

READ, SKIM, SCAN: GAZE-AWARE DOCUMENTS AS IMPLICIT
FEEDBACK FOR MULTIMODAL TYPOGRAPHICAL CUING

ADELIZ KEITH



*A thesis submitted to the
School of Graduate and Postdoctoral Studies in partial
fulfillment of the requirements for the degree of*

Master of Science
in
Computer Science

Faculty of Science
University of Ontario Institute of Technology (Ontario Tech University)

December, 2023 – version 4.2

Adeliz Keith: Read, Skim, Scan: Gaze-Aware Documents as Implicit Feedback for Multimodal Typographical Cuing, © December, 2023

SUPERVISOR:

Dr. Christopher Collins and Dr. Steven Livingstone

COMMITTEE:

Dr. Bill Kapralos

Dr. Loutfouz Zaman

LOCATION:

Oshawa, Ontario, Canada

DATE:

December, 2023

ABSTRACT

Computer-based methods for displaying and formatting texts for speed reading, including the popular Rapid Serial Visual Presentation method, are increasingly popular in both research and commercial applications. However, these techniques tend to be intrusive and optimized only for maximally efficient speed reading. Here, we present a technique for multimodal text layouts utilizing typographical cuing, making the text responsive to the user's reading behavior as observed through eye tracking. We present a quantitative and qualitative evaluation of our work; finding high usability metrics and positive qualitative feedback, but no significant effects on task performance. Our work serves to replicate and extend prior work on gaze-aware, attentive documents.

AUTHOR'S DECLARATION

I, Adeliz Keith, hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

The research work in this thesis was performed in compliance with the regulations of Research Ethics Board/ Animal Care Committee under REB Certificate number file number #17112.

Oshawa, Ontario, Canada, December, 2023

Adeliz Keith

STATEMENT OF CONTRIBUTIONS

I hereby certify that I am the sole author of this thesis and that no part of this thesis has yet been published or submitted for publication as of December, 2023. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis.

Adeliz Keith

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor, Christopher Collins. Even before I officially joined the lab, Chris provided guidance, support, and an enormous amount of his time and energy to help me succeed. He helped me through the most difficult times of my thesis, including strategic and emotional advice when we were forced to make a large-scale pivot more than six months in. Even while working a full-time job at Meta, he has been generous and giving with his time, and I could not have done this without him.

I would like to thank my co-supervisor, Steven Livingstone. I met Steven relatively late into my studies, but he provided excellent advice and review as we finished together. The outside perspective that he provided was crucial for some of the most difficult design decisions that faced us. Like me, he is a newcomer to the university; I give him and his lab warm welcomes and the best of luck.

I would like to thank my friends and colleagues at the Vialab for their guidance, advice, and help. The entire lab was vital to my success, and I'm especially proud of the way we all came together to support one another after the difficult news that Chris would be on hiatus from the university.

I would like to thank Alexandra Elbakyan for her ongoing contribution to the cause of open science.

Finally, I would like to thank my partner, Celyn. Ebie, you always inspire me to try harder. You and our seven dogs are home, wherever home is; you are back-to-back with me on the climb; you make me soup. Until the world is mended.

Thank you.

CONTENTS

List of Figures	xiii
List of Tables	xiv
1 INTRODUCTION	1
1.1 Contributions	2
1.2 Organization	2
2 READING PHYSIOLOGY AND PSYCHOLOGY	5
2.1 Eye Tracking	5
2.1.1 Gaze Data	5
2.1.2 Eye Tracking During Reading	6
2.2 Reading Behaviors	7
3 READING AUGMENTATIONS	13
3.1 Text Layout Methodologies	13
3.1.1 Disruptive Interventions	13
3.1.2 Typographical Cuing	15
3.2 Attentive Documents	17
3.2.1 Reading-Skimming Detection	17
3.2.2 Attentive Documents	20
4 A REALTIME READ-SKIM-SCAN CLASSIFIER	23
4.1 Model Design	23
4.1.1 Reading and Skimming	26
4.1.2 Scanning	27
4.2 Comparison to Existing Models	28
4.2.1 Personalization	29
5 IMPULSE READING: ATTENTIVE TEXT LAYOUT	31
5.1 Design Overview	31
5.1.1 Text Layout Modes	31
5.1.2 User Interface Design	34
5.1.3 Software Components	35
5.2 Control Techniques	35
5.2.1 Automatic Control Technique	35
5.2.2 Manual Control Technique	36
6 EVALUATION	39
6.1 Methods	39
6.1.1 Participants	39
6.1.2 Design and Stimuli	40

6.1.3	Procedures	41
6.1.4	Analyses	44
6.2	Results	45
6.3	Discussion	47
6.3.1	Limitations	53
7	CONCLUSION	55
A	APPENDIX	59
	BIBLIOGRAPHY	63

LIST OF FIGURES

Figure 1	Examples of reading, skimming, and scanning behaviors. .	8
Figure 2	Examples of reading, skimming, and scanning behaviors. .	10
Figure 3	An analysis of a mixed example of reading, skimming, and scanning.	12
Figure 4	An example of typographical cuing used for reading aug- mentation.	16
Figure 5	An example of the Text 2.0 framework.	21
Figure 6	The plaintext text layout used in Impulse Reading.	31
Figure 7	The content words layout used in Impulse Reading.	32
Figure 8	The sentence fadeout layout used in Impulse Reading. . .	32
Figure 9	Raincloud plot of comprehension scores by condition. . . .	48
Figure 10	Raincloud plot of the percentage of fixations on relevant text by condition.	49
Figure 11	Diverging bar chart of Likert scale responses.	50
Figure 12	Survey results for the forced-choice preference questions answered by each participant at the end of the study. . . .	51
Figure 13	Raincloud plot of SUS scores by condition.	52
Figure 14	Tutorial text explaining the Manual control mode.	59
Figure 15	A task introduction containing roleplaying instructions. . .	60
Figure 16	A task screen, as seen by participants during the study. . .	60
Figure 17	The comprehension questions screen for the previous task.	61
Figure 18	A post-survey study.	61

LIST OF TABLES

Table 1	Selected machine learning models performing reading behavior classification.	19
Table 2	Possible saccade classifications, including the point values for each detector for each category.	27

INTRODUCTION

Written language, and therefore the activity of reading it, have existed for thousands of years. Despite this long history, the process of reading has remained relatively unchanged over time. The advent of computerized text displays has allowed for experimentation with new methods of displaying and reading text. Techniques like the popular Rapid Serial Visual Presentation (RSVP) and others have been explored in both scientific and commercial applications.

However, these technologies have yet to find widespread acceptance. We argue that this is due to a myopic focus on *speed reading* as the only metric of value, an opinion commonly displayed in advertising for RSVP and other techniques. Little interest is shown in supporting the highly diverse reading behaviors displayed in the vast spectrum of human experience. Some reading activities are best done slowly and carefully, like studying or copyediting, while others are dependent on more ephemeral qualities; for example, reading for simple pleasure and fun. Not only are most experimental reading augmentations optimized for supporting only a single reading behavior (usually, reading as fast as possible), they force the reader into complying with that behavior by disrupting the existing reading process.

We believe these disruptive techniques are not the most promising avenue of exploration for reading augmentations; instead, we are interested in work like Kobayashi et al. or Biedert and Buscher [12, 48]. These techniques are non-disruptive, because they work by applying minor typographical changes to existing text; despite this restrained approach, they are able to support specific reading behaviors like skimming or scanning.

Furthermore, we also believe non-disruptive reading augmentations that could improve reading comprehension or speed with no tradeoff would be extraordinarily useful across a wide variety of human domains. A population survey of Americans found that 25% of workers spend an hour or more each day reading emails; this means that Americans collectively spend hundreds of millions of hours each year processing text for work communication alone [30]. Professors and students in the university sector read almost constantly across many distinct tasks, in a way that has been described as difficult to quantize because it is “beyond measure” [71]. Even minor improvements in reading efficacy could create immense time savings, leading to increased worker productivity. More subjec-

tively, reading augmentations could potentially lead to reduced frustration and increased satisfaction, or even fun, across both occupational and personal time.

While these non-disruptive augmentations are a promising direction of research, prior research has predominantly investigated their application to only a single reading behavior. The seminal works of Biedert and Buscher extended this by asking compelling questions: what if text could be *attentive* or *responsive*? What if text could react to the reader, and optimize itself for the reader's free choice of reading style?

This work aims to answer these questions. We aim to find non-disruptive reading augmentations that are optimized for multiple distinct reading behaviors, and allow the reader to take control as we provide support tailored for whatever type of reading they choose.

1.1 CONTRIBUTIONS

The main contributions of this work are as follows:

1. A novel classifier that takes in a realtime stream of gaze data and performs a three-way classification into reading, skimming, or scanning (Chapter 4).
2. A reading augmentation named Impulse Reading which uses eye tracking to dynamically switch between multiple text layout methods based on the reader's current task and goal (Chapter 5).
3. An empirical evaluation of task performance and usability of Impulse Reading, including a thematic analysis of qualitative feedback data, as measured in a text search task (Chapter 6).
4. A public dataset of gaze data during a text search task, including conditions for Impulse Reading and control conditions, which can be used by future researchers for training or validation of machine learning models for reading behavior classification (Chapter 6).

1.2 ORGANIZATION

Chapters 2 and 3 present a literature review of prior work. Chapter 2 contains basic research related to eye trackers and reading behaviors. Chapter 3 describes reading augmentations, and documents that are responsive to the process of being read. Chapter 4 discusses the design and implementation of our reading-skimming-scanning classifier. Impulse Reading is described in Chapter 5, followed by the methods and results of our experimental validation in Chapter 6.

Finally, Chapter 7 concludes by describing ideas for future work and limitations of the present thesis.

This chapter, the first of two concerning prior literature, summarizes research on the human behavior of reading as revealed through the use of eye trackers. We review the mechanics of the human eye and the eye trackers that investigate it. We close by reviewing the diversity of reading behaviors displayed by humans undergoing different reading-related tasks.

2.1 EYE TRACKING

The first experiment involving precise tracking of the human eye was performed by Charles Bell, in an experiment that turns 200 this year [10]. Despite this long history, the methods of eye tracking remained difficult, imprecise, and effortful until the introduction of video-based eye trackers [20]. In a 2007 review of eye tracker techniques, Duchowski describes these trackers as using relatively inexpensive cameras to compute the user's point of regard [26]. The video stream is processed to identify eye features such as corneal reflection and the coordinates of the pupil, and these features are used to estimate the angle of the user's eye. Eye trackers that use corneal reflections require the presence of an external light source, which is usually an infrared light built into the eye tracker hardware. Eye tracking devices that do not require infrared light illumination, like widely available commercial webcams, have been explored, but currently have lower accuracy, precision, and validity [72]; instead, most modern eye tracking research is conducted using specialized hardware [26]. These eye trackers can be mounted on a table or directly on the user's head [20]. In this chapter, we use data and evidence collected through eye trackers to describe how the human visual system is used during reading.

2.1.1 Gaze Data

While the human eye can respond to light across a wide angle, its visual acuity is by far the highest in a small region directly in the centre of the of the visual field called the *fovea centralis*. Because only a tiny portion of the visual field can be inspected with maximum acuity at a time, the eye is required to move to sequentially inspect multiple areas in visual scenes with more than one point of

interest [65]. As such, a temporal analysis of the human gaze is characteristically divided into short periods of movements called *saccades* and longer periods of relative stillness called *fixations* [40].

The eye is not perfectly smooth during fixations, with eye movements during fixations being divided into microsaccades, ocular drift, and tremor [65]. However, these fixational eye movements are much smaller than the movements during saccades. Microsaccades correlate less with human attention and information processing, and therefore reveal less information about mental state. Most scientific analysis concerning eye trackers treats fixational eye movements as sources of noise that complicate the problem of fixation identification, rather than an inherently valuable data source. Ignoring these fixational eye movements, normally-functioning eye gaze can be sharply divided into fixations and saccades.

When investigating a visual scene with more than one point of interest, the human eye moves on average every 250-350 ms [60]. This duration can also be thought of as the duration of the fixation that occurs between each saccade. Saccades themselves are highly variable in duration; short saccades last around 20-30 ms, while the largest saccades lasting up to 200 ms.

Finally, attention and gaze are closely interlinked in humans. While it is certainly possible for attention to be placed somewhere in the visual field besides the foveal area, doing so generally either requires constant conscious control or denotes an imminent shift in fixation to the area of attention [60]. As such, eye gaze data as revealed through an eye tracker can be used to investigate human attention.

2.1.2 *Eye Tracking During Reading*

Eye gaze during reading, as during all other activities, are predominantly composed of saccades separated by fixations. However, these fixations and saccades are highly stereotypical during reading. The typical saccade is approximately six to nine letter spaces across a wide range of font sizes [61]. Saccades last approximately 20-30 ms, being shorter and with less extreme variation in distance than human eye gaze in general. The average fixation duration is between 200 to 250 ms, slightly shorter than general eye gaze.

Information intake only occurs while the eyes are still; the angular velocity during saccades are too fast for readers to acquire usable information [73]. Unsurprisingly, saccades are predominantly in the direction of the language (for example, left-to-right in English); however, approximately 10% to 15% of saccades

move backwards in the text (called a *regression*) to look at words or phrases that have already been read [59].

Average fixation duration and saccade distance, while less variable in reading than in general eye gaze, still display considerable variation both between readers and within readers. As a general statement, the within-reader variability is higher than the between-reader variability; a reader's maximum and minimum in these metrics will differ significantly more than the difference of the averages of two readers. While a typical reader might have an average fixation duration of 225ms and an average saccade length of 8 letter spaces, their fixations might range from under 100ms to over 500ms within a single passage of text. Their saccade lengths might range from 1 letter space to over 15 letter spaces [59]. These variations are associated with several features of both the text and the reader, including text difficulty, text layout and font, language fluency, current task, mental fatigue, and presence or absence of different reading behaviors like skimming [14, 58].

2.2 READING BEHAVIORS

Several theories of reading exist, including Just and Carpenter's model focused on the processing load of individual fixations, Kintsch's Constructive-Integrative model, LaBerge and Samuels' theory focused on attention and automaticity of processing, and more [43, 47, 50]. While these theories are in many ways non-contradictory, their focus differs. Most theories concern themselves predominantly with the semantic processing that occurs during thorough reading, including the reflective integration of context as it relates to words, sentences, and passages. However, most theories do not focus on the high variety of reading behaviors; for example, Just and Carpenter's (excellent) paper describes a variety of reading behaviors in less than a page, with little specificity of analysis placed on each example [43]. For the current thesis, it would be beneficial to use an analytical framework that more heavily focuses on the highly heterogeneous nature of different reading behaviors. To do this, we use Robert Carver's theory of *rauding* [22].

Rauding theory focuses on the processing "components" that occur during each fixation while reading; it identifies five separate reading behaviors based on the specific components that occur or do not occur while in that behavior. Rauding theory differs from other theories of reading in its depth of analysis and specificity of predictions over non-traditional (i.e., not thorough reading) reading behaviors. In this section, we describe these disparate reading behaviors, and we provide an overview of the empirical evidence for their existence.

Etymologically, rauding is derived from "auding", the process of comprehending spoken language, as well as "reading". Carver made this unusual choice to highlight that the comprehension processes underlying reading and listening are similar, or according to his claims even identical.

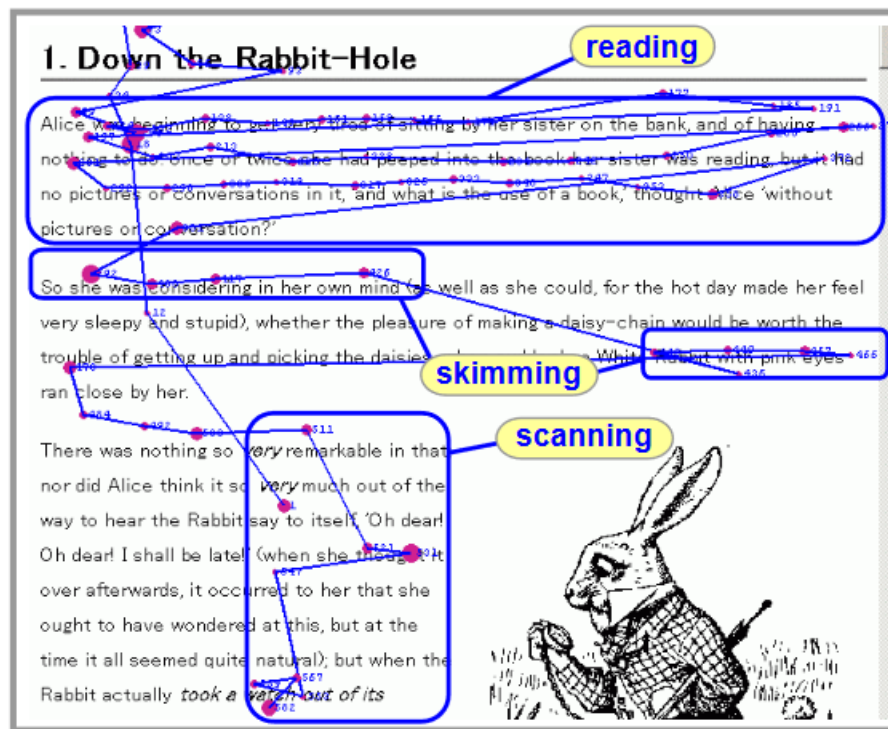


Figure 1: A real-world example of reading, skimming, and scanning behaviors, from Ohno [56].

