Automatic Detection of Cognitive Load and User's Age Using a Machine Learning Eye Tracking System

Mina Shojaeizadeh

A Dissertation
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy
in
Information Technology
April 2018

Approved:

Dr. Soussan Djamasbi
PhD Adviser

Diane Strong

Dr. Diane Strong
Dissertation Committee Member

Dr. Randy Paffenroth
Dissertation Committee Member

Andrew C Trapp

Dr. Andrew Trapp
Dissertation Committee Member

Abstract

As the amount of information captured about users increased over the last decade, interest in personalized user interfaces has surged in the HCI and IS communities. Personalization is an effective means for accommodating for differences between individuals. The fundamental idea behind personalization rests on the notion that if a system can gather useful information about the user, generate a relevant user model and apply it appropriately, it would be possible to adapt the behavior of a system and its interface to the user at the individual level. Personalization of a user interface features can enhance usability. With recent technological advances, personalization can be achieved automatically and unobtrusively. A user interface can deploy a NeuroIS technology such as eye-tracking that learns from the user's visual behavior to provide users an experience most unique to them. The advantage of eye-tracking technology is that subjects cannot consciously manipulate their responses since they are not readily subject to manipulation. The objective of this dissertation is to develop a theoretical framework for user personalization during reading comprehension tasks based on two machine learning (ML) models. The proposed ML-based profiling process consists of user's age characterization and user's cognitive load detection, while the user reads text. To this end, detection of cognitive load through eye-movement features was investigated during different cognitive tasks (see Chapters 3, 4 and 6) with different task conditions. Furthermore, in separate studies (see Chapters 5 and 6) the relationship between user's eye-movements and their age population (e.g., younger and older generations) were carried out during a reading comprehension task. A Tobii X300 eye tracking device was used to record the eye movement data for all studies. Eye-movement data was acquired via Tobii eye tracking software, and then preprocessed and analyzed in R for the aforementioned studies. Machine learning techniques were used to build predictive models. The aggregated results of the studies indicate that machine learning accompanied with a NeuroIS tool like eye-tracking, can be used to model user characteristics like age and user mental states like cognitive load, automatically and implicitly with accuracy above chance (range of 70-92%). The results of this dissertation can be used in a more general framework to adaptively modify content to better serve the users mental and age needs. Text simplification and modification techniques might be developed to be used in various scenarios.

Acknowledgements

It is my opportunity to thank with all my heart people that have made me getting a doctorate possible. First and foremost, I would like to thank my adviser, Dr. Soussan Djamasbi for her professionalism blended with kindness. For helping me gain a strong knowledge in the field and for the effort she put into making me a better researcher. For trying to guide me in a direction where my research could make an impact. For all the help, support, and advice. It has been an honor to be her first PhD student. Next, I am very grateful to my committee members, Dr. Diane Strong, Dr. Andrew Trapp, and Dr. Randy Paffenroth, for their time and insight on my dissertation.

I am very thankful to Foisie Business School at WPI, for providing support throughout my PhD studies. I would like to especially thank Norm Wilkinson, Lorell Tross, Statia Canning, Laurie Stroke and Dr. Hall Philip for helping me recruiting participants for my studies.

I am especially indebted to Dean Ginzburg and Brent French, for providing financial support for incentives for my studies.

My special thanks goes to all the members of User Experience and Decision Making lab at WPI, especially Prateek Jain, Kodey Converse and Chris Chagnon for all their help, and most importantly, their valuable friendship! It’s been a great pleasure working with them in the lab.

I would like to thank my parents, whose unconditional love, prayers and guidance have gotten me where I am today. Without their support, this document and degree would not have been possible. I love you and this dissertation is dedicated to you.

I would like to thank my brother Shahab, and my sisters Rozita and Shima, for their endless courage, support and understanding! For always telling me that they are proud of me.

Last but not least, I would like to thank my husband, Hamed. For being a wonderful partner, fun friend, and excellent mentor. For being kind, patient, and thoughtful and for supporting me going through the toughest times of grad school.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Personalized User Experience

