Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation

Rohail Syed
University of Michigan
Ann Arbor, Michigan
rmsyed@umich.edu

Kevyn Collins-Thompson
University of Michigan
Ann Arbor, Michigan
kevynct@umich.edu

Paul N. Bennett
Microsoft Research AI
Redmond, Washington
pauben@microsoft.com

Mengqiu Teng
University of Michigan
Ann Arbor, Michigan
mengqiu@umich.edu

Shane Williams
Microsoft Research AI
Redmond, Washington
shanewil@microsoft.com

Dr. Wendy W. Tay*
Independent
Montréal, Québec, Canada
wendy.tay@protonmail.ch

Shamsi Iqbal
Microsoft Research AI
Redmond, Washington
shamsi@microsoft.com

ABSTRACT

As AI technology advances, it offers promising opportunities to improve educational outcomes when integrated with an overall learning experience. We investigate forward-looking interactive reading experiences that leverage both automatic question generation and analysis of attention signals, such as gaze tracking, to improve short- and long-term learning outcomes. We aim to expand the known pedagogical benefits of *adjunct questions* to more general reading scenarios, by investigating the benefits of adjunct questions generated after participants attend to passages in an article, based on their gaze behavior. We also compare the effectiveness of manually-written questions with those produced by Automatic Question Generation (AQG). We further investigate gaze and reading patterns indicative of low vs. high learning in both short- and long-term scenarios (one-week followup). We show AQG-generated adjunct questions have promise as a way to scale to a wide variety of reading material where the cost of manually curating questions may be prohibitive.

CCS CONCEPTS

• **Information systems** → *Personalization*; • **Human-centered computing** → *User models*; *Laboratory experiments*; *Interaction techniques*; • **Applied computing** → *Interactive learning environments*.

KEYWORDS

Education/Learning, Gaze tracking, Lab study, Personalization, User modeling

*Work performed while at Microsoft Research.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380240

ACM Reference Format:
Rohail Syed, Kevyn Collins-Thompson, Paul N. Bennett, Mengqiu Teng, Shane Williams, Dr. Wendy W. Tay, and Shamsi Iqbal. 2020. Improving Learning Outcomes with Gaze Tracking and Automatic Question Generation. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3366423.3380240

1 INTRODUCTION

The advance of AI technology offers an opportunity to improve educational outcomes by transforming fundamental learning experiences such as reading. Moreover, given that the Web is one of the world's primary information sources, online users spend a significant amount of time and effort searching [3] and in-depth reading of content on the Web in pursuit of their learning goals, using a multitude of sites, especially reference sites such as Wikipedia [27].

In addition, past educational studies in controlled settings have demonstrated the multiple benefits of *adjunct questions*, which are questions inserted into text to draw attention to important textual material [13] as a form of active learning [2, 6, 16, 28]. While the specifics of these studies vary, the general results are consistent: people who learn in a way that requires active processing of the subject matter tend to show significantly better learning outcomes. Thus, providing readers with relevant adjunct questions is a potentially powerful mechanism to increase learning in a range of reading-to-learn scenarios.

Using adjunct questions broadly, however, involves several challenges. Foremost, how could questions be sourced at scale for the vast amount of reading material available on the Web and other open information repositories? Given that people frequently skim text, as well as read it more deeply, could attention signals be used to choose which portion of text to draw questions from, so as to maximize the potential for learning that is both effective and efficient? Are some types of questions more effective for improving learning outcomes? Finally, in hopes of developing implicit metrics rather than more time-consuming instruments, are there gaze patterns that are more predictive of learning outcomes?

The goal of this study is to provide a basis for significant progress on these questions, by investigating the potential for adaptive, interactive reading experiences that can leverage the powerful pedagogical benefits of adjunct questions in a way that could also scale to the Web. Our approach combines recent advances in deep models for automatic question generation with gaze tracking for reading behavior analysis.