As described by Carver's theory of reading, skimming and scanning are the two reading behaviors with a higher reading speed and lower comprehension than thorough reading. Both behaviors characteristically differ from thorough reading in lacking *sentential integration*; i.e., they do not integrate the complete thought of most sentences. Skimming consists of *lexical access*, or word recognition, along with *semantic encoding*, or determining the meaning of a word as it relates to the words around it. Of course, lexical access and semantic encoding are applied only to some words of a sentence in a skimming behavior, as performing semantic encoding on every word of a sentence would suffice to also perform sentential integration.

Scanning behavior differs from skimming in that it lacks semantic encoding; consisting solely of lexical access, scanning behaviors can search for and recognize specific words or types of words, but cannot apply sentence-level or word-level context to those words. Scanning is the reading behavior with the highest overall reading speed (measured in words per minute) described by Carver.

Because readers use these behaviors to increase their reading speed at the cost of comprehension, Carver describes the overall speed for these behaviors for fluent college students. He describes 300 WPM as a typical college rate for reading. Skimming is on average 50% faster, at 450 WPM. Scanning is twice as fast as reading, at 600 WPM on average for college students. However, it should

be noted that these numbers vary quite heavily by individual. Carver describes these specific numbers without reference to any standard deviation or error of measurement based on his research summarized in a 1990 book, but the exact ratio of speedup experienced when switching to skimming or scanning should not be considered well-known based on this research alone [21].

Skimming has been found to be adaptive, meaning that it is not a strictly inferior version of thorough reading [27]. Specifically, skimming while under time pressure can improve performance on a memory task compared to a strategy maintaining strictly thorough reading. Skimming is predominantly a *satisficing* process: the reader will skim continuously over a block of text until their information intake is reduced below a certain threshold, at which point they will skip to the next block of text [28]. This skipping process may skip to the next paragraph, page, or section, depending on the reader's prediction of the value of the current local area. Importantly, this paper points out that most literature investigating skimming uses very short texts, usually less than 500 words. As such, there is a dearth of research on behaviors in long documents. We wish to note that this correspondingly means that there may be a lack of research into scanning behaviors, as scanning requires a certain length of text.

In a 2004 paper, Ohno used an eye tracker to trace a reader's gaze in long (chapter length) documents [56]. Their contribution consists of two major parts—an interactive reading experience and an observational description of readers' behaviors while interacting with the text. The former part will be described further in Section 3.2.2. The latter notably includes a three-way categorization between reading, skimming, and scanning. They describe scanning as the fastest process used to acquire information; physically, the fixations and saccades proceed mostly vertically, with gaze data usually appearing only once or twice in a given line. An example is visible in Figure 1.

Other studies by Chanijani et al. and by Gwizdka (sourced from human participants) show similar patterns for reading, skimming, and scanning behaviors [23, 36]. Visual examples of these behaviors, taken from real-world data, are available in Figure 2 and Figure 3.

Another work investigating different behaviors over the course of a single task is Symons and Pressley's 1993 paper; this paper is also noteworthy as one of the few papers investigating reading behaviors over long texts. They track behaviors during *text search* [70]. Text search, as defined in this work, is a task involving reading long texts to locate specific information. Specifically, the task under investigation was finding low-inference information (directly stated with no critical thinking required once located) in psychology and earth sciences textbooks. The authors find that readers separate their high-level goal into smaller discrete actions; they repeatedly transfer back and forth between local reading

A: Reading

The plane took off from Santa Cruz, Bolivia and crashed near the airport in Medellín, Colombia on 28 November. Only six of the 77 passengers on board survived. The dead included 19 members of the Chapecoense soccer club from southern Brazil and 20 of the journalists covering the team.

B: Skimming

The plane took off from Santa Cruz, Bolivia and crashed near the airport in Medellín, Colombia on 28 November. Only six of the 77 passengers on board survived. The dead included 19 members of the Chapecoense soccer club from southern Brazil and 20 of the journalists covering the team.

C: Scanning

The plane took off from Santa Cruz, Bolivia and crashed near the airport in Medellín, Colombia on 28 November. Only six of the 77 passengers on board survived. The dead included 19 members of the Chapecoense soccer club from southern Brazil and 20 of the journalists covering the team.

Figure 2: A real-world example of reading, skimming, and scanning, from Chanijani et al. [23]. This example is sourced from readers who were fluent in English as a second language.

for comprehension and global searching for relevant areas. It should be noted that this study used human observers, rather than eye trackers, to investigate reader actions; we focus on this study to introduce the task of text search, and reinforce that multiple moment-to-moment reading behaviors can be found in a single high-level task.

A paper by Cole et al. further investigates reading behaviors in text search tasks [25]. They find that all readers switch back and forth between reading and scanning behaviors, with the switch predominantly taking place at the end of text segments. However, they also find that individuals significantly differ in the probability and frequency of these transitions; some readers are overall biased towards scanning, only switching to reading for particularly relevant segments, while others spend far more of their time in reading. This bias was consistent for a given participant across multiple search tasks concerning different domains. Interestingly, no significant correlation was found between this bias and the participant's cognitive ability, prior domain knowledge on either task, or overall task performance.

Liang and Huang's 2014 paper investigates the diversity of reading patterns in elementary school students over long reading tasks of 15, 30, or 45 minutes [53]. They find that their sample of students divides into roughly two categories, which they call fluctuant readers and coherent readers. Fluctuant readers switch between reading, skimming, and scanning behaviors, while coherent readers predominantly stay in a reading state. They found no significant difference in information retrieval between the two reading groups, reinforcing that skimming and scanning are not strictly inferior versions of reading.

While we wish to highlight that skimming and scanning can be adaptive, it is also important to note that they are not strictly *better* versions of reading. The idea of *speed reading*, or reading at an increased speed with no corresponding loss of comprehension, has been thoroughly shown to not exist [62]. Speed reading is, in fact, skimming or scanning.

It should be noted that this is not an exhaustive list of observed reading behaviors. Other behaviors include spell checking, as in Strukelj and Niehorster, "barking" (reading text in a second language without enough proficiency to perform sentential integration or semantic encoding), as seen in Beelders and Stott, and more [9, 68]. However, this thesis will mainly focus on three behaviors: reading, skimming, and scanning.

READING AUGMENTATIONS

In this chapter we present a literature review on *reading augmentations*, computer-based techniques for supporting or improving the process of reading. This definition (which intentionally encompasses a wide purview) includes techniques and technologies that aim to support the diverse goals and tasks related to reading, ranging from productivity increases for speed readers to interventions supporting subjective fun and pleasure.

3.1 TEXT LAYOUT METHODOLOGIES

We begin with an overview of *text layout* methodologies. Traditional reading consists of the reader tracing their eyes over static, rectangular text blocks with relatively uniform typography; this is the traditional text layout. Alternative text layouts investigate the possibilities of altering this paradigm. We roughly separate these alternative text layouts into two categories: disruptive interventions, which replace the existing reading process entirely, and non-disruptive interventions, which augment and support the existing reading process.

3.1.1 *Disruptive Interventions*

In this section we overview the most popular techniques that are united in their attempt to *replace* the existing reading process, usually by reducing or eliminating the requirement for saccades. Any technique that replaces the usual reading process (saccades and fixations between fixed-position lines of text) would fall under this definition, but in practice these techniques are closely related: the most popular disruptive interventions, and perhaps the most popular reading augmentations in general, are a family of techniques that aim to improve reading speed without a corresponding loss of comprehension.

The text display mode known as Rapid Serial Visual Presentation (RSVP) is a common technique used in a wide variety of speed reading software, both commercial and free to use. Although it was originally invented in 1959 as a scientific method for investigating the role of saccades in reading, the technique was later reused as a method claimed to allow for significantly faster speed reading [34]. Although its proponents hold that products that implement RSVP

reduce visual fatigue and increase speed without a corresponding decrease in comprehension, non-affiliated studies have instead found that reading comprehension is decreased significantly, just the same as if the reader was skimming or scanning normally [11]. The technique could even potentially be harmful to the reader's health; eye blinks during RSVP were found to be reduced to such an extent that the researchers warn it could contribute to dry eye syndrome and visual fatigue.

A key limitation of RSVP is its inability to adjust to suit the reader's information processing ability. As such, we find it important to highlight an intriguing recent work by Kosch et al. which investigates the feasibility of using electroencephalography (EEG) to calibrate text alignment and presentation speed in RSVP [49]. They find that EEG data is strongly correlated with reading speed, subjective workload, and text comprehension, and that there are parameters of RSVP that will overall maximize these metrics. It is important to note that this study does not yet attempt to optimize RSVP parameters, nor does it claim that RSVP with optimized parameters has been shown to be superior to traditional reading. Nonetheless, we find this work an important step towards determining whether RSVP can be made beneficial, and a reflection of the critical importance of reader-specific calibration for supporting the wide diversity of reading styles.

A technique known as Times Square aims to minimize saccades by scrolling the text right-to-left over time (the name, of course, being derived from the most famous implementation of this technique). As the reader advances left-to-right, the text moves in the opposite direction, with the reader's gaze overall staying relatively fixed. Experimental evidence is overall mixed; while some studies report low comprehension without an increase in reading speed, more recent studies show the technique functioning about as well as RSVP [35, 44, 67]. However, even the most positive studies do not show Times Square as performing *better* than RSVP, a technique which has been shown to decrease reading comprehension.

A third technique by Kawashima et al. is inspired by both RSVP and Times Square, but remains distinct from both. The technique continually raises a single word of text slightly above the baseline, with the selection of word continually moving forward at a set reading speed [45]. The effect is thus to attempt to constrain the reader to a set reading speed as they follow the raised text. The authors claim that their technique increases reading speed with no decrease in comprehension, but it should be noted that their data presentation does not include basic elements such as error bars or hypothesis testing. They also have not been experimentally replicated by a third party. Because the technique works by forcibly applying a set reading speed without pausing for blinks, it is possible

that the same risks would apply to it as RSVP. Visual fatigue measurements were not reported in this work.

Overall, we are not aware of any intrusive interventions that have been reliably shown by trustworthy experiments to enhance any single aspect of reading performance without a corresponding decrease to another aspect. For example, calibrating an RSVP display for high WPM may increase reading speed, but this speed necessarily comes with a corresponding decrease in comprehension. We humorously call this the *no free lunch* theory of reading. While we give it a name here, we are by no means the only researchers to have made this observation [62].

This result is not surprising according to rauding theory; according to Carver, natural reading at an intuitive pace already maximizes the information processing that a particular reader is capable of [22]. Because the limitation preventing most readers from reading at a faster rate is not found in the physical movements of the eyes but rather in the brain, interventions like RSVP may be poorly targeted. Instead of attempting (apparently in vain) to increase the information processing capability of the human brain, it may be more advantageous to investigate interventions that allow a reader to accomplish the same task by processing less information.

3.1.2 *Typographical Cuing*

Typographical cuing is an umbrella term for reading augmentations that draw attention to certain areas of text using the typographical properties of that text; examples include bold face, italics, font weight, underlining, and background coloring [32, 38]. In this section, we review the use of typographical cuing, specifically focusing on typographical cuing as a reading augmentation.

Typographical cuing increases comprehension of cued sentences and decreases comprehension of non-cued sentences, with no overall change in the comprehension of the entire text [29, 32]. Similarly, typographical cuing does not increase the total amount of material learned while studying, but does increase the relevance of learned material [38]. Light amounts of typographical cuing (5-10% of the words in a text being cued) perform better at directing attention than heavy amounts (50%), which may be due to an increasing cognitive load caused by more cuing or more complex text layouts [54]. Typographical cuing is only effective at directing attention when applied before reading the text, reinforcing that its key benefit lies only in *directing* information processing and not *improving* information processing.

Importantly, reading augmentations using typographical cuing allow the reader to choose their own reading pace and direction, instead of attempting to force a

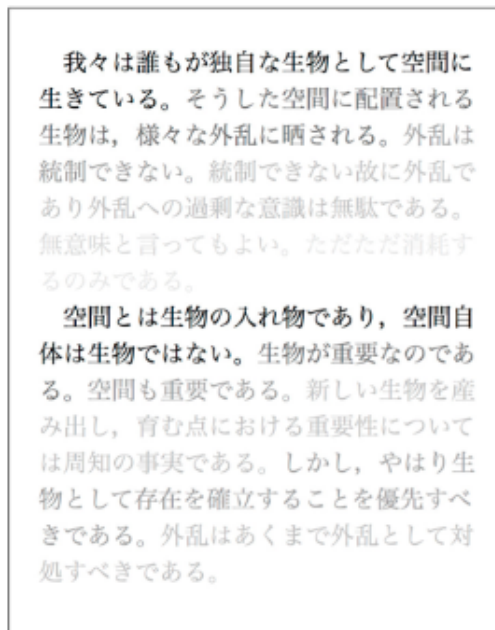


Figure 4: Typographical cuing used as a reading augmentation, by Kobayashi et al. [48].

set reading style upon them. Additionally, when done properly they can draw the reader's attention to more relevant areas of text and away from less relevant areas. Because the possibility of increasing reading performance with no tradeoff seems doubtful, typographical cuing stands out as an interesting area because it may allow a reader to overall perform more efficiently without necessitating a higher overall rate of information processing.

With this overview of typographical cuing complete, we present the most relevant computerized reading augmentations based on typographical cuing.

Kobayashi et al. present a text layout that sequentially fades out the text, sentence-by-sentence, starting from each paragraph of the passage [48]. Accordingly, the early sentences in a paragraph, including the initial topic sentence in most traditional expository paragraphs, are more visually salient. They found that this formatting increased overall reading comprehension on a task that required the user to identify key passages and ideas from a text. Additionally, it led to slower reading speeds and better recognition rates for the most important sentences in paragraphs. Correspondingly, less time and attention was spent on the least important sentences. A visual example can be seen in Figure 4.

Interest exists across both academic and non-academic sources in the area of drawing the reader's attention to *content words*. Content words (sometimes called *lexical words*) are those from the parts of speech that carry the most meaning in a sentence: verbs, nouns, adjectives, and adverbs. Content words are separate from *function words* (sometimes called *grammatical words*), which are articles, pronouns, conjunctions, and prepositions. Focusing on content words is a common speed

reading technique; even some non-academic sources have recommended the use of typographical cuing to increase the emphasis of content words [1].

However, the only academic source we are aware of to investigate typographical cuing for content words is Biedert et al.'s work concerning an augmentation named QuickSkim [12]. They describe a continuous and continually changing boldface degree based on the unigram frequency of each word (a metric by which content words would be far more bolded than function words) and the reader's current reading behavior. QuickSkim, alongside other augmentations, was shown to 16 users in a preliminary study; user feedback was described as quite positive, but unfortunately the QuickSkim augmentation was not thoroughly studied in isolation. Furthermore, we are unaware of any future work that followed up on these preliminary investigations into QuickSkim.

Overall, we are unaware of any reliable evidence for tradeoff-free improvements to reading speed or comprehension from typographical cuing, as predicted by the *no free lunch* theory of reading. This theory does not mean that reading augmentations are useless; we have also shown that interventions like typographical cuing can direct attention away from irrelevant and towards relevant sections, improving overall performance in text search tasks.

Interestingly, the augmentation described in the Text 2.0 work is slightly different from what is shown in the demo video on the Text 2.0 website - it is unknown which version was used in the demos for preliminary feedback [2]. More details on the differences can be found in chapter 5.

3.2 ATTENTIVE DOCUMENTS

We close our review of prior literature by describing a series of works that aim to create *attentive documents*, or documents that are responsive to the reading process. We first describe how eye trackers can be used to glean information about the user's current reading behavior and mental state, and then proceed to the augmentations and interventions that can be accomplished using this information. We place special importance on the work of German researchers Ralf Biedert and Georg Buscher, whose seminal works have directly inspired the present thesis.

3.2.1 Reading-Skimming Detection

In 2001, Cambell and Maglio created an algorithm for detecting reading using a video eye tracker. Their simple heuristics-based algorithm served as a simple reading detector and could not yet differentiate skimming, but this is the first work we are aware of that identified a use case for reading or skimming detection. While discussing the implications and future work, Campbell and Maglio recommend future researchers attempt both skimming detection and scanning detection.