Nowadays, we are surrounded by a multitude of interactive devices and smart objects. In this situation, applications need to adapt to continuous changes of contexts, but only end users know the specific adaptations that they would like to have in their applications. In 1999, David Weinberger, a technologist and co-author of The Cluetrain Manifesto (Levine, Locke, Searls, & Weinberger, 1999), wrote, “Personalization: the automatic tailoring of sites and messages to the individuals viewing them so that we can feel that somewhere there’s a piece of software that loves us for who we are.” Interestingly, nearly 20 years later, personalization is being used by companies attempting to make the online experience more human. Personalization has grown rapidly since Weinberger’s statement, so that personalized experiences have become the norm, not an occasion. Personalization takes place by adjusting the system to suit the needs and preferences of a particular user. A useful and satisfying experience includes tailoring the experience to the individual and making it easier for a user to find relevant information or reach their intended goal. A properly personalized user interface improves users’ satisfaction and performance, compared to traditional manually designed “one size fits all” interfaces. In general, there are two basic kinds of personalization: adaptable and adaptive (Eichler, 2014). Both adaptable and adaptive approaches try to improve usability through personalization of interface from its default configuration. The adaptable approach means personalization performed by the user which is also known as customization, whereas the adaptive approach stands for personalization performed by the computer (Eichler, 2014).

Personalization of user experience requires collection of accurate and adequate user feedback. There are two primary techniques for collecting user feedback. Information can be obtained explicitly, by directly requesting the user to specify their feedback in a given piece of data on a specified rating scale, or implicitly, by observing user actions and inferring the user characteristics and cognitive behaviors (Mac Aoidh et al., 2009). One implicit method of collecting user's feedback is by measuring a user's cognitive load while interacting with an interface. Assessing cognitive load has long been an area of interest in the Information Systems and Human Computer Interaction literature (Stassen etal., 1990; Riedl, 2010; Buettner, 2014; Buettner et al., 2015). Cognitive load is defined as the relationship between the cognitive demands placed on a user due to a cognitive task and the user's cognitive resources (Wickens, 2002).

To design more intuitive systems that are easier to learn, and free of performance error, the first step is to model user's natural behavior (Oviatt, 2006). According to Human-Centered Design principles, cognitive load associated with extraneous complexity of system output needs to be minimized (Oviatt, 2006).

Personalization of a user interface is aligned with Intelligent Human Computer Interface (IHCI). An IHCI emphasizes that human behavior encompasses both apparent human behavior and the hidden mental state behind behavioral performance (Duric et al., 2002). IHCI integrates parsing and interpretation of nonverbal information with a computational cognitive model of the user, which, in turn, feeds into processes that adapt the interface to enhance operator performance and provide for rational decision-making (Duric et al., 2002).

One objective of this study is to collect nonverbal, objective and unobtrusive user's behavior via using eye tracking technology, and then interpret the user's behavior to build a user's mental model of interaction with an interface. In this study I focus on building a model that can detect user's cognitive load and user's age characteristics. Such models can be used further in enhancing the IHCI or Adaptive User Interfaces (AUIs).

Once the user's mental model is built, the next step for an intelligent or adaptive system is to diagnose the problems that influence user's experience negatively. Lastly, the final step is to adapt the interface to the user's need appropriately. Adaptive and intelligent HCI are important for novel applications of computing, including universal and human-centered computing (Duric et al., 2002). Recently, psychophysiological monitoring and neuro-feedback using bio-sensors have been recognized to provide accurate feedback from users' emotion and cognitive status, when interacting with interfaces (Dimoka, 2012). For example, an electrocardiogram could directly measure whether a certain interface increases the user's heart rate, thus inferring anxiety or stress. Eye tracking tools can implicitly measure user's cognitive load level and capture whether a user finds it difficult to interact with the interface by observing how the eyes move on a computer screen (Dimoka, 2012). Further details on how this technology is used to collect unobtrusive user feedback is provided in the next chapter section 2.1.

In this dissertation, eye tracking technology is used to implicitly and unobtrusively collect user feedback when the user is fulfilling different cognitive tasks such as problem solving and reading comprehension. Machine learning models are used to infer and predict user's age group and user's cognitive load, only through the eye movements. The use of predictive analytics and machine learning makes it possible to extend UX principles beyond personas all the way down to the individual user. This level of specificity can help guide users to content that is most preferable to their unique requirements. Many top companies have used this approach to deliver bull's-eye-precision targeted marketing while providing an environment that is engaging and unique to each user. The most pervasive examples of this are Amazon and Netflix. Amazon is able to track individual shopping habits, and suggest items that are highly relevant to each user. Amazon doesn't rely on focus groups or persona matching to find these items; they use machine learning (ML) and individual usage data to provide useful product recommendations. Until very recently, personalization has required face-to-face communication to identify user needs and adaptively revise the interface's content or appearance, but massive amounts of user data and machine-learning techniques can now augment these techniques to provide a better user experience.

Development of a personalized UX model that will be implemented in Adaptive User Interfaces (AUI), is the main focus of the present dissertation. "An AUI is a design that improves its ability to interact with a user by constructing a user model based on partial experience with that user" (Stathopoulos et al. 2002). AUI are at the intersection of Information System (IS), Human Computer Interaction (HCI) and Machine Learning (ML).