With this technology, we conduct an extensive interactive user study to investigate the benefits of adjunct questions where the questions are presented adaptively, i.e. only after, and based on, what the participant is known to have attended to when reading an article (see Figure 1). We compare both manually human-curated and Automatic Question Generation (AQG) as a potential source of such questions. In contrast to prior work, ours is the first to study either human-curated or AQG questions in an attention-informed setting; as such, finding a resulting improvement in learning outcomes regardless of question source would be interesting – though we note the AQG setting offers better promise for scaling to all information sources. We evaluate whether the use of adjunct questions that are not directly in text (and therefore visible to the learner before reading) provides benefits in terms of both short- and long-term learning outcomes. We further analyze whether there are gaze patterns or reading behaviors that are indicative of the potential presence or absence of learning gains. In sum, we demonstrate successful use of adaptive, auto-generated adjunct questions: a finding that has significant implications for effective, large-scale application of this technique to a wide variety of reading material where the cost of manually curating questions may be prohibitive.

2 RELATED WORK

We first discuss prior work in active learning and how it relates to interactive reading experiences. We then discuss prior gaze tracking research for both understanding and supporting learning, and how this work informs our study design.

2.1 Improving Learning Through Active Learning Strategies

Active learning practices, which aim to engage students to participate more in the learning process, have been shown to have positive outcomes in classroom environments [5, 8, 16, 17]. Various methods of active learning have been explored to understand such benefits. For example, Jenkins *et al.* [17] showed that asking students to summarize paragraphs they read substantially improved their reading comprehension compared to students in a control condition. As another example, work by Frey & Fisher [16] found that teachers whose students performed at high assessment standards engaged students by asking a range of different types of questions as a regular part of their classes.

Other relevant active learning literature that directly informs our study is Peverly and Wood's work on the *adjunct questions effect* [22]. In that paper, the authors reported that augmenting reading material with questions in text led to improved learning. Other work has also found that the use of questions as part of the reading process produced benefits in learning outcomes. Callender & McDaniel [7] found significantly better learning outcomes among participants whose learning materials included embedded questions. In a study by Dornisch and Sperling [14], adjunct questions presented alongside the reading material resulted in both short-term and long-term learning gains.

Similarly, we also leverage adjunct questions as a mechanism for promoting positive learning outcomes. However, we build upon and extend the aforementioned prior work in two major ways. First, rather than present questions at predetermined points in the text, we adaptively present questions to the reader based on what they have just read, using gaze tracking to detect the location and nature of the reader's visual attention to words in the text. In this way, we leverage knowledge about content that the reader has merely skimmed, versus read more carefully. Second, we investigate the effects of different types of questions on adjunct learning, comparing outcomes based on questions from an automatic question generator (AQG) to questions that are manually created. This allows us to understand the potential for applying automatic question generation effectively at scale to the Web or other open collections.

2.2 Leveraging Eye-Gaze Patterns to Understand Learning

Following earlier work exploring the connections between eye-gaze patterns and information processing [18, 29], a number of studies have focused in recent years on the use of gaze-tracking to understand the relationship between a person's eye movement patterns and their knowledge state [4, 9–11, 19]. For example, Bhattacharya & Gwizdka [4] investigated how *fixations* (fixed eye gaze on one area for a stable period of time) and *regressions* (backward directed eye movements) differed between those who showed low and high levels of knowledge gain during a Web-based learning task. Their study found that those who showed higher knowledge gain also tended to have fewer fixations in sequences of fixations, and spent less time per fixation on average. There was also evidence that the low knowledge gain group tended to show more and longer backwards regressions. Earlier work by Cole *et al.* [9] showed that perceptual span and time spent reading was strongly predictive of a person's prior knowledge in the medical domain. Later work by Mao *et al.* [19] further applied the main eye movement variables from [9] to also investigate the link between domain-specific knowledge and eye movement behaviors and reached similar conclusions. Eye movement behavior can also indicate different types of learning strategies in different scenarios. Work by Copeland & Gedeon [10] found that learners 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 revisit it.