Buscher et al. created a simple heuristics-based algorithm to detect reading using only an eye tracker [18]. The algorithm classified saccades into a list of categories based on the distance and direction of the saccade. Each category of saccade then contributed a set number of points to two detectors, one for reading and the other for skimming. Limitations of this early work include a lack of experimental validation and an inability to calibrate the algorithm to account for individual differences in eye movements. However, it was a seminal work in reading detection.

Together, these early works by Cambell and Maglio, along with Buscher et al., present the possibility of attentive documents which are responsive to reading behaviors. However, because these works were but preliminary investigations, both works do not attempt to create or evaluate a reading augmentation based on reading detection or skimming detection. These results inspire a small subfield of machine learning researchers to create models with better performance on these tasks. Somewhat surprisingly, most of these researchers did not connect their models to any particular use case for reading or skimming detection. As such, we will briefly review the most relevant machine learning models for reading behavior classification, before returning to a review of the research on attentive documents built on this category of classifier.

A later work by Buscher et al. advanced the field of reading behavior classification by training a machine-learning-based classifier on approximately 1400 saccades taken from 12 users [15]. Saccades were manually classified as skimming or reading by two human judges, and an SVM classifier was trained on those labels. The classifier achieved an overall accuracy of 88% on the test set under the optimal window size. This window size, representing the context window of previous saccades that the model has access to, achieved optimal performance at 3 saccades. Under this context window, the model has a response time of approximately 750 ms on average; equivalently, it “forgets” about prior saccades after less than a second.

Several other works have presented machine learning models for the task of reading behavior classification. Works with lower relevance to the present thesis will be briefly described in Table 1.

We have focused so far on the area of reading behavior classification; a related yet higher-level task is that of inferring reading comprehension from scanpaths. Reich et al. constructed a neural network that estimated reading comprehension and subjective text difficulty [64]. Most relevant for our work, their evaluation process did not simply aggregate a test set by leaving out random data on the saccade level, but instead experimented with leaving out single pages, books, or readers to generate the test set. They found that leaving out readers led to the worst accuracy levels, suggesting that machine learning models tend to focus

Authors	Task	Overall Accuracy	Notes
Chen and Srivastava	Reading and skimming detection	82%	Performs reading and skimming classification for both desktop and mobile reading. Both modes used a remote eye tracker. [24]
Ishimaru et al.	Reading detection	74%	Performs reading detection using electrooculography instead of eye trackers, with the advantage of being comfortably wearable. [41]
Islam et al.	Reading detection in English and Japanese	93%	Four-way classification between reading English, reading Japanese vertically, reading Japanese horizontally, and not reading. Uses self-supervised learning to classify without manual labeling. [42]
Kelton et al.	Reading and skimming detection	72-95%	Performs both local (second-to-second) and global (article-length) classification. 82.5% global classification accuracy, with local accuracy varying based on the model's degree of fine-tuning for global classification. [46]
Landsmann	Reading detection	93%	Uses a fully-supervised algorithm to create a binary decision tree. Labels may be inaccurate due to a reliance on participants to accurately report their own reading behavior. [51]

Table 1: Selected machine learning models performing reading behavior classification.

on reader-specific patterns. One possible implication of this fact is that reading augmentations dependent on machine learning models for reading detection may require a calibration process for each reader, or else suffer from reduced accuracy. Although this result was not the main focus of the work, we highlight this finding because it was influential for the design of our system.

Finally, we highlight the work of Chanijani et al. This work makes the noteworthy choice to separate out scanning from skimming, and develop both a generative model and RNN-based classifier for this three-way categorization [23]. This model is the only one we are aware of that makes this three-way division. They find an overall classification accuracy of approximately 95%. Interestingly, they found the highest degree of confusion was between the skimming and scanning, indicating that these two may be difficult to differentiate even for modern machine learning models. We found this work intriguing because of its decision to include scanning in its classification; despite the seminal 2001 work by Cambell and Maglio explicitly naming both skimming and scanning detection as compelling future research directions, scanning detection was otherwise ignored until this work [19]. Unfortunately, this work also does not provide access to its code, model weights, or sufficient detail to its training process to reliably replicate.

Overall, the past decade has seen a moderate amount of work in developing new and improved machine learning models for reading behavior classification. In the next section, we will review the existing literature for the *use cases* for these models.

3.2.2 *Attentive Documents*

As previously defined, we refer to attentive documents as documents that are responsive to the reading process, particularly by being gaze-aware. We close our literature review by describing the extant techniques in this space.

As mentioned in Section 2.2.1, Ohno created an interactive document display system that responded to the reader's gaze [56]. Specifically, their text display changed the background color elements of text once the reader had thoroughly read them (a form of typographical cuing). This color would deepen as the element was re-read multiple times, with the intent of allowing the reader to remember and intuitively understand when they had skipped or repetitively re-read areas of text. They found that their system did not significantly improve the overall comprehension of the entire text. However, it did increase the readers' precision (percentage of text read that was relevant) in a text search task. (It should be noted that this result is consistent with the *no free lunch* theory of reading.)

Aside from this early work, only one team has significantly explored the area of gaze-aware, attentive documents. Over a three-year period from 2009 to 2012, German researchers Ralf Biedert and Georg Buscher, alongside their advisor Andreas Dengel and other collaborators, developed and presented a series of seminal publications. In these works, they create a series of bodies of text that are responsive to the reading process. These researchers and their work are vital to the development of the present thesis.

In 2009, Buscher, Biedert, and Dengel created the eyeBook [13]. Designed to be an immersive and interactive "reading experience", the project implemented gaze-based interactive features, such as multimedia display triggered by reading specific phrases. They present copies of *The Little Prince* and *Dracula* that play thematically appropriate music, display images representing the text upon reading certain phrases, and change the background color theme to match the current setting of the story. An example is given in Figure 5. They also describe an algorithm identical to the one presented in their 2008 work for reading detection and skimming detection [18]. While this algorithm is not yet used in their work, the researchers make it clear that they wish to make their work responsive to the text's relevance as revealed through the user's skimming behavior.

The same authors expanded on their work by creating the *Text 2.0* framework [12]. This work, described in more detail in a previous section, includes more technical implementation details concerning the eyeBook. They also describe the *QuickSkim* text layout.

Finally, their 2012 work *Attentive Documents* gives this section its title [17]. In it, participants read documents that were either relevant or irrelevant. Several

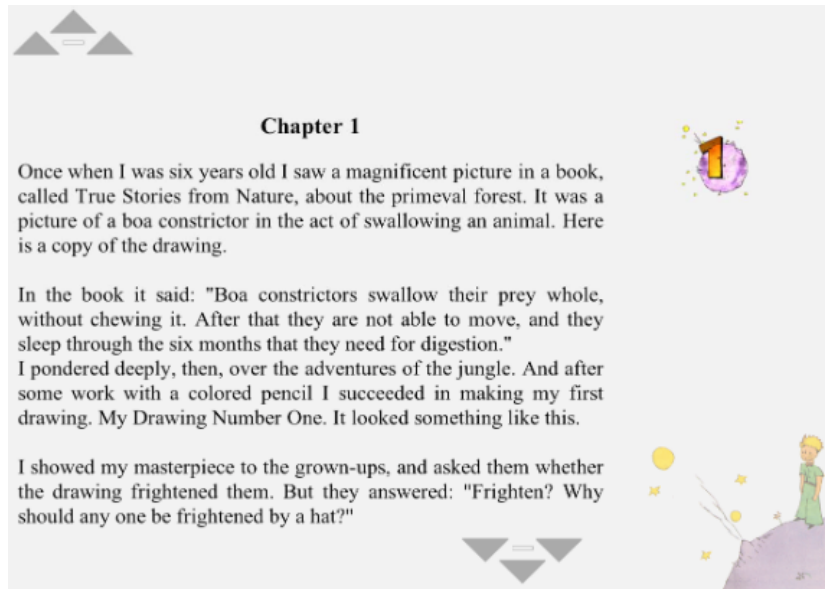


Figure 5: An excerpt from *The Little Prince* modified using the Text 2.0 framework [12, 13]. The visible images appear in response to the reader's gaze fixating on certain passages of text. A video showing other realtime effects, including audio effects not possible to include in text, can be seen at the authors' website [2].

eye tracking metrics, including mean forward saccade length, thorough reading ratio, coherently read text length, were measured. Finally, two additional metrics were included in personalized form: regression ratio and average fixation duration. The researchers found that these two metrics significantly correlated with relevance, but *only* after being personalized. The personalization process, as described in the work, has a noteworthy limitation in that it must be done post-facto using a user's complete corpus of gaze metrics on all documents. That said, this work reinforces the importance of personalizing gaze metrics to best understand a reader's reading behavior.

The work closes by outlining a detailed path for promising future research into attentive documents, including ideas for supporting web search, text search, providing support for resuming reading after a loss of attention, and supporting the scanning of long documents.

Unfortunately, this positive outlook on the future of attentive documents has not yet come to pass. Neither researcher followed up on their work concerning attentive documents; this gap is not due to a lack of compelling research directions, but rather because Biedert and Buscher both transitioned from academia to industry. Although their academic works are frequently cited as a compelling reason to create machine learning models that perform reading detection and skimming detection [23, 24, 46], we are aware of no works that aim to replicate, experimentally verify, or expand on their work. In other words, the past decade

has seen significant amounts of publications aiming to create these models, but almost no research into *using* them.

It is the goal of the present thesis to continue the work started by Biedert and Buscher. We aim to expand their work on attentive documents, and provide an empirical evaluation of a gaze-aware reading system where Biedert and Buscher did not.

A REALTIME READ-SKIM-SCAN CLASSIFIER

We present a novel classifier that takes in a real-time stream of gaze data and performs a three-way classification into reading, skimming, or scanning. The model recognizes fixations and saccades from the raw gaze data, then applies a set of rules-based heuristics to determine which reading behavior is most consistent with the user’s current gaze data. We close by comparing our classifier and its design choices to other existing models.

4.1 MODEL DESIGN

The core idea of the classifier is similar to Buscher, Dengel, and van Elst’s earlier work [18]. First, fixations are detected from gaze data. Second, the saccades between these fixations are classified. Third, scores associated with these classifications are accumulated. Finally, the category with the highest point value is returned.

The classifier takes as input a realtime stream of gaze coordinates. We then use the same algorithm described by Buscher, Dengel, and van Elst to detect new fixations in our realtime gaze data [18].

For each gaze point received in the realtime stream sent by the eye tracker, we apply a main loop, briefly described in pseudocode in Algorithm 1.

We now describe the functions used in Algorithm 1 in more detail. In these sections, we will at times use constant values (for example, a certain distance in pixels). Unless otherwise mentioned, we use the same constants as in Buscher, Dengel, and van Elst’s work [18]. Some minor differences exist for the sake of hardware compatibility, as described further in Section 4.2.

In `checkForNewFixation()`, we maintain a sliding window of three sequential gaze locations. A new fixation is detected if all of the past three gaze locations were within a rectangle of 30 x 30 pixels. If a new fixation is detected, its position is returned for use later in the algorithm.

In `isCurrentFixationEnded()`, we check if the newly received gaze point is consistent with our positional estimate of the user’s current fixation. We do so by comparing the x , y coordinates of the new gaze point to the center of the current fixation. To add tolerance for microsaccades and drift, it is beneficial to increase the allowed area (from its previous 30 x 30 pixel rectangle) once a fixation has been found. As such, the set of all points making up this fixation are

Algorithm 1 Fixation Detection and Classification

```

1: procedure MAINLOOP
2:    $x, y \leftarrow$  a non-negative integer gaze position
3:    $currentFixation \leftarrow$  the position of the current fixation, or Null
4:   if  $currentFixation = \text{Null}$  then
5:      $newFixation \leftarrow \text{checkForNewFixation}(x, y)$ 
6:   else
7:      $newFixation \leftarrow \text{isCurrentFixationEnded}(x, y)$ 
8:   end if
9:   if  $newFixation$  then
10:     $transitionType \leftarrow \text{classifyTransition}(newFixation)$ 
11:     $\text{updateDetectors}(transitionType)$ 
12:  end if
13: end procedure

```

required to fit in a 50 x 50 pixel rectangle. If this new gaze point cannot fit in a 50 x 50 pixel rectangle with the set of previously observed points that make up the current fixation, it is declared inconsistent with the current fixation. To prevent fixations being erroneously ended due to outliers or transient equipment error, three inconsistent gaze points in a row are required to end the current fixation. At this point, these three gaze points may create a new fixation.

When a new fixation is found, we classify the transition between the last fixation and the new fixation in `classifyTransition()`. We use a simple rules-based classification according to the amplitude and direction of the saccade, measured according to the change in x and y . The list of rules for this classification can be seen in Algorithm 2. Each classification represents a commonly seen movement during reading, with the exception of UNCLASSIFIED_MOVE, a category that serves as a catch-all for movements that are highly uncharacteristic for reading (for example, a horizontal saccade significantly longer than the width of the text). A static constant representing the classification category of the previous saccade is returned to be used in the rest of the algorithm.

Finally, the previous saccade's classification is used in `updateDetectors()`. Each classification provides an amount of positive or negative evidence for reading, skimming, and scanning. To represent this, we maintain three *detectors*, one for each behavior, which maintains a point value for their corresponding behavior. The definitions and point values of the saccade classifications are given in Table 2. Finally, these detectors are continually compared to each other; whichever detector has the highest score determines the current classification.

The above pseudocode algorithms and descriptions present a simplified but essentially accurate portrait of our real-time classifier. However, some additional modifiers and factors provide additional influence on the precise score totals.

Algorithm 2 Saccade Classification

```

1: procedure CLASSIFY_SACCADE
2:   changeX  $\leftarrow$  the change in horizontal position, measured in letter widths
3:   changeY  $\leftarrow$  the change in vertical position, measured in line heights
4:   if abs(changeY) > 2.5 then
5:     return VERTICAL_JUMP
6:   else if 0 < changeX  $\leq$  personalizedSkimBoundary then
7:     return READ_FORWARD
8:   else if personalizedSkimBoundary < changeX  $\leq$  21 then
9:     return SKIM_FORWARD
10:  else if 21 < changeX  $\leq$  66 then
11:    return LONG_SKIM_JUMP
12:  else if -6  $\leq$  changeX < 0 then
13:    return SHORT_REGRESSION
14:  else if -16  $\leq$  changeX < -6 then
15:    return LONG_REGRESSION
16:  else if changeX < -16 and changeY > 0.6 then
17:    return RESET_JUMP
18:  else
19:    return UNCLASSIFIED_MOVE
20:  end if
21: end procedure

```

In addition to the saccade classifications, each detector's point value is modified continually. Every time the eye tracker provides a position update (therefore, every $1/33$ seconds), each detector's point value is multiplied by 0.993. The overall point values are thus subject to exponential decay; after a second, each score will decay to approximately 0.793 times its previous value. These constants were derived by iterative design to create a subjectively appropriate time-response curve. This allows for a non-fixed context window that does not instantly forget strong signals of user behavior while still being mostly determined by the most recent second or two.

When one detector's point value overtakes another, their point values will necessarily be close to each other. This can create a behavior wherein mode transitions are undone immediately after they occur, which is undesirable as our model is designed to trigger user-visible behavior as a result of mode transitions. Rapid flickering between modes would increase cognitive load and cause frustration, a behavior we refer to as *thrashing* by metaphor to the problem of layout thrashing in web development. We apply two methods to reduce this behavior: a one-time multiplicative bonus and a permanent constant bonus.

As a one-time bonus, the new higher detector has its current point value (which is realistically always positive) multiplied by 1.3x on a mode transition.

This one-time bonus quickly decays away, but significantly decreases the likelihood that the mode will transition over the next 1–3 seconds.

Additionally, a permanent bonus is added to the current winning detector: for as long as a mode is active, it has a virtual 10 points added to its total for the purposes of deciding whether other detectors will overtake it. This acts as *hysteresis*: mode shifts are dependent on their own past history. To avoid artificially penalizing the current winning detector, these virtual points are not used when calculating the detector’s exponential decay.

Both the one-time multiplicative bonus and the permanent constant bonus reduce the likelihood of frequent mode switches in ambiguous or quickly-changing scenarios, without preventing mode switches in response to large, unambiguous behavioral patterns.

To provide a concrete example, we analyze the point values of a hypothetical case where the classifier switches from reading to skimming. The scores for the reading, skimming, and scanning detectors are currently at 85, 83, and 20, respectively. At this point, a saccade of type `LONG_SKIM_JUMP` is detected. According to the point values seen in Table 2, the scores are updated by -5, 8, and 5 to a new total of 80, 91, and 25. Even with the 10 point bonus for being the winning detector, the reading detector has been overtaken by the skimming detector. The skimming detector then becomes the new winning detector. Its score is multiplied by 1.3 to become 118.3. Thereafter, for another classifier to win it would have to overtake this score by 10.

