Wilson. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 291–306.

Christopher R Wolfe, Valerie F Reyna, Colin L Widmer, Elizabeth M Cedillos, Christopher R Fisher, Priscila G Brust-Renck, and Audrey M Weil. 2015. Efficacy of a web-based intelligent tutoring system for communicating genetic risk of breast cancer: A fuzzy-trace theory approach. *Medical Decision Making* 35, 1 (2015), 46–59.

Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium*. ACM, 254–257.

Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. 2014. Relevance and Effort: An Analysis of Document Utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 91–100.

Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting User Knowledge Gain in Informational Search Sessions. In *Proceedings of the 41th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Cheng Xiang Zhai, William W Cohen, and John Lafferty. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 10–17.

Meilan Zhang and Chris Quintana. 2012. Scaffolding strategies for supporting middle school students' online inquiry processes. *Computers & Education* 58, 1 (2012), 181–196.

Shuai Zhang, Kai Lu, and Bin Wang. 2011. ICTIR Subtopic Mining System at NTCIR-9 INTENT Task.. In *Proceedings of NTCIR-9 Workshop Meeting*. 106–110.

Xiangmin Zhang, Jingjing Liu, Michael Cole, and Nicholas Belkin. 2015. Predicting users' domain knowledge in information retrieval using multiple regression analysis of search behaviors. *Journal of the Association for Information Science and Technology* 66, 5 (2015), 980–1000.

Xiaojin Zhu. 2013. Machine teaching for bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems*. 1905–1913.

Joerg Zumbach and Maryam Mohraz. 2008. Cognitive load in hypermedia reading comprehension: Influence of text type and linearity. *Computers in Human Behavior* 24, 3 (2008), 875–887.