An adaptive UI changes dynamically in response to its experience with users (McTear, 2000). Adaptation to user requires a user model containing attributes of the user to which adaptations are sought. Such an effect is typically achieved by using dedicated machine learning algorithms responsible for acquiring and maintaining user characteristics and behaviors towards suitable interface adaptations (Goecks & Shavlik, 2000). The ML techniques are capable of expressing a rich variety of non-linear decision models. Such techniques, in general, process training/input data and attempt to make decisions or classification based on this input.

As mentioned earlier, the goal of present research is to develop a UX personalization model that consists of user age characterization and cognitive load detection. Figure.1.1 represents the general diagram of the theoretical framework that is proposed to address the user personalization while the user is interacting with an interface such as a website or a mobile application. Eye tracking is used to collect user's eye-movement while interacting with the interface. The eye-movement data is preprocessed and used to detect the user's cognitive state (e.g., high or low cognitive load) as well as user's age characteristics (e.g., young or old). Detecting user's level of cognitive load, and user's characteristics, will help in providing adaptive user personalization that can provide a personalized experience for individual users. With the advent of technology, adaptive user personalization will be applied in implementing adaptive user interfaces.

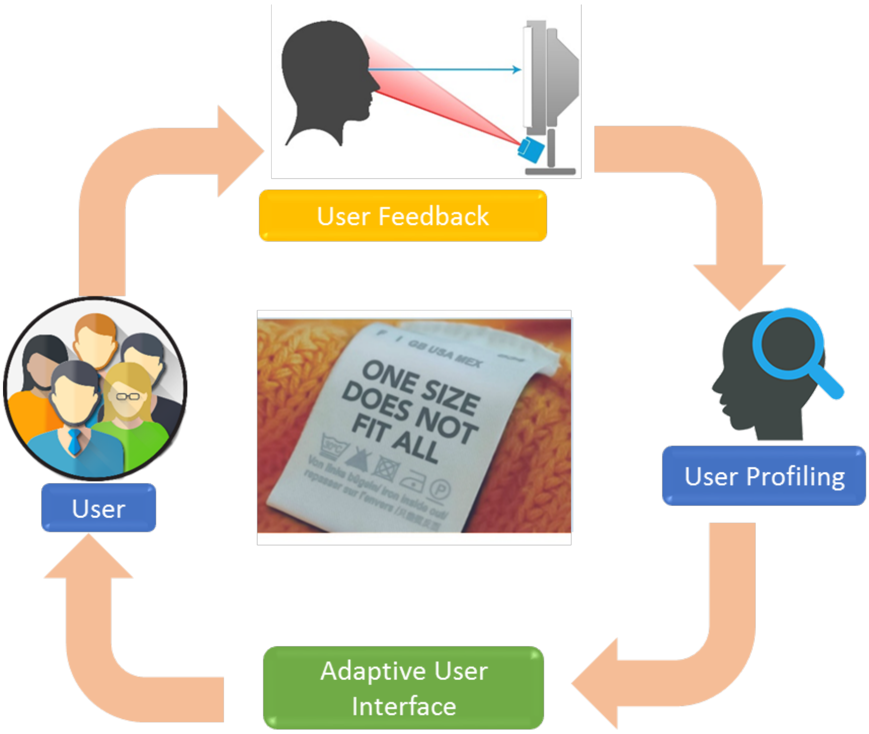


Figure 1. 1 UX personalization Model Diagram

1.2 Dissertation Contributions

To support the dissertation statement, my interdisciplinary research integrated the fields of machine learning, human computer interactions, eye tracking, statistics, and cognitive science. It makes several contributions that lay a foundation for HCI research by bringing cognitive load measurement through eye movements for HCI. In particular, with this dissertation, I make the following contributions:

1. Eye-movement Analysis: I describe novel methods of analyzing eye movement data, use newly defined eye movement features and develop classification methods and tools for eye movement analysis.

2. Cognitive Load Prediction and Machine Learning: I apply machine learning models and methods to analyze and predict cognitive load using eye-movement data. I show that cognitive load level due to different task demands is detectable and predictable in a complex cognitive task. I explored the two cognitive tasks of problem solving and reading comprehension because they have direct relevance to many HCI research studies.

3. Detecting Reading Difficulty Level Using Eye Movements: Using machine learning models I show that eye movements can be used as a proxy to detect the level of comprehension difficulty during a reading task.

4. Detecting Age Group of a User through Eye Movements: Using machine learning models I show that eye movements can be used to predict the age group of a user in a reading comprehension task.