4.1.1 *Reading and Skimming*

Because our core algorithm is similar to Buscher, Dengel, and van Elst’s earlier classifier, our point values for the reading and skimming detectors are identical to their constants. The sole exception to this is that we set a personalized boundary between the `READ_FORWARD` and `SKIM_FORWARD` classifications. Because these represent forward horizontal saccades, this personalization effectively personalizes the average forward saccade length, an eye gaze metric that is known to vary heavily between users [17, 58, 59].

We describe a simple personalization technique that can be used to personalize the read-skim boundary suitable for real-time HCI interventions. We modify Buscher, Dengel, and van Elst’s reading detection algorithm to allow for personalization of average forward saccade length [18]. We do so by asking the user to attempt to skim approximately 2 pages of text, and thoroughly read approximately 1 page of text. We then treat each corpus of eye movements for these pages as a ground truth for skimming and reading behavior, respectively. While

Saccade type	Reading detector score	Skimming detector score	Scanning detector score
READ_FORWARD	10	5	0
SKIM_FORWARD	5	10	5
LONG_SKIM_JUMP	-5	8	5
SHORT_REGRESSION	-5	-5	-8
LONG_REGRESSION	-5	-3	5
RESET_JUMP	5	5	-5
VERTICAL_JUMP	0	0	1 per 6.5 pixels
UNCLASSIFIED_MOVE	0	0	5
Scroll event	0	0	1 per 40 pixels

Table 2: Point update amounts for each possible event. The scores for each detector are incremented or decremented according to the classification of the saccade. While a scroll event is not a saccade classification, it is included in this table for ease of reference for point values.

the user reads these texts, each saccade is classified according to the schema previously described in Algorithm 2. Any saccade that is predominantly horizontal to the right, with at most a small vertical component (READ_FORWARD, SKIM_FORWARD, and LONG_SKIM_JUMP) are saved. At the end of the calibration process, we take the mean of the average forward saccade distance of these saccades. That mean is set as the new boundary between the READ_FORWARD and SKIM_FORWARD classifications. The results of this personalization process are described in Section 6.2.2.

4.1.2 Scanning

Because our core algorithm is similar to Buscher, Dengel, and van Elst’s earlier classifier, our point values for the reading and skimming detectors are identical to their constants. However, their classifier is a two-way classifier between reading and skimming. To implement our scanning detector, we must determine its point value constants for each saccade classification.

The point values, visible in Table 2, were motivated by the literature on scanning behavior [22, 25, 53, 63]. Their precise values were adjusted over the course of an iterative process wherein initial prototypes were continually tested in pilot user studies, then modified based on user feedback.

As visible in Table 2, the scanning detector change for the VERTICAL_JUMP classification does not have a single constant value. This is because the degree of evidence gained for scanning behavior from a vertical jump scales strongly with the amount of text skipped [17]. The scanning detector is incremented by 1 point per 6.5 pixels (approximately 6 points per line). Because a classification of VERTICAL_JUMP requires a minimum of 2.5 line heights of vertical change, the increment will therefore have a minimum value of 15 points.

While the reading and skimming detectors increase point values solely through points added from saccades, the scanning detector additionally uses scroll behavior as implicit feedback for scanning behavior. The scanning detector acquires points each time the window viewport is scrolled in any direction, at a rate of 1 point per 40 pixels (approximately one point per line). Note that this rate is significantly smaller than the previous score for the `VERTICAL_JUMP` classification, as scrolling is more common when scanning but still occurs during skimming and reading. This point source is meant to capture the case where a user quickly scrolls through text without significant eye movement; this behavior is strongly associated with scanning but is not directly captured in gaze data. In practice these points have a relatively small effect on the behavior of the model, except for when the user scrolls more than a single page in a few seconds.

4.2 COMPARISON TO EXISTING MODELS

We have made the noteworthy choice to implement our classifier using a rules-based model similar to the earliest work in reading detection like Campbell and Maglio did in 2001 [19] or Buscher and Biedert did in 2008 [18]. This choice is opposed by the clear alternative of implementing a machine learning based model. We chose to implement a simpler rules-based model for two key reasons: lack of reproducibility in the most relevant machine learning works, and a possibility to avoid thrashing.

Concerning reproducibility, we would need to build off of research by Chaniyani et al. to build a three-way reading-skimming-scanning classifier, as it is the only work we are aware of to include scanning as a separate category [23]. However, this work did not include their model weights, or even sufficient detail to recreate their training process. Reproducing their research would be possible, but it would necessitate starting almost entirely from scratch; doing so while also implementing our text layout described in Chapter 5 and experimentally validating it as described in Chapter 6 would be beyond the scope of a Master's thesis.

Concerning thrashing, we found it beneficial to use a rules-based model because it allows us to easily and reliably decrease the frequency of thrashing due to the implementation of hysteresis. Machine learning models will cause thrashing in ambiguous behavior because it leads to the highest classification accuracy, but that accuracy comes with a real-world cost to user experience when it repeatedly triggers user-visible behavior. It is clearly possible to implement a new machine learning model with penalties to mode shifts that would reduce thrash-

ing; however, we again believe it to be beyond the scope of this thesis. For these reasons, we chose to implement a rules-based model for our classifier.

In comparison to the rules-based model we based our classifier on, our fixation detection algorithm requires three sequential gaze locations, as opposed to the four required by Buscher, Dengel, and van Elst [18]. This is because our eye tracker, the Tobii 5, has a 33 hz refresh rate, compared to their 50 hz. As mentioned in their work, four gaze locations at 50 hz represents approximately 80 ms; our three gaze points at 33 hz represent approximately 90 ms, as close as is possible on our hardware.

4.2.1 *Personalization*

Buscher et al. describe a method for personalizing two metrics of eye gaze: regression ratio and average fixation duration [17]. That work uses the existing heuristics described in their earlier work to detect and classify reading and skimming behaviors [18]. However, this algorithm is not personalized with respect to its input gaze metrics; for example, the border between a “read forward” saccade and a “skim forward” saccade is universally set at 8 character lengths. This is despite it being known that average forward saccade length varies heavily between users [17, 58, 59].

Rather than personalizing the forward saccade length which is known to vary heavily, they instead personalize the read-skim percentage by normalizing the raw percentages to the user’s observed maximum and minimum thoroughly-read-text percentage. This personalization is not mentioned in their 2012 work and it is unclear if it was applied in that case. This technique has some potential flaws related to ceiling and floor effects (i.e., it is unclear what should happen if a user’s variation causes the algorithm to believe they read near 100% or near 0% of all text).

Regardless of that personalization technique’s efficacy, it cannot be used to trigger real-time user-facing behavior. This is because the technique is post-facto: it requires a complete corpus of existing eye gaze data. Additionally, normalizing percentages only after the read-skim classification means that the algorithm is no longer classifying any individual point in time as either reading or skimming. As such, we deviate from their personalization technique, instead using the one we describe in Section 4.1.1.

IMPULSE READING: ATTENTIVE TEXT LAYOUT

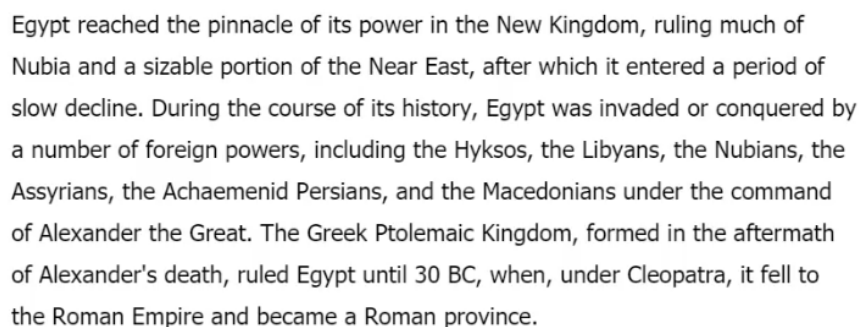
We describe our design and implementation of Impulse Reading, a reading augmentation that facilitates the diverse goals of reading by switching between multiple text layout methods based on the reader's current task and goal. Our software captures the reader's gaze using an eye tracker and uses it to estimate their behavior as either reading, skimming, or scanning. Each of these three behaviors is associated with a specific text layout method that aims to support the unique traits of that behavior.

5.1 DESIGN OVERVIEW

We begin with an overview of the features of Impulse Reading. Most critically, we describe the text layouts we use as the core of our reading augmentation; other design details include our typographical choices and a brief summary of the technical design.

5.1.1 *Text Layout Modes*

The first mode is simply plaintext, which is the text layout when the user is reading text thoroughly. Plaintext was chosen in this case because we wanted a non-disruptive text layout that would not interfere with the participant's presumably well-practiced process of reading text thoroughly.

A rectangular box containing a paragraph of text in a standard serif font, representing the plaintext layout mode. The text is left-aligned and occupies the width of the box.

Egypt reached the pinnacle of its power in the New Kingdom, ruling much of Nubia and a sizable portion of the Near East, after which it entered a period of slow decline. During the course of its history, Egypt was invaded or conquered by a number of foreign powers, including the Hyksos, the Libyans, the Nubians, the Assyrians, the Achaemenid Persians, and the Macedonians under the command of Alexander the Great. The Greek Ptolemaic Kingdom, formed in the aftermath of Alexander's death, ruled Egypt until 30 BC, when, under Cleopatra, it fell to the Roman Empire and became a Roman province.

Figure 6: The plaintext text layout, meant to support a reader who is reading thoroughly. Note that no additional typographical cuing has been applied in this layout.

The Nile has been the lifeline of its region for much of human history. The fertile floodplain of the Nile gave humans the opportunity to develop a settled agricultural economy and a more sophisticated, centralized society that became a cornerstone in the history of human civilization. Nomadic modern human hunter-gatherers began living in the Nile valley through the end of the Middle Pleistocene some 120,000 years ago. By the late Paleolithic period, the arid climate of Northern Africa had become increasingly hot and dry, forcing the populations of the area to concentrate along the river region.

Figure 7: The content words layout, meant to support a reader who is skimming. Function words like prepositions and articles are decreased in contrast, while content words like verbs and nouns remain at full contrast.

By about 5500 BC, small tribes living in the Nile valley had developed into a series of cultures demonstrating firm control of agriculture and animal husbandry, and identifiable by their pottery and personal items, such as combs, bracelets, and beads. The largest of these early cultures in upper (Southern) Egypt was the Badarian culture, which probably originated in the Western Desert; it was known for its high-quality ceramics, stone tools, and its use of copper.

The Badari was followed by the Naqada culture: the Amratian (Naqada I), the Gerzeh (Naqada II), and Semainean (Naqada III). [page needed] These brought a number of technological improvements. As early as the Naqada I Period, predynastic Egyptians imported obsidian from Ethiopia, used to shape blades and other objects from flakes. In Naqada II times, early evidence exists of contact with the Near East, particularly Canaan and the Byblos coast. Over a period of about 1,000 years, the Naqada culture developed from a few small farming communities into a powerful civilization whose leaders were in complete control of the people and resources of the Nile valley. Establishing a power center at Nekhen (in Greek, Hierakonpolis), and later at Abydos, Naqada III leaders expanded their control of Egypt northwards along the Nile. They also traded with Nubia to the south, the oases of the western desert to the west, and the cultures of the eastern Mediterranean and Near East to the east, initiating a period of Egypt-Mesopotamia relations.[when?]

Figure 8: The sentence fadeout layout, meant to support a reader who is scanning. Sentences are progressively decreased in contrast as each paragraph progresses. This draws attention to areas of text correlated with higher relevance, like topic sentences and shorter paragraphs.

We alternatively considered some other layouts that would allow for a similarly non-disruptive reading process, but eventually rejected them. In our first design prototypes, the text layout in the reading mode would not have been simple plaintext, but rather Bionic Reading, an experimental layout created by Swiss designer Renato Casutt. The Bionic Reading formatting consists a slight increase in contrast on the first few letters of a word, with the intent of making the user's eye fixate on these letters. Its designer claims that the text layout simultaneously increases the user's reading speed and comprehension [3]. However, we find this claim dubious due to the reasons discussed in Chapter 3. We considered this layout due to its unobtrusive formatting and positive word-of-mouth reviews. However, we eventually rejected it because it lacks scientific evidence for its efficacy; the website claims that an unreferenced Swiss university is performing a study on the technique, but this claim is unsourced and no study has been published at the time of writing.

When the user is skimming text, we use the *content words* layout. This layout is a modified version of the QuickSkim layout described in Biedart et al.'s *Text 2.0* [12]. In this text layout, content words like verbs and nouns are at maximum contrast, while function words like prepositions and articles are decreased in contrast. This layout is appropriate for skimming because content words become more important as the user begins to skim. The choice was additionally made to use the same layout as Text 2.0, instead of any potential alternatives, because the Text 2.0 layout is a seminal work in this research area that has not yet been empirically studied. When considering this choice, however, we noticed that multiple implementations of the QuickSkim layout existed; we therefore needed to choose between these different layouts for our software.

The original layout in QuickSkim, as described in the conference proceedings, involved a continuous and continually changing highlight degree [12]. However, the demo video on the Text 2.0 website clearly shows a discrete boundary between reading and skimming [2]. Because we wished to switch between different interventions in a discrete fashion, it is beneficial for the text layout to be static and easily understandable at a glance. We therefore decided to use the modified version seen on the website, with discrete reading and skimming modes. For clarity, and because this text layout is not exactly the same as the QuickSkim layout originally described by Biedart et al., we rename our layout as the *content words* layout. While we use this new name for clarity in this thesis, we would be amiss not to credit Biedart et al. as the original creators of the text layout.

Finally, when the user is scanning text, we use the *sentence fadeout* layout. This layout is the same layout described by Kobayashi and Kawashima [48]. In this text layout, each sentence in a paragraph is increasingly faded going deeper into the paragraph. Because the original work describes the layout in general terms

without naming it, for sake of clarity we name this layout as *sentence fadeout*. Similarly, we exclusively credit Kobayashi and Kawashima for the creation of the layout.

Examples of the text layouts are available in Figures 6, 7, and 8. A video demo is available at <https://youtu.be/h-747bS0GtI>.

5.1.2 User Interface Design

The colour, size, and relative positioning of elements is important for any user interface, but this truism is particularly relevant for reading text [39]. Accordingly, we describe our typographical choices.

For our software, the background was set to white (#FFFFFF) and the text to black (#000000) for maximum contrast. Typographically, the text was laid out in a single left-justified column set in the center of the screen with approximately 6-inch margins on either side. The font size was set to 22 px, with the height of each line set to 1.5x the text height. This font size is larger than most modern websites; the large size was chosen to increase the accuracy of the eyetracker compared to each individual line of text. The text width was set at 795 px, resulting in an average characters-per-line of approximately 78.

Body text used the Tahoma font, which is a semi-condensed form of the Verdana font family. Verdana was initially chosen due to its high readability, as recommended by several non-scientific designer guides and scientifically studied in Hojjati & Muniandy [39]. However, in pilot testing we noticed that Verdana's large horizontal spacing creates text separation issues for very large text like ours. That is, the spacing of the letters can be so large that words are no longer perceptually grouped; spaces between letters become ambiguous with spaces between words. As such, we changed the font to Tahoma, a design with the readability of Verdana but with slightly condensed horizontal spacing. Headers used the Georgia font, a serif sister-font to Verdana.

While text was set at black by default, other colors were used for the skimming and scanning text layouts. The scanning mode used the paragraph fadeout layout described in Kobayashi and Kawashima [48]. We use the same colors described in that work: #000000 (black) for fading degree 0, and #545454, #888888, #B0B0B0, #CFCFCF, and #E0E0E0 for increasing fading degrees. For the skimming mode, we selected a color of #989898 (approximately 60% of the intensity of pure black) for the faded function words based on an iterative design process using the feedback from qualitative pilot studies.

5.1.3 *Software Components*

The application consists of two parts: a simple backend server to capture eye tracker data, and a frontend application.

The backend is a C# script that interfaces with the Tobii SDK to acquire and transmit its gaze data. This script then creates a local UDP server to transmit the gaze data to. The Tobii 5 has a 133 hz sample rate, but internally performs a downsampling process for noise reduction before the raw data is made available to consuming software at 33 hz; as such, the UDP server transmits 33 messages a second. Each message contains the x-coordinate in pixels, y-coordinate in pixels, an attention boolean, and a Unix timestamp, in the following format:

```
{"attention":true, "x":1532, "y":263, "timestamp":183474646}
```

The frontend is an Electron webapp, with the JavaScript portion of the webapp written using the React library. When launched, the application creates a listener to capture the UDP messages containing the Tobii gaze data.