2.3 Using Gaze Tracking to Support Learning

In addition to using gaze-tracking as an outward signal to assess knowledge or learning, other studies have investigated how signals from gaze tracking could be used as an input to systems for supporting learning. Copeland *et al.* [12] presented a framework for providing adaptive difficulty in text content as a function of estimated comprehension, which would be determined through gaze tracking. In work by Eskenazi & Folk, reading regressions were found to be not only a sign of simple oculomotor correction,

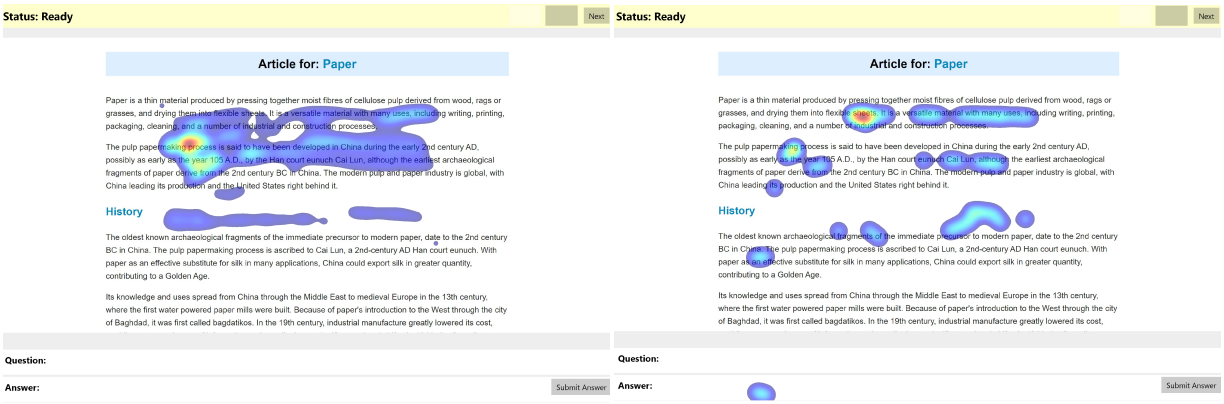

Figure 1: Gaze fixation heatmap on article page for a participant on the topic 'Paper'. Question/response area is below the content area. Left: Fixation heatmap before a question was asked. Right: Fixation heatmap after a question was asked: "What is a common use for paper?".

but also an indicator of re-finding behavior as well as comprehension difficulty [15]. Earlier work by Sibert *et al.* [26] introduced The Reading Assistant, which was an adaptive tutoring system that helps learners struggling with understanding a particular word by detecting their eye movements and taking appropriate personalized action in the form of auditory feedback.

In contrast to previous work on gaze tracking and learning, we use gaze primarily as an attention signal to choose when and how to present questions dynamically (either human-curated or automatically generated, depending on condition). We also investigate low-level gaze patterns that are associated with learning in this interactive setting, with particular focus on the differences between low- and high-knowledge participants. To the best of our knowledge, there have been no studies which have compared the use of automatically-generated questions versus human-curated questions in the context of the adjunct questions effect. We are also unaware of prior studies that have investigated the adjunct questions effect in a context where the questions appear adaptively during the reading in response to content the learner is known to have read or skimmed.

Finally, we note that gaze tracking is just one method for estimating users' attention to content: a number of studies have shown (e.g. [25], [20]) that cursor/mouse movements might be used as an approximate proxy for gaze in some reading scenarios, while other approaches have used common webcams to estimate the gaze positions of users without requiring that video be sent to a server [21]. Thus, providing adaptive questions during reading could potentially be implemented at scale within a browser using one or a combination of these generic approaches, without the need for commercial gaze tracking devices or any other specialized hardware.

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 automatically 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.

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 [1], 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 automatically generated questions, we used a generative model based on the work by Wang et al. [30] and provided through an external API by Microsoft Research. We will refer to this service as our AQG API.

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 learning outcomes?

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)?

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 Sec. 4).

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 [23], 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 auto-generated questions with crowdsourced questions in a learning task. For example, for the topic 'Pain', for a particular paragraph we had the following human-curated (Human) and AQG-generated (AQG) questions:

Human: What year was peripheral pattern theory developed?

AQG: In what year did DC Sinclair and G Weddell develop peripheral pattern theory?

[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 [23].

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) [11]. 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 low ($0 < \text{NNF} < 0.70$). We chose the threshold of 0.70 based on prior work [11]. We considered a paragraph for generating 'Focus-Reading' questions if its NNF was at or above this threshold ($\text{NNF} \geq 0.70$).

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 using 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 of 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.