5. Time Series Analyses of Pupil Dilation: I show that, pupillary data can be used continuously to understand the ongoing cognitive processes during reading. Further, I show that dividing the information processing and decision period of a task into smaller intervals provides valuable insight for examining the information processing behavior in real time.

1.3 Dissertation Roadmap

This dissertation is organized as follows: Chapter 2 provides an overview of related work that lays the foundation for this dissertation. The literature is presented which includes related background on the eye tracking technology and how the technology is used to collect eye movements of a user while fulfilling a task in front of a computer screen. I present the research works related to the difference between older and young users in online web experience, and how this differences emphasize the need for implementing adaptive user interfaces. Further, I will elaborate on the definition and theory of cognitive load, and will present related research that focus on measurement of cognitive load, using eye movements. This chapter also covers application of machine learning in eye-tracking HCI research.

As discussed above, detection of cognitive load and user's age characterizations are the main focus of my PhD studies. To this end, I conducted three different studies which will be presented in Chapters 3 to 5 of this dissertation.

Chapter 3 (study 1) focuses on detection of cognitive load (due to task condition) in a problem solving task. Cognitive load is manipulated as time limit in solving math problems (treatment group). A machine learning model is developed to classify the users into treatment and control (no time limit) conditions using their eye movements.

Chapter 4 (study 2) covers analysis of time-domain pupillary data to investigate cognitive load during a reading comprehension and decision making task. It is shown that pupil dilation time series during reading is not significantly different between groups of users who read passages with different task conditions. However, it was found that examining pupillary data in various time intervals can provide additional information for assessment of cognitive load.

The relationship between user's age and eye-movements during reading comprehension task is presented in Chapter 5 (study 3). Regression analysis is used to investigate how certain eye-movement metrics maps to user's age characteristics.

The final chapter (study 4) of the dissertation addresses the major goals of my PhD thesis, that is to develop two machine learning models using eye-movement features, one for user's age characterization (e.g. Generation Y, Baby boomers), and another for cognitive load during reading comprehension task.

2 Theoretical Background

This chapter presents the background material and research work relevant to this dissertation. It specifically looks at the motivation behind this research, a literature review on the eye tracking research, the state of the art in eye tracking, and how we can use eye tracking as a means of collecting implicit user feedback in real time. In this chapter I also look at the differences between older and younger generations, the two most important and distinct populations of the web users in the United States, and then conclude the chapter by reviewing how machine learning can be applied in eye tracking research.

2.1 Eye Tracking: An Unobtrusive Technology for Objective User Experience Assessment

Eye tracking refers to capturing the focus of a viewer's gaze on a stimulus at a given time. This is done by tracking a viewer's eye movements (Djamasbi, 2014). The predominant means to collect information from our environment is through the visual system, hence eye tracking provides an excellent tool for examining how people focus their attention and process information (Djamasbi, 2014). Eye tracking has been possible for many years, however, the earlier versions of eye tracker were very difficult to use because of large head mounted cameras which needed long set up times and difficult calibration procedures. With the extant growth of eye tracking technology, this technology is becoming increasingly popular in user experience research for investigating the interaction of users with user interfaces such as mobile and website (Albert & Tullis, 2013; Bergstrom & Schall, 2014). Eye tracking has been used by a number of researchers in the areas of Human-Computer Interactions, Marketing, Cognitive Psychology, and the new field of Neuro-IS, to detect where users look at a point in time, how long they look at something, and the path their eyes travel (Bergstrom & Schall, 2014).

An overview on how eye tracking technology works and how it is used in user experience research for capturing eye movements is given in the following. In addition, details are presented on the accuracy and precision of the eye tracking devices.

2.1.1 How does eye tracking work?

Most modern eye-trackers capture the eye-movements unobtrusively, using a method called video-based corneal reflection. The corneal reflection (or glint) is created by projecting infrared light into the eye, which also turns the pupil into a bright disc that makes its detection easier. The video-based corneal reflection method of eye movement detection captures the corneal reflection appearing as a small bright glint on the surface of the eye, as well as the center of the pupil (Djamasbi, 2014). The gaze location on the screen is calculated based on the relative position of glint and pupil center (Djamasbi, 2014).

Calibration

To measure the accurate location of the eyes on the screen, eye trackers need to be calibrated for each person. During the calibration process, the eye tracker learns how certain coordinates on the stimulus correspond to a person's eye position.

By asking the viewer to look at several dots on the computer screen or a calibration plate, the calibration process allows the eye tracker to associate the viewer's glint/pupil data with known locations on the stimulus. Figure 2.1 shows an example of a calibration process on a desktop computer. During calibration the user is asked to follow the movement of a red dot on the screen, without guessing where the future location of the dot is. The red spot in Figure 2.1 shows where the user's eye is fixated during calibration on a computer screen.