For the application to be able to display text formatted in the content words and sentence fadeout modes, we must identify the function words and sentence breaks in a text. For performance reasons it is beneficial to pre-process the texts rather than doing so in JavaScript. Accordingly, we implement two scripts written in Python to perform this preprocessing. The first script, `highlight_contentwords.py`, outputs a modified version of the input text wherein each word is surrounded by an HTML span element: the content words with a class marking them as high contrast, and the function words as low contrast. The second script, named `fadeout_sentences.py`, uses the NLTK library to programmatically identify the sentences breaks in a text. Each sentence is then surrounded with an HTML span element with a class marking them with the appropriate fadeout level, as described in the work by Kobayashi and Kawashima [48]. We ran these scripts on all texts used in our software to acquire text files that can be directly inserted into an HTML document tree.

5.2 CONTROL TECHNIQUES

Two control techniques have been developed to transition between the different text layout modes: Automatic and Manual.

5.2.1 *Automatic Control Technique*

The Automatic control technique uses the read-skim-scan classifier described in Chapter 4 to determine the appropriate text layout mode. Because the classifier

was designed to provide generic functionality that can be applied in many different situations, we set certain constants for best applicability to our implemented software and hardware. The character width is set to 15 px and the line height is set to 39 px, matching the measured dimensions of the text used in our software. Additionally, we set the refresh rate to 33 hz to match the Tobii 5.

To minimize user disruption during the mode shifts, the color of the text changes over the course of 1.1 seconds using a linear interpolation function. Implementing this transition requires a somewhat unusual HTML layout, as this animation would not be performant to apply to each HTML element individually. Instead of having one HTML element per word and varying the opacity for each word independently, three HTML elements each contain a copy of the entire text, one per text layout. The three HTML elements, representing the plain, content words, and sentence fadeout versions of the full text, are *overlaid* on exactly the same position. The currently active text has full opacity, while the other two are set to opacity: 0. The mode transitions are activated by toggling the opacity of the given element, causing the previously active text to fade out over 1.1 seconds and the newly active text to fade in over 1.1 seconds. Because we use a linear interpolation function, and because the text elements are positioned exactly over each other, the transition visually appears to be a smooth fade between two different versions of the same text.

In development and testing we occasionally observed an unusual issue where the three elements, despite being exactly the same size and being positioned exactly identically, had different line breaks. These small differences would cause the text to jump on mode transitions, an unacceptable issue due to the disruption it would cause to the user's reading. We did not isolate the root cause of this behavior (although it probably occurs due to inconsistent behavior in rounding fractional pixel sizes). However, because the issue was deterministic for a given HTML layout, we simply ensured through manual testing that it did not occur in the final version of our software, as confirmed using screen recordings.

5.2.2 *Manual Control Technique*

By contrast to the Automatic control technique, the Manual control technique places all responsibility for transitioning between text layout modes in the hands of the user. Under this technique, three buttons are placed on the left-hand side of the screen, labeled "Remove Formatting", "Highlight Content Words", and "Fadeout Sentences". Each button, when clicked, triggers a text layout shift. Because the shifts are fully user-triggered in this mode, they take place instantly (without a fade animation). This is implemented in JavaScript simply by replac-

ing the HTML element that is currently in the document tree; unlike in the Automatic control technique, only one version of the text is in the document tree at any given time. This transition occurs in a single frame on our 60fps recordings; i.e., it is effectively instant.

In addition to the three buttons, keyboard shortcuts allow the user to change modes without using the mouse. Ctrl+1 changes to the plaintext layout, Ctrl+2 to the content words layout, and Ctrl+3 to the fadeout sentences layout.

As it may be nonobvious, we now describe the design rationale behind including the Manual control technique. The natural point of comparison for a novel reading augmentation technique is simply the lack of a technique: traditional, unmodified text. However, if this was our only point of comparison, we would not be able to tease out the effects of the *individual text layout modes* from the effects of the *gaze-aware swapping* between those layout modes. By implementing a traditional user-initiated control technique with identical text layout modes, we can estimate the effects of the Automatic control technique separately. The Manual control technique, therefore, represents our best-effort attempt to create an appealing and usable design; choices like the addition of keyboard shortcuts were decided upon based on our professional expertise to increase the usability of the design. Our user study, described in the next chapter, takes advantage of the fine-grained comparisons that the implementation of the Manual control technique allows.

EVALUATION

We conducted a user study with the goal of evaluating qualitative task performance metrics and usability metrics of Impulse Reading. Participants completed a text search task under Control, Automatic, and Manual conditions. In this chapter, quantitative metrics of eye gaze data, task performance, and software usability are described, and a thematic analysis of the qualitative feedback results is presented.

6.1 METHODS

We begin by describing the study setup, including our participant recruitment procedure, the study format, the procedure of the study sessions, and our choice of metrics.

6.1.1 *Participants*

30 participants were recruited, using two means: a recruitment email sent to all undergraduate students at Ontario Tech, and recruitment posters placed on Durham College and Ontario Tech campuses. While our recruitment methods were most likely to be seen by students, participants were not required to be students at either Durham College or Ontario Tech, or students at all. Participant gender was self-reported in a free-text format; 10 participants were male, 19 participants were female, and 1 participant was nonbinary. All participants were between the ages of 18 and 29.

Participants were required to affirm that they did not have dyslexia, photosensitive epilepsy, or eye problems, and had normal or corrected-to-normal vision. Participants were required to affirm that they were proficient in reading English, but were not required to be native English speakers. All participants described themselves as reading English “well” or “very well” (10/30 “well”, 20/30 “very well”).

6.1.2 Design and Stimuli

Our study used a repeated-measures, within-subjects design under three conditions: Control, Automatic, and Manual. The task selected was a text search task under time pressure. For each participant, the task was repeated three times with three different texts, once for each condition. Condition order was counterbalanced to reduce order effects; in other words, the subjects were divided into 6 groups of equal population, who experienced the three tasks in order CAM, CMA, AMC, ACM, MCA, and MAC. Similarly, text-condition pairing was counterbalanced to reduce any effects of text heterogeneity. Sample screenshots of the study software are available in Appendix A.

The text search task involved searching over long documents for a specific type of information (hereafter the *target concept*). For instance, the target concept of a text concerning a historical event might be a specific person; information about this person would be relevant, and information about any other person or persons would be irrelevant.

Each task was under a 5 minute timer. At the conclusion of the task period, the study software automatically proceeded to a set of comprehension questions concerning the target concept. The documents contained a mix of relevant and irrelevant information. Documents were quite long; fully reading each article in the allotted time would require an unrealistically high reading speed of over 800 WPM. As such, participants needed to skim or scan over the document to find relevant information. This task format, namely text search under time pressure, was chosen because it would necessitate naturalistic switches between reading, skimming, and scanning behavior, switches that were initiated based on the participant's best judgement and natural choice of reading strategies.

Texts were selected from Wikipedia's list of Featured Articles, articles that are considered to be of high quality. The list of Featured Articles was sorted by length and 3 articles between 3900-4300 words were selected. Article selection was based on the researcher's subjective judgement to find articles that were on varied topics, relatively obscure, non-controversial, non-technical, and were relatively easy to read. Finally, for each article the researchers selected the target concept. Texts were chosen such that a mix of relevant and irrelevant text would be spread relatively evenly throughout the article. Further details about the selected texts are available in Appendix A.

For each text, 4 comprehension questions were written by the researchers. Each question was multiple choice with 4 choices. The researchers followed the guidelines given by Fuhrman to develop fair and unambiguous questions that do not provide unintentional clues while remaining easily readable and comprehensible [33]. Additional guidance was found in an excellent work by Boland, Lester,

and Williams, although it was not used as a primary guide because its advice is primarily directed at academic psychiatrists [16]. For example, the following was a comprehension question used for the article *The Great Gold Robbery*, an article for which the target concept was the actions of James Burgess, a conspirator in the robbery:

What was Burgess's role in the robbery?

- a. He would deliver the stolen gold bars to a safe location once the train had arrived.
- b. He would notify the thieves of a shipment being made and let them into the guard's van.
- c. He would use wedges to break the iron rivets on the boxes of bullions once another thief had picked the safe lock.
- d. He would hide the evidence of the thieves' activities after the gold had been removed from the train.

We used a Tobii 5 eye tracker to collect and record gaze data. The operating distance (between the user's eyes and the screen) was about 50 to 70 cm. The screen was a 27-inch screen with a resolution of 1920 x 1080. The Tobii Experience application was used to calibrate the eye tracker. All hardware and software used for the eye tracker and its calibration were selected to be inexpensive, commercially-available options.

Study sessions were screen recorded using the Open Broadcaster Software (OBS), with an additional overlay for the user's gaze position provided by the Tobii Ghost software. The gaze position indicator was not visible to the participant in realtime, but was encoded into the captured screen recording.

6.1.3 Procedures

Each sessions began by calibrating the eye tracker using Tobii Experience's 7-point calibration procedure. Participants were then given instruction on the three text layouts used in the study, and were allowed unlimited time to use the Automatic and Manual techniques to gain familiarity. Participants were allowed to end the training period and begin the tasks at a time of their choosing. After the training period, participants completed three tasks under the Automatic, Manual, and Control conditions, in counterbalanced order. After finishing all tasks, participants answered a brief questionnaire, then were thanked for their time and given \$20 in remuneration.

After calibration, participants were trained on the text layouts. To train on the three text layout modes, the participants were shown a screen describing the intervention and its purpose. At the bottom of this screen, they were informed that upon advancing the study software, they would be shown an example of that text layout. Participants were asked to skim this page of text for the content words and sentence fadout layout modes, and asked to read the page of text thoroughly for the plaintext layout mode. Participants were asked to skim twice (instead of scanning) to acquire additional sample size for the skimming personalization described in Section 4.1.1, as more text is required for an equivalent amount of fixations while the user is skimming compared to reading.

Upon advancing to the next page, a text was displayed using the text format that the participant was currently being trained on. While this page was active, the participant's saccades were recorded, and used as input to the read-skim boundary personalization algorithm. The training texts were taken from the Wikipedia article *Ancient Egypt*. This article was chosen for its interesting subject matter and high likelihood of participant familiarity; these qualities would not be desirable for the study texts but were desirable for a training text to increase participant engagement.

Our choice to use participant training as a way to simultaneously collect personalization data is a noteworthy one. We believe it to be beneficial because it increases the time efficiency of design. We wish to note that we deliberately chose to not inform participants that this training process was also used for the skimming personalization. This was to increase the likelihood of acquiring naturalistic data, as participants may have felt unable to "act natural" if they knew their rate of skimming was affecting the experiment.

After the participant finished the three training pages for the text layouts, they were trained on the Manual control technique. A page briefly described the technique and showed the three buttons that allowed switching between the text layouts. Participants were allowed to experiment with switching between the different layouts, and advance at their own pace.

Afterwards, the participants were trained on the Automatic control technique. This page consisted of three paragraphs, each containing a description of the type of reading behavior that would trigger each text layout in order (i.e., reading, skimming, and scanning). Because we did not wish to assume participants remembered the names of the layouts, each paragraph displayed the text layout for the reading behavior it described in a static, unchanging formatting. After these paragraphs, the participant was asked to experiment with the Automatic control technique by switching between the three modes using their eyes. A long text, also taken from the Wikipedia article *Ancient Egypt*, followed. This text was dynamically formatted using the Automatic control technique, allowing the par-

ticipant to experiment with reading, skimming, and scanning. Participants were allowed to advance to the next page at their own pace.

After the training process, participants were asked to complete three tasks in order, one for each of the Control, Automatic, and Manual conditions. Because each task was completed in order by each participant, both order effects and sequence effects were possible. To guard against this, condition order was counterbalanced. 6 permutations are possible from 3 conditions, and 30 participants were recruited; as such, each possible order was seen 5 times. Each condition required a text to perform the information search task over; to avoid interaction effects the text-condition pairs were also counterbalanced. (Similarly, each possible text-condition permutation was seen 5 times.) The condition orders and text-condition permutations were counterbalanced separately.

Each task consisted of an introduction paragraph explaining the task, a 5-minute period to perform the text search task, then a set of four comprehension questions. After the Automatic and Manual conditions, there was additionally a 14-question usability questionnaire; as no reading augmentation was present in the Control condition, no usability questionnaire was included.

The introduction paragraph briefly described a roleplaying scenario that the participant was asked to act out. For example, one task asked the participants to read an article about a historical event, searching for information only on a single historical person; participants were asked to imagine themselves as a biographer writing a book about this person. Participants received a description of the content of the article, along with a description of the target concept they were searching for. They were informed of which control technique would be in use for the search task. Finally, they were informed of the 5 minute task timer, and reminded that they would have to skim quickly or skip portions of text in order to cover the entire text. Participants were allowed to begin the task (and the 5 minute timer) at their own pace.

During the text search task, the text was placed in the middle of the screen as described in Section 5.1.2. Additionally, a timer was present on the left side of the screen, along with a reminder of the target concept. At the bottom-right of the screen, a button was placed that allowed the participant to forfeit their remaining time and skip to the comprehension questions. Screenshots of the task screen are available in Appendix A.

The post-task survey consisted of two portions. The first portion is the System Usability Scale (SUS), a well-known measure of usability [52]. The System Usability Scale consists of 10 statements, with the participant choosing their agreement on a 5-point Likert scale between *Strongly Agree* and *Strongly Disagree*. After these 10 statements, participants were asked to rate their agreement with 4 statements,

with the participants choosing on the same 5-point Likert scale. The text of the custom questions were as follows:

- My strategy for finding information used the text highlighting.
- I found the text formatting useful when it highlighted content words, like verbs and nouns.
- I found the text formatting useful when it faded out the sentences in a paragraph.
- Changing the text highlighting distracted me.

The first question investigates whether the participant consciously changed their strategic choices in regards to the task. The second and third questions investigate whether the individual text layouts were perceived as useful. Finally, the fourth question investigates whether the *change* between modes was disruptive, separate from whether the modes themselves were useful.

Once participants had finished all three tasks, they answered a post-study survey. This survey included three forced-choice questions comparing the participant's experience in the Manual and Automatic conditions. Finally, a free-text field allowed the participants to offer any additional comments they wished. After submitting the final survey, the participants were thanked for their time and given their remuneration.

The duration of the experiment, including the training and tasks, was dependent on how quickly participants chose to proceed. In practice, no sessions were shorter than 20 minutes or longer than 45 minutes.

6.1.4 Analyses

Quantitative task performance was scored using comprehension question scores. Each condition was measured by the participant's percentage correct on the 4 multiple choice questions for each condition. Unanswered questions were coded as incorrect.

Other quantitative metrics include the number of unanswered questions, the SUS scores for the Automatic and Manual conditions, the 4 additional Likert scale responses for each condition, and the results of the three forced-choice preference questions at the end survey. Additionally, gaze data was used to estimate the percent of fixations that were on relevant text. As this metric is notably more complex and potentially ill-defined than our other quantitative metrics, we first describe in precise detail how we calculated it.

To investigate the degree to which participants could find relevant text, we count the number of fixations for each participant that were on relevant text. We compare this to the number of fixations that were on irrelevant text. For each of the three texts, we manually labelled each sentence as relevant or irrelevant based on whether it relates to the target concept for that text. While relevance was almost entirely determined by sentences (as opposed to words or paragraphs), we additionally allowed for boundary breaks at clause boundaries of compound sentences if one clause was determined to be relevant and the other irrelevant. We extracted the x,y bounding boxes of each sentence in Javascript, and compared the coordinates to the Tobii eye gaze coordinates.

The Javascript x,y coordinates are measured in pixels relative to the top-left of the webpage; however, the Tobii gaze coordinates are in pixels relative to the top-left of the computer monitor. To convert between these coordinates, we add the current scroll position in pixels to the Tobii y coordinate. No change is made to the Tobii x coordinate because the webpage could not be scrolled horizontally. Finally, 21 pixels are subtracted from the y coordinate, as the webpage's origin coordinate is 21 pixels below the top of the monitor due to a toolbar inserted by Electron. This converted coordinate is then compared to all bounding boxes for relevant text; if the point is within a 15-pixel margin of any bounding box, it is labeled relevant. Otherwise, it is labeled irrelevant. A margin is necessary due to the imperfect accuracy of the eye tracker. 15 pixels of margin were chosen because it is half of the size of the fixation window described in Section 4.1; that is, it is exactly the same distance from the center of a fixation that would still be accepted as part of that fixation.

6.2 RESULTS

For reading comprehension scores, we conducted a one-way repeated-measures ANOVA to compare the overall means (out of 4) of the Control ($M = 2.03$, $SD = 1.13$), Automatic ($M = 2.07$, $SD = 1.05$), and Manual ($M = 2.10$, $SD = 1.21$) conditions. There were no statistically significant difference between conditions, $F(2,58) = 0.03$, $p = 0.97$. A raincloud chart showing the overall distribution is given in Figure 9.

For the rate of unanswered questions, we conducted a one-way repeated-measures ANOVA to compare the overall means (out of 4) of the Control ($M = 0.30$, $SD = 0.65$), Automatic ($M = 0.20$, $SD = 0.55$), and Manual ($M = 0.1$, $SD = 0.31$) conditions. There were no statistically significant difference between conditions, $F(2,58) = 1.10$, $p = 0.34$.