3.6 Measuring Learning Outcomes

We measured how well participants had learned the content by asking questions at two time points: immediately after they finished reading (*post-test*), and then a week later (*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 delayed-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.

4 METHODOLOGY

4.1 Study Interface

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 internally assessed for each paragraph if the reader had likely skimmed (S) the paragraph or performed focused reading (F). In all of the conditions that involved asking 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 1). The condition assigned at any given point determined *which* questions (if any) would be asked during the reading phase. The experiment design involved the following four conditions 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 from paragraphs the learner showed focused-reading over.) Each participant answered *exactly* four questions based on 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 any topic.

Each participant in the study completed four learning tasks.[2] There was one task per 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 to help ensure ordering effects were balanced on average across participants with respect to topic and condition.

4.2 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. Subjects were recruited through a recruitment email to a large distribution list 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 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 one week later.

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.

[2] In a pilot study we chose six topics. Participants reported the experiment took too long and individual articles were too long. We reduced to four topics and reduced content length by 25% for the full study.

4.3 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, i.e. 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.

4.4 Grading

Since users gave 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) 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 grade.

4.5 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 examine the first topic/condition that a participant completed, as well as perform 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]

[3]Formerly Crowdflower.
[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.

Measure	Q_{None}	Q_{Auto}	Q_{Human}	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 1: Average values for different learning measures by condition. Marked values indicate significant differences b/w that condition and Q_{None}. Also shown is breakdown by question type: Base (seen in pre-test), New (post-test only), and All (Base+New).

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% of 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.

5 RESULTS - LEARNING OUTCOMES

We present an analysis of learning outcomes here, and an analysis of real-time reading behavior patterns in Sec. 6.

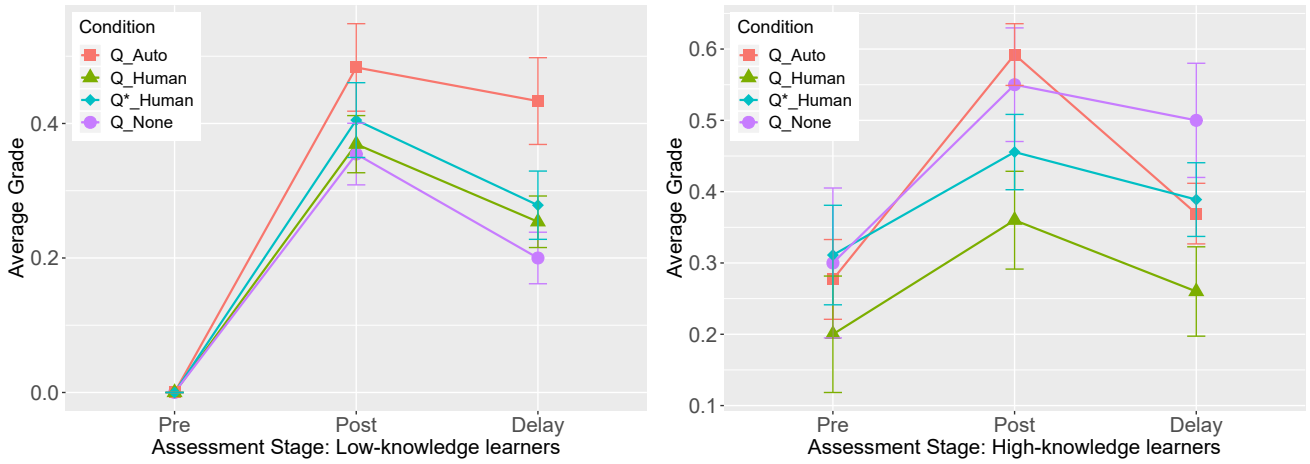

Figure 2: Breakdown of average test item scores at each stage, by condition, showing that in general both short-term and long-term learning is happening for all conditions. Left: Low-knowledge (LK) learners. Right: High-knowledge (HK) learners. Error bars are standard errors.

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 2. 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 confirm that participants on average are indeed learning.

Table 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.

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.