For the percentage of fixations that were on relevant text, we conducted a one-way repeated-measures ANOVA to compare the overall means of the Control ($M = 37.2$, $SD = 19.0$), Automatic ($M = 37.5$, $SD = 15.7$), and Manual ($M = 37.8$, $SD = 18.4$) conditions. There were no statistically significant difference between conditions, $F(2,58) = 0.01$, $p = 0.99$. A raincloud chart showing the overall distribution is given in Figure 10.

While our study was not primarily designed to evaluate the individual text layouts, we also report our statistics for the Likert scale questions described in Section 6.1.3, concerning participants' reactions to the different text layout modes. Scores were converted to a numeric scale, with *Strongly Disagree* being a 1 and *Strongly Agree* being a 5. When asked in the Automatic condition, the content words layout received a median response of 4 and a mode of 5. The sentence fadeout layout received a median response of 3 and two modes of 3 and 4. When asked in the Manual condition, the content words layout received a median response of 5 and a mode response of 5. The sentence fadeout layout received a median response of 3 and a mode response of 2. Note that we report medians and modes instead of means, and do not include standard deviations, due to the ordinal nature of Likert response data.

The correct method of applying statistical analysis to ordinal data like Likert scales is a commonly debated topic; in particular, no strong agreement has been reached on whether parametric data analysis is acceptable [37]. For our analysis, we choose to apply non-parametric tests like the Wilcoxon signed-rank test. For the Automatic condition, we conducted a Wilcoxon signed-rank test to compare the Likert scale responses for the content words layout to the responses for the sentence fadeout layout. The difference was significant, $Z = 3.17$, $p < 0.005$. In the Manual condition, the contrast was even more stark; a Wilcoxon signed-rank test found the difference to be significant, $Z = 3.85$, $p < 0.0005$. That is, participants found the content words layout significantly more useful than the sentence fadeout layout in both conditions. A diverging bar chart plot showing the overall distribution is given in Figure 11.

For the question "My strategy for finding information used the text highlighting" which used a 5-point Likert scale, the median response for the Automatic condition was a 4 and the mode was a 4. For the Manual condition, the median response was a 4.5 (i.e., halfway between a 4 and a 5, due to our sample size being an even number) and the mode was a 5. A Wilcoxon signed-rank test found no significant difference, $Z = 0.66$, $p = 0.51$.

For the question "Changing the text highlighting distracted me" which used a 5-point Likert scale, the median response for the Automatic condition was a 3 and the mode was a 4. For the Manual condition, the median response was a 2 and the mode was a 1. A Wilcoxon signed-rank test found the difference to

be significant, $Z = 2.08$, $p < 0.05$. That is, participants found changes in the text highlighting to be significantly more distracting in the Automatic condition.

The post-study survey included three forced-choice preference questions. For the question “Which technique did you like most”, 17/30 (56.7%) participants chose Manual and 13/30 (43.3%) chose Automatic. For the question “Which technique do you feel made the task easier to complete”, 15/30 (50%) participants chose Manual and 15/30 (50%) of participants chose Automatic. For the question “In which technique do you feel you were best able to answer the questions”, 21/30 (70%) chose Manual and 9/30 (30%) chose Automatic. A visualization is given in Figure 12.

SUS scores were taken for the Manual control technique and the Automatic control technique. The mean SUS score for the Manual condition was 80.4. The mean SUS score for the Automatic condition was 76.2. According to Sauro’s commercially-available guide to the SUS, raw SUS scores should be accompanied with a percentile score [66]. By their scale, the Manual condition’s raw score of 80.4 is approximately 90th percentile, while the Automatic condition’s raw score is approximately 70th percentile. Both scores indicate high usability; under Bangor et al.’s adjective rating scale, the Manual condition’s raw score is classified as “Excellent”, while the Automatic condition’s raw score is classified as “Good” [8]. A raincloud chart showing the overall distribution is given in Figure 13.

6.3 DISCUSSION

First and foremost, it is concerning to note that our quantitative task performance outcomes did not find any significant effects. However, one key difficulty for our study, and potentially a contributing factor to our lack of significant findings, is our relatively small sample size of 30 participants. While this sample size may be sufficient to detect large, obvious effects, it is possible for false negatives to occur when effect sizes are small [69]. Because it is difficult or impossible to improve reading speed or comprehension without a corresponding tradeoff, a small effect size should not be surprising for a non-disruptive reading augmentation [62].

As such, it is entirely possible that our t-tests were investigating real effects; our lack of significance should not be taken as definitive proof that no real effect exists on our quantitative task performance metrics. It is noteworthy that there are 6 pairwise comparisons of Automatic or Manual conditions with the Control condition, and in all 6 cases the Automatic or Manual condition is on average the better of the two. To be clear, we are not claiming that this fact proves anything about our data; however, we do claim that our lack of significant findings should be interpreted as a weak or missing signal, rather than a strong negative one.

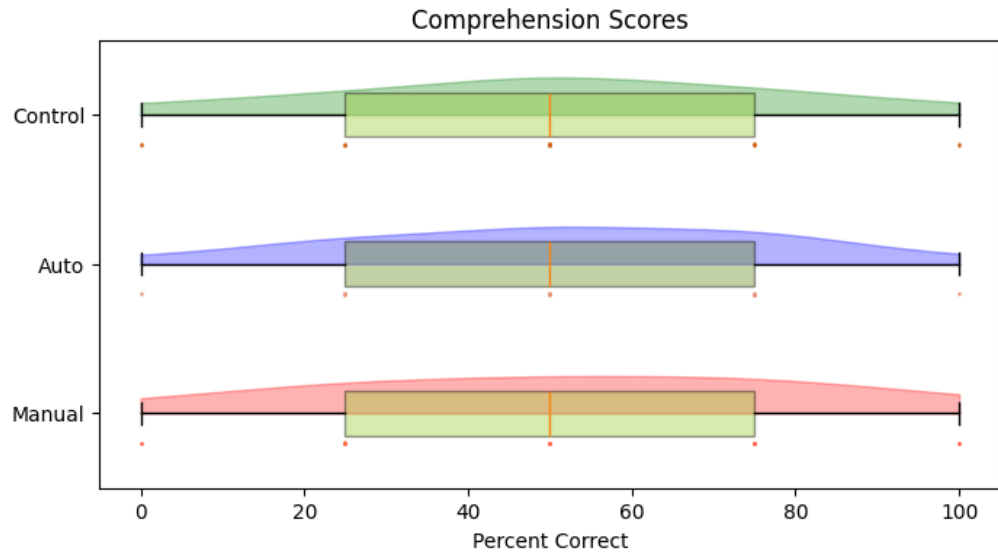


Figure 9: Raincloud plot of comprehension scores by condition.

In our results, we reported that 37.5% of all fixations were relevant, compared to 21.4% of words in the text that were relevant. These numbers may be slightly misleading because not all words take up an identical amount of pixels on the screen, and also are affected by the 15 pixels of margin applied to each fixation. That said, the large difference between these numbers may suggest that participants were able to skip irrelevant text and focus on relevant text to some degree.

We found that the usability scores, as measured using the SUS, were overall higher for the Manual condition than the Automatic. A closer analysis of the distribution of scores reveals an enlightening pattern: the high end of the distribution were quite similar for both conditions, but the lowest scores for the Automatic condition were much lower than the Manual condition. The 50th percentile for both conditions was exactly the same, at a score of 81.3, but the 25th percentile for Automatic was 65.0 compared to Manual’s 70.6. This may indicate a “love it or hate it” response to the Automatic condition. This may shed some interesting light on the results of the forced-choice preference questions; although Automatic seemed to be slightly less well-received overall, still more than 40% of participants indicated an active preference for Automatic over Manual.

The personalization process described in Section 4.2 was used in our study. Accordingly, we describe the high-level calibration results, and compare them to the non-personalized algorithm given in Buscher 2008. Our participants had a mean boundary of 7.69 character spaces, which is quite similar to Buscher’s constant boundary of 8. However, significant variety between participants was observed. The standard deviation was quite large, at 2.17, with the lowest boundary being 4.59 character spaces and the highest being 14.79 character spaces. Given

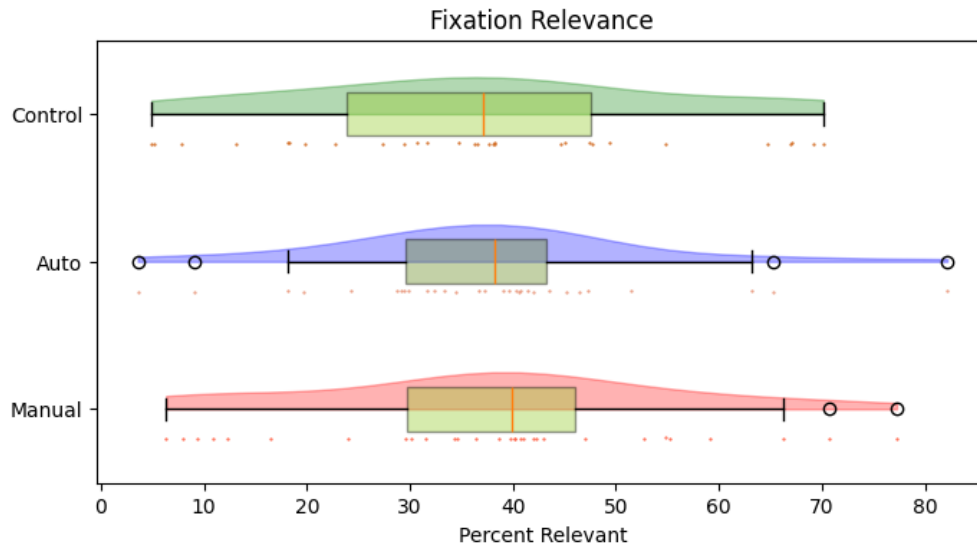


Figure 10: Raincloud plot of the percentage of fixations on relevant text by condition.

this large variation, it is clear that a constant boundary would not have been appropriate in our case.

However, it must also be investigated whether our calibration process performed adequately. Because skimming is associated with a higher average saccade distance, the difference in average saccade distance between skimming and reading should be positive. Out of our 30 participants, 6 had a negative difference, and 3 had a highly negative difference (< -1). These participants therefore experienced a suboptimal calibration process. Given that no loss of eye tracking was observed in these study sessions, one possible explanation is that these participants did not understand or did not comply with the instruction to skim or thoroughly read the calibration texts. We did not intervene in these cases; we believe that future work inspired by our experimental protocol should instead intervene and re-do calibration.

Given that the calibration process only affects the Automatic condition, any negative effects of a suboptimal calibration process would be constrained to that condition. Compared to the overall mean, these 6 participants had an overall worse user experience with the Automatic condition (average SUS score of 72.9 vs. 76.2 overall), a worse reading comprehension on the Automatic condition (average of 1.3 questions correct vs. 2.1 overall), and a lower percentage of fixations on relevant text (36.2% vs 37.5%). That is, participants with suboptimal calibration performed worse in all relevant metrics in the Automatic condition. This reinforces the importance of the personalization process. However, it should be noted that we cannot rule out that these differences were due to inherent differences in the 6 users with suboptimal calibration, rather than a causal effect of the suboptimal calibration.

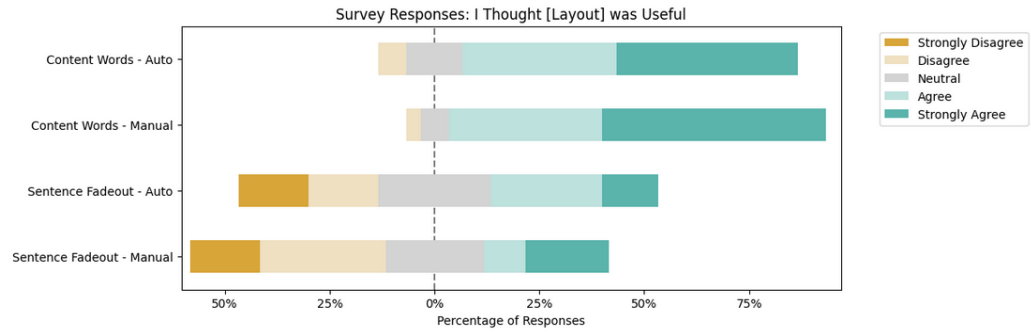


Figure 11: Diverging bar chart of 5-point Likert Scale responses to the four questions “I found the text formatting useful when [layout description].”.

The free-text comments participants offered at the end of the study may provide thematic insight into our usability results. We begin by analyzing the comments from participants who offered negative feedback on the Automatic control technique, including the 4-digit unique ID used to track the participants in a non-identifiable fashion. The most common negative feedback was that the changes between text layout modes was distracting. Comments on this theme include “Automatic switching made it easier to find the information, but it was distracting” by participant 5946, and “...it can be a bit distracting when it switches between methods in the middle of reading a paragraph” by participant 6956. Potentially related to the theme of distraction, negative responses highlighted the lack of predictability in the mode switches: “When I scrolled down it would switch when I didn’t want it to switch” by participant 3588, and “the system is not stable enough to switch between the modes seamlessly” by participant 9349. This theme is consistent with our finding from the Likert scale questions that participants found change more distracting in the Automatic condition.

Positive comments related to the Automatic control technique tended to focus on its ease of use relative to the Manual control technique. Comments on this theme include “The automatic method was good because I did not have to focus on evaluating when I should change the mode” by participant 1057, and “The changes on the screen were slightly distracting at first, but I got used to it quickly. Those initial distractions were much less significant than deciding when to switch while reading in manual mode”, also by participant 1057. Other comments focused on the Automatic control technique’s utility for performing the task: “I believe automatic switching made the task easier to complete” by participant 3426, and “I think automatic switching was a great help” by participant 9344. The same comment by participant 9344 followed that “if I could tune the automated system... I would use it all the time.” Given our continuing emphasis on personalization, we agree with this participant that a user-initiated calibration process is an important next step.

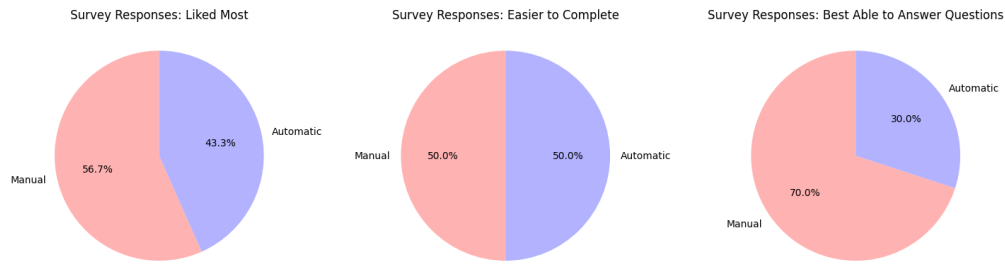


Figure 12: Survey results for the forced-choice preference questions answered by each participant at the end of the study.

Our study was predominantly intended to empirically evaluate the control techniques, not the individual text layouts; as such, we were intrigued to see that the single most common theme in participant comments was feedback on the content words layout mode. Furthermore, this feedback was unanimously positive. 5 out of the 12 received comments were on this theme: “[I found] highlighting verbs and nouns to being [sic] more efficient and easier to skim” by participant 5094, “I found the ‘Highlighting Content Words’ in the Manual Switch Technique was an extremely useful way to find necessary information. It helped me a lot, and it made it much easier for me to find information” by participant 6956, “I think the paragraph fading format is less useful compared to the highlighting content words format” by participant 7726, “I tend to favor the term-based highlighting method” by participant 9344, and “I like the text highlighting for sure” by participant 1325. This surprising result merits further investigation.

The extremely low p value found in Section 6.2 comparing the participant reactions to the content words layout and the sentence fadeout layout is even more intriguing. Even leaving aside statistical tests, the overall distribution given in Figure 11 shows a huge difference in reaction between the two layouts, clearly visible to the naked eye. This clear signal is particularly surprising because the sentence fadeout layout has a stronger prior evidence base supporting its use than the content words layout; the content words layout (or QuickSkim, as originally named by Biedert and Buscher) has never before been empirically evaluated [12]. While this evaluation is not a true comparison to a control condition and should be interpreted with caution, it seems clear that the content words layout deserves further research in future work.

This closes our analysis of our data; however, in the spirit of open science we wish to give other researchers the opportunity to perform their own analyses. We have made the decision to release all non-identifiable data, publicly available at the Open Science Framework at https://osf.io/bnqkd/?view_only=052792b9f6e543a7bea378de1dd88178. The decision to share our data publicly

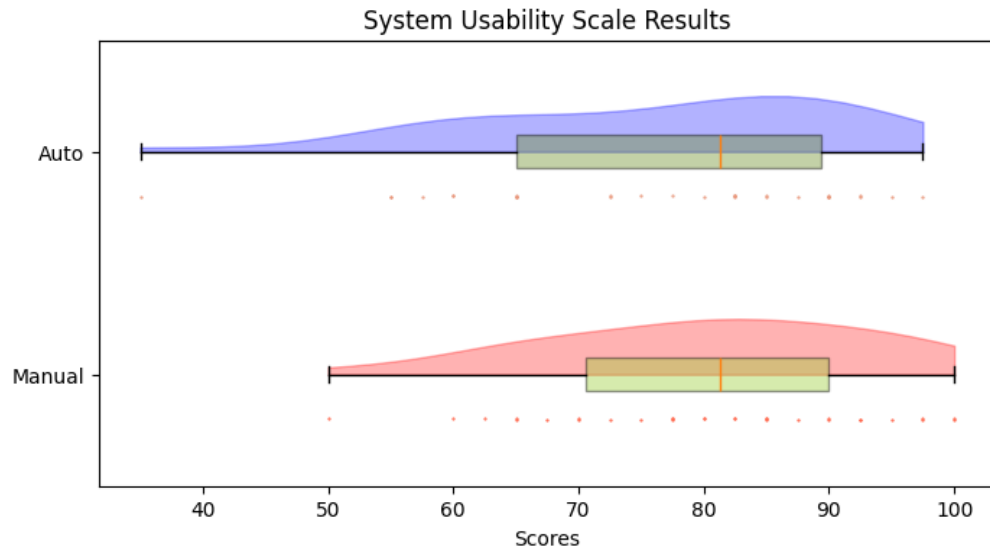


Figure 13: Raincloud plot of SUS scores by condition.

serves two key purposes: it allows our research to be replicated and validated by future researchers, and it provides a scientific dataset of gaze data during a text search task.

It is noteworthy that we are aware of no existing public dataset of gaze data during a text search task. Existing public datasets, like Strukelj and Niehorster’s work, cover a wide range of tasks without including text search [68]. Some papers like Symons and Pressley’s have described reader behavior during text search based on the recollections of human observers, but without the use of eye trackers to capture gaze data [70]. The closest work we are aware of, by Cole et al., investigates task and user effects in a text search task using gaze data; however, their publication does not include public access to their data [25].

While the creation of this dataset is not intended to serve as the main contribution of this thesis, we would be remiss not to briefly describe the attributes and qualities of the dataset. Analysis and statistics for the present thesis was generated using a Jupyter Notebook `analysis.ipynb`, allowing for simple re-analysis of our data. Study logs for each participant are included; log data is formatted as simple text files containing one log event per line. Log events are generated for each fixation, saccade, text layout switch, window scroll event, and task switch. Additionally, the participant’s reading comprehension results and survey answers are available, including both quantitative results and any free-text comments.

While the most obvious reason for making this dataset publicly available is due to the clear and obvious value of open science, the dataset may additionally be useful for future work that is not directly related to this thesis. Because we include all saccades and fixations of a naturalistic reading task, future machine

learning models for reading behavior classification can use our data for training or validation; logs of mode shifts in the Automatic condition may be useful for work looking to measure and reduce mode thrashing. Future research into text search tasks can also use our dataset to create or validate theories without duplication of effort.

6.3.1 *Limitations*

One of the key limitations of our work is the lack of experimental validation of our rules-based reading-skimming-scanning classifier. We made this choice because we did not see it as the main contribution of the research; while the classifier is required for our implementation of Impulse Reading, we believe the interaction design of Impulse Reading described in Chapter 5 and its experimental validation described in Chapter 6 to be the main contributions. Additionally, the limited timespan of a Master's thesis required us to perform only one experiment; this experiment was required to allow for naturalistic flow between multiple reading behaviors to support our multiple text layouts. As such, no clear ground truth existed for our participant's reading behaviors without requiring significant additional investment for manual labeling, and we could not accurately estimate our classifier's accuracy. We strongly recommend that future work using our models perform experiments to experimentally validate the classifier's accuracy.

Although we were able to collect qualitative feedback from the participants through free-text comments, we did not include a post-study interview with our participants due to time constraints. This choice reduced our ability to collect qualitative feedback from participants, and correspondingly reduces our certainty in the grouping of lessons learned from the thematic analysis we applied in Chapter 6.

The open dataset available at the Open Science Framework contains gaze data from naturalistic participant behavior in a participant-guided text search task. These properties have both benefits and drawbacks; the lack of labels or ground truth imposes a limitation on its use for training machine learning models for the task of reading behavior classification. In other words, because our participants were allowed to freely and naturalistically switch between the reading behaviors of reading, skimming, and scanning, we do not have a known-good label for the specific behavior they exhibited at any specific moment in time. While some previous studies have successfully applied self-supervised learning to eliminate the need for labelling and would therefore be unaffected by this limitation, other models may require manual labeling to use our data for training [42].

CONCLUSION

In this chapter, we conclude our thesis by summarizing our contributions and presenting ideas for future work.

Our first contribution, as described in Chapter 4, is a classifier for reading, skimming, and scanning detection. Our classifier expands the state of the art by solving the problem of mode thrashing using a novel score-based implementation of hysteresis. Compared to existing machine learning models, our rules-based model increases the usability of the tool by preventing undesirable user-facing behavior caused by frequent mode shifts.

Our second contribution, which we believe to be the most important contribution of the present thesis, is the design, development, and experimental validation of Impulse Reading, as described in Chapters 5 and 6. We have presented the first empirical analysis of what Buscher and Biedert call *attentive documents*, as implemented in our Impulse Reading system. We have shown for the first time that there is value in Buscher and Biedert’s early work on gaze-aware reading systems by extending their prototypes and evaluating them to show high system usability and good user feedback.

Finally, we have created and made public a scientific dataset of gaze data during a text search task. This dataset is the first of its kind that we are aware of, as no existing gaze data datasets include specifically text search tasks. This open dataset is an important contribution to future machine learning models for the task of reading behavior classification, as well as allowing for the validation of future research in reading behaviors during text search without unnecessary duplication of effort.

While our current implementation uses relatively cheap, commercially available hardware in the form of the Tobii 5 eye tracker, we believe future work in this area should investigate the possibility of using camera-based eye trackers, such as traditional webcams or the front camera of a mobile device. Webcam eye tracking is uncommon, both in research and commercial applications, due to their lower accuracy and sample rate [72]. However, a key point of interest of our reading detection algorithm described in Chapter 4 is that it does not require precise *positional* tracking of the reader’s gaze, only *relative* tracking. That is, while calibration errors can entirely preclude the usefulness of many eye-tracker-based techniques, our algorithm is mostly unaffected by absolute error and drift. As such, a near-identical version of our software could be implemented using

only webcam eye tracking with only minor degradation in performance. While the exact market penetration of webcams is not currently known, some sources estimate it at approximately 80% [4]. A webcam-based implementation would therefore make our reading augmentation available to potentially more than a billion people without any additional purchases or hardware required.

More speculatively, we would like to explore integrating Large Language Models (LLMs) like ChatGPT into our work. With such integration, we would be able to draw the reader's attention to words, phrases, or sentences based on their semantic relation to the passages that they have previously focused most on. This could take the form of additional inserted content, like a LLM-generated summary of the most relevant information inserted into the sidebar of the text, or use unobtrusive augmentations like typographical cuing. For example, the content words or sentence fadeout layouts could instead increase the saliency of the most relevant sentences in the text, based on the sentences that the reader has focused most on. A key note is that this concept is that it requires the identification of which word a reader is fixated on, instead of our current algorithm which simply estimates the direction and magnitude of saccades without any care for absolute positioning. As such, it is mutually incompatible with the prior suggestion for webcam eye tracking; webcam eye trackers have yet to achieve sufficient accuracy and precision to be able to consistently identify which word a user is reading [72].

We see potential in the idea of expanding our work into the domain of language acquisition. Because reader fluency, comprehension, and attention can be inferred from gaze data, eye-tracking-based augmentations for language learners are especially appealing [14, 58]. Using similar algorithms to ours described in chapter 4, we could provide support for the language learner tailored to their current mental state and reading behavior. For example, we see possible value in the idea of allowing the user to read in their second language while they are performing thorough reading, but seamlessly replacing it with text in their native language while they are skimming and scanning. More generally, inferring the reader's degree of comprehension, current reading behavior, and current task or goal could be used to provide customized support tailored to the unique situation of each user.

We would also like to encourage future work into increasing the accessibility of this and similar tools. For example, while prior work exists in creating an analogue to skimming in blind users through technological interventions [5–7], we are unaware of any work that aims to infer the pre-existing *listening behavior* of these users. We are interested in the possibility of detecting the equivalent of skimming and scanning in blind users, along with the corresponding potential to provide support tailored for these diverse listening behaviors.

To further explore the topic of accessibility, we would like to encourage future investigation into the applicability of this work to people with specific disabilities or mental conditions. In particular, we would like to highlight dementia, dyslexia, and attention deficit hyperactivity disorder as possibilities. Patient reading ability and behavior is known to correlate with the presence and severity of dementia, suggesting that reading augmentations may be useful for either the diagnosis or treatment of the condition [55]. Meanwhile, dyslexia and attention deficit hyperactivity disorder are known to affect reading behaviors; we would like to encourage future work to investigate the efficacy of reading detection algorithms in people with these conditions [31, 57].

Overall, the present thesis has presented an extension and evaluation of a gaze-aware reading augmentation based on the work of Biedert and Buscher, including for the first time an experimental validation of positive user experience in the use of attentive document systems. We believe our strong qualitative and quantitative user feedback supports the idea that multimodal reading augmentations supporting diverse reading behaviors are a promising area for future research. We hope this research inspires a renewed interest in a research area that has gone sadly underexplored since the seminal works of Biedert and Buscher.

APPENDIX

In this appendix, we provide more details and screenshots of the text search task described in Chapter 6. For the text search task, three texts were selected from Wikipedia: *Brownhills*, *Water Rail*, and *The Great Gold Robbery*.

Brownhills describes a small English town with a history of coal mining. The target concept for this article was historical events related to the mines of the town.

Water Rail describes a species of waterfowl widespread across Europe, Asia, and North Africa. The target concept for this article was any information related to the breeding and nesting of the bird.

The Great Gold Robbery describes a historical event wherein gold bullion was stolen from a moving train en route to Paris. The target concept for this article was James Burgess, one of the four conspirators of the plot.

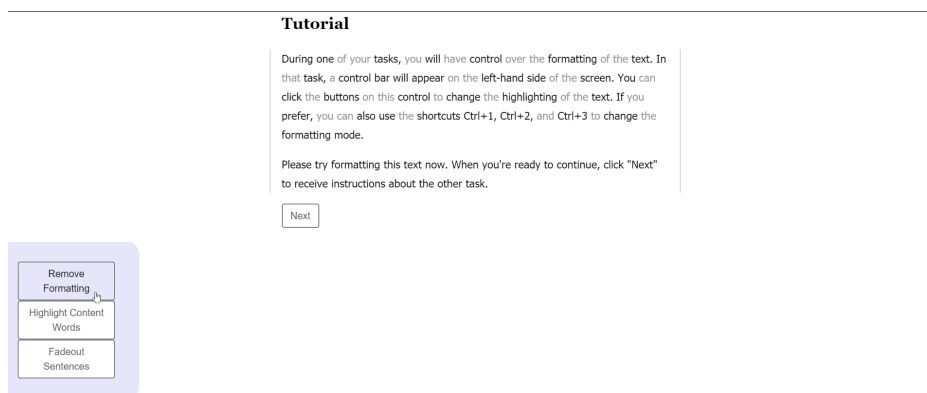


Figure 14: The tutorial on the Manual control mode. Participants were allowed to freely experiment with changing the text layout modes before advancing at their own pace.

Control Task

For this task, you will be roleplaying as a biographer interested in the life of a specific historical person. Namely, you are interested in a person named James Burgess who was involved in a gold robbery. To do this, you will read a passage describing the events and aftermath of the gold robbery. For your report, only some of the information in this passage will be useful: you will need to find information on **James Burgess**. The details of any other person can be ignored. The text is quite long, so it is recommended to skim the text quickly to find the information you need.

This task will not format the text in any special way.

Once you begin, you will have 5 minutes to read. After these 5 minutes are up, we'll ask you some questions about the passage. You won't be able to go back to the passage once time is up, so do your best to read quickly and find the most relevant information. These questions will ask only about James Burgess, so be on the lookout for those details.

Start

Figure 15: A task introduction containing the roleplaying instructions, target concept, and task format. In particular, this screenshot introduces the task for *The Great Gold Robbery* under the Control condition.

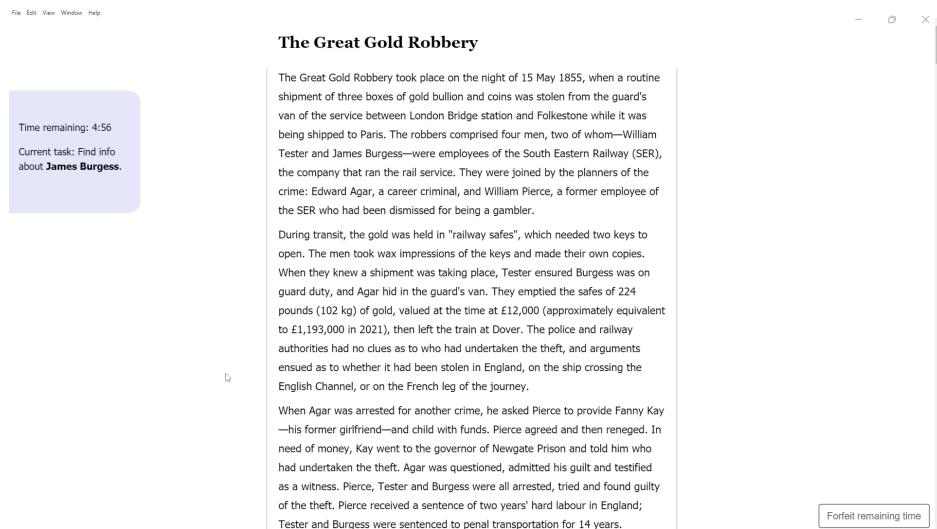


Figure 16: The task screen. A continuation of the previous image, this task is also for *The Great Gold Robbery* under the Control condition.

Questions

Time is up for the task. Before we move on, please answer the following comprehension questions about the passage you just read.

As a reminder, your performance is not being evaluated. It's okay if you don't know the answer to a question. You may leave questions blank if you don't wish to answer them or don't know the answer.

1. What best describes Burgess's prior history at the South Eastern Railway company?

- ☐ He had a history of being reprimanded for gambling and drinking on the job.
- ☐ He was regarded with suspicion due to his short tenure.
- ☐ He was a respectable man who had worked as a guard for over a decade.
- ☐ His employment with the railway traffic department meant that he had previously interacted with the company.

2. What was Burgess's role in the robbery?

- ☐ He would deliver the stolen gold bars to a safe location once the train had arrived.
- ☐ He would notify the thieves of a shipment being made and let them into the guard's van.
- ☐ He would use wedges to break the iron rivets on the boxes of bullions once another thief had picked the safe lock.
- ☐ He would hide the evidence of the thieves' activities after the gold had been removed from the train.

Figure 17: The comprehension questions screen for the previous task.

Survey

In today's study, two of the tasks used text formatting. During those tasks, we used two different techniques to switch the formatting of text. In the Manual Switching technique, you used buttons to switch the formatting at a time of your choosing. In the Automatic Switching technique, the computer system did its best to figure out which formatting was useful using eye tracking.

1. Which technique did you like the most?

- ☐ Manual Switching
- ☒ Automatic Switching

2. Which technique do you feel made the task easier to complete?

- ☒ Manual Switching
- ☐ Automatic Switching

3. In which technique do you feel you were best able to answer the questions?

- ☒ Manual Switching
- ☐ Automatic Switching

Do you have any other comments on your preference of technique? (If you can't type in this text box, please press Command+Tab twice, then try again.)

Figure 18: The final survey, including three forced-choice preference questions and a free-text comment box. This survey was shown at the conclusion of the study, after all three tasks had been completed.