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 LK learners who received adjunct questions while reading had significantly higher long-term retention (grade in delayed post questions) than $\mathbf{Q_{None}}$ participants (M=.30 vs M=.20, p=.04). These results are shown in Figure 3. 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. This suggests that adjunct questions have a positive association with long-term retention of content for those with no prior knowledge on the topic; however, adjunct questions may not be as beneficial for those with some prior knowledge of the topic, and may perhaps impede their natural reading flow.

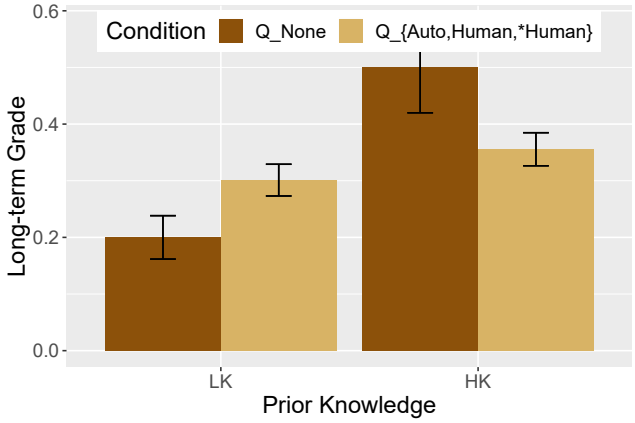

Figure 3: Long-term grades by non-question vs. question condition and prior knowledge level. LK participants particularly benefit in the case of interactive questions.

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

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

Result	Short-term learning	Long-term retention
Adjunct Questions improved grades better than $\mathbf{Q_{None}}$ (Section 5.2)	No	Yes (for low-knowledge learners)
$\mathbf{Q_{Auto}}$ performed comparable to $\mathbf{Q_{Human}}$ (Section 5.3)	Yes	Yes
Synthesis question affected grades (Section 5.4)	No	No
Focus-based question selection improved grades (Section 5.5)	No	Yes
Gaze behavior was different for those who would answer questions correctly (Section 6.4)	Yes (for low-knowledge learners)	Yes (for low-knowledge learners)

Table 2: Major conclusions regarding learning outcomes and reading behaviors/treatments.

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 driven 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) [24], and this evidence helps support our hypothesis that these detailed questions gave rise to more focused reading and thus a difference in performance. Interestingly, at first glance AQG questions may appear too detailed and simplistic (as simple textual rewrites of input passages) but in a learning scenario these same properties may help readers quickly find the right passage in the document and then require focused reading, facilitating learning.

5.4 Effects of 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 1). This may be in part due to the extra time on task (see Section 6.1) that is spent when a synthesis question is asked.

5.5 Skim- vs Focus-Reading Adjunct Questions

In **RQ4** we asked if learning outcomes varied based on questions that were selected using gaze focus patterns. More specifically, we wanted to see if differences existed in learning outcomes when participants had skimmed over content, versus performed focused reading, as determined by gaze-tracking.

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 whether those participants who received at least one skim-reading question showed a different learning outcome than those who didn't. We refer to this binary variable as **GotSkim**, and in particular we 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 re-read more carefully to answer the question - which resulted in better delayed post-test scores, indicating the potential importance of adaptive, focus-based adjunct questions for long-term retention.

6 RESULTS - READING/TIME PATTERNS

In Sec. 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.

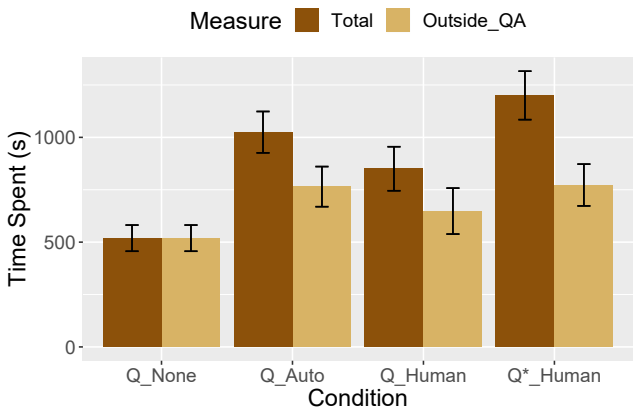

Figure 4: Breakdown of average reading time by treatment. Outside_QA is the reading time not spent answering questions. Results suggest that being given questions encourages participants to spend more time reading excluding time needed to answer questions.

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 the timestamp difference between the first and last gaze event on the article. As expected, $\mathbf{Q_{None}}$ had the lowest average time, $\mathbf{Q_{Auto}}$ 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 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.

6.2 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).

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

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).

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 learning outcomes?

we investigated the relationship between normalized number of fixations (NNF) and post-test scores. As a reminder, 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 3 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 [11]. 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 *et al.* [24]. 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 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.

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 3: 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.

7 DISCUSSION

A summary of our study findings is shown in Table 2. 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 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) than $\mathbf{Q_{Human}}$, 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 that participants did show significantly better long-term results when asked at least one question about a paragraph that was 'focus-read', compared to those who got only questions about 'skimmed' content. This highlights the potential importance of asking questions adapted to content that participants did, or did not, pay attention to – something we achieved via real-time gaze tracking.

In addressing **RQ5**, we found strong evidence that a measure of gaze fixations, the *normalized number of fixations* or NNF, was significantly higher when participants answered post-test and delayed post-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 any of these NNF types either in short- or 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. In sum, gaze fixation behavior may help indicate, for learners with low prior knowledge, how much they are actually learning, including estimation of their likely long-term retention of knowledge.

In our experiment implementation, there was a potential concern that the gaze tracking software's calibration may have needed recalibration, 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, we found limited benefit to introducing adjunct questions, and in some cases, potentially detrimental effects. Thus, we suggest that using adjunct questions may not be appropriate 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 a topic).

For low-knowledge learners, we observed higher learning performance in both short-term and long-term outcomes. For long-term outcomes, 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.

8 CONCLUSIONS

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; and (2) where the questions are generated through an automatic question generation (AQG) API versus more traditional manual curation. Our results reinforce earlier findings on the learning benefits of adjunct questions, though in our study we found these were limited to learners with low prior knowledge of a topic. We further found evidence that automatic question generation can perform comparably to – and in some cases, better than – human-curated questions in this scenario. Our results demonstrate the promising potential that applying the benefits of the adjunct questions effect might have for large-scale learning-oriented applications, such as embedding questions directly into arbitrary web pages, encyclopedia entries or digital textbooks. We also showed that gaze tracking signals based on reading fixations 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.

REFERENCES

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

[2] Tessa M Andrews, Michael J Leonard, Clinton A Colgrove, and Steven T Kalinowski. 2011. Active learning not associated with student learning in a random sample of college biology courses. *CBE-Life Sciences Education* 10, 4 (2011), 394–405.

[3] 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.*

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

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

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

[7] 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.

[8] Julie Campbell and Richard E Mayer. 2009. Questioning as an instructional method: Does it affect learning from lectures? *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 23, 6 (2009), 747–759.

[9] 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.

[10] 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.

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

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

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

[14] 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.

[15] 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.

[16] Nancy Frey and Douglas Fisher. 2010. Identifying instructional moves during guided learning. *The Reading Teacher* 64, 2 (2010), 84–95.

[17] Joseph R Jenkins, James D Heliotis, Marcy L Stein, and Mariana C Haynes. 1987. Improving reading comprehension by using paragraph restatements. *Exceptional children* 54, 1 (1987), 54–59.

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

[19] 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.

[20] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. 2013. Measurement and Modeling of Eye-mouse Behavior in the Presence of Nonlinear Page Layouts. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13).* ACM, New York, NY, USA, 953–964. DOI: http://dx.doi.org/10.1145/2488388.2488471

[21] Alexandra Papoutsaki, James Laskey, and Jeff Huang. 2017. SearchGazer: Webcam Eye Tracking for Remote Studies of Web Search. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR).* ACM.

[22] 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.

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

[24] 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.

[25] 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. DOI: http://dx.doi.org/10.1145/1358628.1358797

[26] 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.

[27] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1591–1600. DOI: http://dx.doi.org/10.1145/3038912.3052716

[28] 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.

[29] 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.

[30] Tong Wang, Xingdi Yuan, and Adam Trischler. 2017. A Joint Model for Question Answering and Question Generation. *arXiv preprint arXiv:1706.01450* (2017).