BIBLIOGRAPHY

- [1] Sept. 2023. URL: <https://www.eslfluency.com/language-skills/reading/skimming-and-scanning/6066/>.
- [2] June 2023. URL: <https://text20.net/>.
- [3] June 2023. URL: <https://bionic-reading.com/>.
- [4] Sept. 2023. URL: <http://zugara.com/webcam-penetration-rates-adoption>.
- [5] Faisal Ahmed, Yevgen Borodin, Andrii Soviak, Muhammad Islam, IV Ramakrishnan, and Terri Hedgpeth. "Accessible skimming: faster screen reading of web pages." In: *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 2012, pp. 367–378.
- [6] Faisal Ahmed, Andrii Soviak, Yevgen Borodin, and IV Ramakrishnan. "Non-visual skimming on touch-screen devices." In: *Proceedings of the 2013 international conference on Intelligent user interfaces*. 2013, pp. 435–444.
- [7] Ameer Armaly, Paige Rodeghero, and Collin McMillan. "Audiohighlight: Code skimming for blind programmers." In: *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE. 2018, pp. 206–216.
- [8] Aaron Bangor, Philip Kortum, and James Miller. "Determining what individual SUS scores mean: Adding an adjective rating scale." In: *Journal of usability studies* 4.3 (2009), pp. 114–123.
- [9] Tanya Beelders and Angela Stott. "Eye Movements during Barking at Print." In: *Visual Impairment and Blindness*. Ed. by Giuseppe Lo Giudice and Angel Catalá. Rijeka: IntechOpen, 2018. Chap. 21. DOI: [10.5772/intechopen.81898](https://doi.org/10.5772/intechopen.81898). URL: <https://doi.org/10.5772/intechopen.81898>.
- [10] Bell, Charles E. "On the motions of the eye, in illustration of the uses of the muscles and nerves of the orbit." In: *Philosophical Transactions of the Royal Society of London* 113 (1823), pp. 166–186.
- [11] Simone Benedetto, Andrea Carbone, Marco Pedrotti, Kevin Le Fevre, Linda Amel Yahia Bey, and Thierry Baccino. "Rapid serial visual presentation in reading: The case of Spritz." In: *Computers in Human Behavior* 45 (2015), pp. 352–358.

- [12] Biedert, Ralf, Buscher, Georg, Schwarz, Sven, Jörn Hees, and Dengel, Andreas. "Text 2.0." In: *Proceedings of the 28th of the International Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '10*. 2010.
- [13] Biedert, Ralf, Buscher, Georg, and Dengel, Andreas. "The Eye Book." In: *Informatik-Spektrum* 33.3 (2009), pp. 272–281.
- [14] Ralf Biedert, Andreas Dengel, Mostafa Elshamy, and Georg Buscher. "Towards robust gaze-based objective quality measures for text." In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. 2012, pp. 201–204.
- [15] Ralf Biedert, Jörn Hees, Andreas Dengel, and Georg Buscher. "A robust realtime reading-skimming classifier." In: *Proceedings of the symposium on eye tracking research and applications*. 2012, pp. 123–130.
- [16] Robert J Boland, Natalie A Lester, and Eric Williams. "Writing multiple-choice questions." In: *Academic Psychiatry* 34 (2010), pp. 310–316.
- [17] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V. Elst. *Attentive documents*. en. 2012. DOI: [10.1145/2070719.2070722](https://doi.org/10.1145/2070719.2070722). URL: <http://dx.doi.org/10.1145/2070719.2070722>.
- [18] Georg Buscher, Andreas Dengel, and Ludger van Elst. *Eye movements as implicit relevance feedback*. 2008. DOI: [10.1145/1358628.1358796](https://doi.org/10.1145/1358628.1358796). URL: <http://dx.doi.org/10.1145/1358628.1358796>.
- [19] Christopher S. Campbell and Paul P. Maglio. *A robust algorithm for reading detection*. 2001. DOI: [10.1145/971478.971503](https://doi.org/10.1145/971478.971503). URL: <http://dx.doi.org/10.1145/971478.971503>.
- [20] Carter, Benjamin T., and Luke, Steven G. "Best practices in eye tracking research." In: *International Journal of Psychophysiology* 155 (2020), pp. 49–62.
- [21] Ronald P Carver. *Reading rate: A review of research and theory*. Academic Press, 1990.
- [22] Ronald P. Carver. "Reading Rate: Theory, Research, and Practical Implications." In: *Journal of Reading* 36.2 (1992), pp. 84–95. ISSN: 00224103. URL: <http://www.jstor.org/stable/40016440> (visited on 07/08/2023).
- [23] Seyyed Saleh Mozaffari Chanijani, Federico Raue, Saeid Dashti Hassan-zadeh, Stefan Agne, Syed Saqib Bukhari, and Andreas Dengel. "Reading type classification based on generative models and bidirectional long short-term memory." In: *Proc. IUI Workshops*. 2018, pp. 1–5.

- [24] Xiuge Chen, Namrata Srivastava, Rajiv Jain, Jennifer Healey, and Tilman Dingler. *Characteristics of Deep and Skim Reading on Smartphones vs. Desktop: A Comparative Study*. 2023. DOI: [10.1145/3544548.3581174](https://doi.org/10.1145/3544548.3581174). URL: <http://dx.doi.org/10.1145/3544548.3581174>.
- [25] Michael J. Cole, Jacek Gwizdka, Chang Liu, Ralf Bierig, Nicholas J. Belkin, and Xiangmin Zhang. *Task and user effects on reading patterns in information search*. en. 2011. DOI: [10.1016/j.intcom.2011.04.007](https://doi.org/10.1016/j.intcom.2011.04.007). URL: <http://dx.doi.org/10.1016/j.intcom.2011.04.007>.
- [26] Duchowski, Andrew. "Eye tracking techniques." In: *Eye Tracking Methodology: Theory and Practice* (2007), pp. 51–59.
- [27] Geoffrey B. Duggan and Stephen J. Payne. *Text skimming: The process and effectiveness of foraging through text under time pressure*. en. 2009. DOI: [10.1037/a0016995](https://doi.org/10.1037/a0016995). URL: <http://dx.doi.org/10.1037/a0016995>.
- [28] Geoffrey B. Duggan and Stephen J. Payne. *Skim reading by satisficing*. 2011. DOI: [10.1145/1978942.1979114](https://doi.org/10.1145/1978942.1979114). URL: <http://dx.doi.org/10.1145/1978942.1979114>.
- [29] John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. *Improving Students' Learning With Effective Learning Techniques*. en. 2013. DOI: [10.1177/1529100612453266](https://doi.org/10.1177/1529100612453266). URL: <http://dx.doi.org/10.1177/1529100612453266>.
- [30] Deborah Fallows. *Email at work*. Pew Internet & American Life Project, 2002.
- [31] Rebecca H Felton and Frank B Wood. "Cognitive deficits in reading disability and attention deficit disorder." In: *Journal of learning disabilities* 22.1 (1989), pp. 3–13.
- [32] Robert L. Fowler and Anne S. Barker. *Effectiveness of highlighting for retention of text material*. en. 1974. DOI: [10.1037/h0036750](https://doi.org/10.1037/h0036750). URL: <http://dx.doi.org/10.1037/h0036750>.
- [33] Miriam Fuhrman. "Developing good multiple-choice tests and test questions." In: *Journal of Geoscience Education* 44.4 (1996), pp. 379–384.
- [34] Luther C Gilbert. "Saccadic movements as a factor in visual perception in reading." In: *Journal of Educational Psychology* 50.1 (1959), p. 15.
- [35] Michael M Granaas, Timothy D McKay, R Darrell Laham, Lance D Hurt, and James F Juola. "Reading moving text on a CRT screen." In: *Human Factors* 26.1 (1984), pp. 97–104.

- [36] Jacek Gwizdka. *Characterizing relevance with eye-tracking measures*. 2014. DOI: [10.1145/2637002.2637011](https://doi.org/10.1145/2637002.2637011). URL: <http://dx.doi.org/10.1145/2637002.2637011>.
- [37] Spencer E Harpe. "How to analyze Likert and other rating scale data." In: *Currents in pharmacy teaching and learning* 7.6 (2015), pp. 836–850.
- [38] Wayne A. Hersberger and Donald F. Terry. *Typographical cuing in conventional and programed texts*. en. 1965. DOI: [10.1037/h0021690](https://doi.org/10.1037/h0021690). URL: <http://dx.doi.org/10.1037/h0021690>.
- [39] Nafiseh Hojjati and Balakrishnan Muniandy. *The Effects of Font Type and Spacing of Text for Online Readability and Performance*. 2014. DOI: [10.30935/cedtech/6122](https://doi.org/10.30935/cedtech/6122). URL: <http://dx.doi.org/10.30935/cedtech/6122>.
- [40] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.
- [41] Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise, and Andreas Dengel. *Towards reading trackers in the wild*. 2017. DOI: [10.1145/3123024.3129271](https://doi.org/10.1145/3123024.3129271). URL: <http://dx.doi.org/10.1145/3123024.3129271>.
- [42] Md. Rabiul Islam, Shuji Sakamoto, Yoshihiro Yamada, Andrew W. Vargo, Motoi Iwata, Masakazu Iwamura, and Koichi Kise. *Self-supervised Learning for Reading Activity Classification*. en. 2021. DOI: [10.1145/3478088](https://doi.org/10.1145/3478088). URL: <http://dx.doi.org/10.1145/3478088>.
- [43] Marcel A. Just and Patricia A. Carpenter. *A theory of reading: From eye fixations to comprehension*. en. 1980. DOI: [10.1037/0033-295x.87.4.329](https://doi.org/10.1037/0033-295x.87.4.329). URL: <http://dx.doi.org/10.1037/0033-295X.87.4.329>.
- [44] T Jin Kang and Paul Muter. "Reading dynamically displayed text." In: *Behaviour & information technology* 8.1 (1989), pp. 33–42.
- [45] Toshio Kawashima, Takanori Terashima, Takeshi Nagasaki, and Masashi Toda. "Enhancing visual perception using dynamic updating of display." In: *Intuitive Human Interfaces for Organizing and Accessing Intellectual Assets: International Workshop, Dagstuhl Castle, Germany, March 1-5, 2004, Revised Selected Papers*. Springer. 2005, pp. 127–141.
- [46] Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir R. Das, Dimitris Samaras, and Gregory Zelinsky. *Reading detection in real-time*. 2019. DOI: [10.1145/3314111.3319916](https://doi.org/10.1145/3314111.3319916). URL: <http://dx.doi.org/10.1145/3314111.3319916>.
- [47] Walter Kintsch. "The role of knowledge in discourse comprehension: a construction-integration model." In: *Psychological review* 95.2 (1988), p. 163.

- [48] Jumpei Kobayashi and Toshio Kawashima. *Paragraph-based Faded Text Facilitates Reading Comprehension*. 2019. DOI: [10.1145/3290605.3300392](https://doi.org/10.1145/3290605.3300392). URL: <http://dx.doi.org/10.1145/3290605.3300392>.
- [49] Thomas Kosch, Albrecht Schmidt, Simon Thanheiser, and Lewis L. Chuang. "One Does Not Simply RSVP: Mental Workload to Select Speed Reading Parameters Using Electroencephalography." In: CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–13. ISBN: 9781450367080. DOI: [10.1145/3313831.3376766](https://doi.org/10.1145/3313831.3376766). URL: <https://doi.org/10.1145/3313831.3376766>.
- [50] David LaBerge and S Jay Samuels. "Toward a theory of automatic information processing in reading." In: *Cognitive psychology* 6.2 (1974), pp. 293–323.
- [51] Manuel Landsmann, Olivier Augereau, and Koichi Kise. *Classification of reading and not reading behavior based on eye movement analysis*. 2019. DOI: [10.1145/3341162.3343811](http://dx.doi.org/10.1145/3341162.3343811). URL: <http://dx.doi.org/10.1145/3341162.3343811>.
- [52] James R. Lewis. *The System Usability Scale: Past, Present, and Future*. en. 2018. DOI: [10.1080/10447318.2018.1455307](http://dx.doi.org/10.1080/10447318.2018.1455307). URL: <http://dx.doi.org/10.1080/10447318.2018.1455307>.
- [53] Tsung-Ho Liang and Yueh-Min Huang. "An Investigation of Reading Rate Patterns and Retrieval Outcomes of Elementary School Students with E-books." In: *Journal of Educational Technology & Society* 17.1 (2014), pp. 218–230. ISSN: 11763647, 14364522. URL: <http://www.jstor.org/stable/jeductechsoci.17.1.218> (visited on 07/08/2023).
- [54] Robert F. Lorch Jr., Elizabeth Puzgles Lorch, and Madeline A. Klusewitz. *Effects of Typographical Cues on Reading and Recall of Text*. en. 1995. DOI: [10.1006/ceps.1995.1003](http://dx.doi.org/10.1006/ceps.1995.1003). URL: <http://dx.doi.org/10.1006/ceps.1995.1003>.
- [55] Hazel E Nelson and PAT McKenna. "The use of current reading ability in the assessment of dementia." In: *British journal of social and clinical psychology* 14.3 (1975), pp. 259–267.
- [56] Takehiko Ohno. "Eyeprint: support of document browsing with eye gaze trace." In: *Proceedings of the 6th international conference on Multimodal interfaces*. 2004, pp. 16–23.
- [57] Marilyn J Ransby and H Lee Swanson. "Reading comprehension skills of young adults with childhood diagnoses of dyslexia." In: *Journal of learning disabilities* 36.6 (2003), pp. 538–555.

- [58] Rayner, Keith. "Eye movements in reading and information processing: 20 years of research." In: *Psychological Bulletin* 124.3 (1998), pp. 372–422.
- [59] Keith Rayner. "Understanding eye movements in reading." In: *Scientific Studies of Reading* 1.4 (1997), pp. 317–339.
- [60] Keith Rayner and Monica S Castelhana. "Eye movements during reading, scene perception, visual search, and while looking at print advertisements." In: *Visual marketing: From attention to action* 1.1 (2008), pp. 9–42.
- [61] Keith Rayner, Albrecht Werner Inhoff, Robert E Morrison, Maria L Slowiaczek, and James H Bertera. "Masking of foveal and parafoveal vision during eye fixations in reading." In: *Journal of Experimental Psychology: Human perception and performance* 7.1 (1981), p. 167.
- [62] Keith Rayner, Elizabeth R. Schotter, Michael E. J. Masson, Mary C. Potter, and Rebecca Treiman. *So Much to Read, So Little Time*. en. 2016. DOI: [10.1177/1529100615623267](https://doi.org/10.1177/1529100615623267). URL: <http://dx.doi.org/10.1177/1529100615623267>.
- [63] William R. Reader and Stephen J. Payne. "Allocating Time Across Multiple Texts: Sampling and Satisficing." In: *Human–Computer Interaction* 22.3 (2007), pp. 263–298. DOI: [10.1080/07370020701493376](https://doi.org/10.1080/07370020701493376). eprint: <https://www.tandfonline.com/doi/pdf/10.1080/07370020701493376>. URL: <https://www.tandfonline.com/doi/abs/10.1080/07370020701493376>.
- [64] David Robert Reich, Paul Prasse, Chiara Tschirner, Patrick Haller, Frank Goldhammer, and Lena A. Jäger. *Inferring Native and Non-Native Human Reading Comprehension and Subjective Text Difficulty from Scanpaths in Reading*. 2022. DOI: [10.1145/3517031.3529639](https://doi.org/10.1145/3517031.3529639). URL: <http://dx.doi.org/10.1145/3517031.3529639>.
- [65] Michele Rucci, Paul V. McGraw, and Richard J. Krauzlis. "Fixational eye movements and perception." In: *Vision Research* 118 (2016). Fixational eye movements and perception, pp. 1–4. ISSN: 0042-6989. DOI: <https://doi.org/10.1016/j.visres.2015.12.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0042698915003648>.
- [66] Jeff Sauro. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC, 2011.
- [67] Andrew Sekey and Jerome Tietz. "Text display by saccadic scrolling." In: *Visible Language* 16.1 (1982), pp. 62–77.
- [68] Alexander Strukelj and Diederick C Niehorster. *One page of text: Eye movements during regular and thorough reading, skimming, and spell checking*. 2018. DOI: [10.16910/jemr.11.1.1](https://doi.org/10.16910/jemr.11.1.1). URL: <http://dx.doi.org/10.16910/jemr.11.1.1>.

- [69] Gail M Sullivan and Richard Feinn. "Using effect size—or why the P value is not enough." In: *Journal of graduate medical education* 4.3 (2012), pp. 279–282.
- [70] Sonya Symons and Michael Pressley. *Prior Knowledge Affects Text Search Success and Extraction of Information*. 1993. DOI: [10.2307/747997](https://doi.org/10.2307/747997). URL: <http://dx.doi.org/10.2307/747997>.
- [71] Juhana Venäläinen. "Getting texts done: affective rhythms of reading in quantified academia." In: *Affective Capitalism in Academia*. Policy Press, 2023, pp. 196–215.
- [72] Katarzyna Wisiecka, Krzysztof Krejtz, Izabela Krejtz, Damian Sromek, Adam Cellary, Beata Lewandowska, and Andrew Duchowski. "Comparison of webcam and remote eye tracking." In: *2022 Symposium on Eye Tracking Research and Applications*. 2022, pp. 1–7.
- [73] Gary S Wolverton and David Zola. "The temporal characteristics of visual information extraction during reading." In: *Eye movements in reading*. Elsevier, 1983, pp. 41–51.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both \LaTeX and \LyX :

<https://bitbucket.org/amiede/classicthesis/>

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

<http://postcards.miede.de/>

Final Version as of December 14, 2023 (`classicthesis` version 4.2).