When looking at a stimulus, our eyes constantly move around to help construct a complete schema of what we are looking at. This process results in formation of two major types of eye movements: fixation and saccades (Djamasbi, 2014). Further details on the definition of different types of eye movements and the related literature is provided in section 2.2.2.

In the following I will present details about important attributes of an eye tracking device, which needs to be considered by researchers before conducting an eye tracking study. One of the most important features of an eye tracking device is its sampling rate (frequency) of the device.

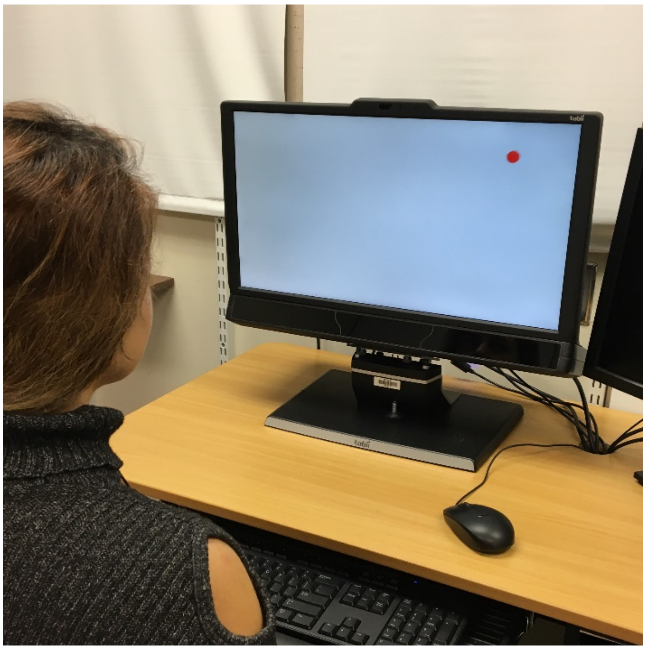


Figure 2. 1 The Calibration Procedure

Sampling Rate

An important attribute of an eye tracking device, which facilitates a continuous recording of eye movements, is sampling rate of the device. Frequency of the eye tracking device corresponds to how many times per second the eye position is measured. For example, an eye tracking device with a 60Hz sampling rate can record gaze points every 16.6 milliseconds, which is an adequate sampling rate for web studies (Djamasbi, 2014). However, for reading studies a higher sampling rate is required because higher sampling rates produce better temporal accuracy (also called temporal resolution) when measuring eye movements such as duration of fixations and saccades (Campbell & Bovee, J. C, 2014). Common sampling rates include 1,000 Hz, 300 Hz, 250 Hz, and 60 Hz. The average temporal error will be approximately half the duration of the time between samples (Campbell & Bovee, J. C, 2014). For example, a sampling rate of 300 Hz samples the eye position every 3.3 msec, which will lead to an average error of 1.65 msec. A sampling rate of 60 Hz, which samples eye position every 16.7 msec, will lead to an average error of approximately 8 msec. An 8 msec error in temporal resolution might be considered too large to study some of the eye metrics during reading. Thirty years ago most reading research was conducted using eye trackers with 60 Hz sampling rates. Most research on reading is now performed using eye trackers capable of sampling at 300 Hz or above.

Accuracy

Another important feature of an eye tracking device is the level of accuracy that it delivers in tracking the gaze. Accuracy is defined as the average difference between the real stimuli position and the measured gaze position. In other words, accuracy refers to how well the calculated fixation location matches actual fixation location ("Specification of Gaze Precision and Gaze Accuracy," 2016). This is measured in degrees of visual angle (a half circle has 180º of visual angle). Accuracy of the eye tracker is dependent on different factors such as illumination, gaze angle, and the distance of eyes to the screen. The average accuracy of the Tobii X-300 system when using a white stimuli background, and with average distance of 65cm to 75 cm to the screen, and with gaze angels of maximum 30º, is between 0.4 to 0.6º of visual angle ("Specification of Gaze Precision and Gaze Accuracy," 2016). When looking at a 17-20-inch computer monitor at a normal viewing distance, the width of the monitor covers 20-30º of visual angle. The degree of accuracy needed depends on the research goals. In a reading eye tracking study, if the goal is to measure which character on a line is fixated, then character position accuracy is needed. If the goal is to measure which word on a line is fixated, then word position accuracy is needed (Campbell & Bovee, J. C, 2014).

Precision

Precision is defined as the ability of the eye tracker to reliably reproduce the same gaze point measurement ("Specification of Gaze Precision and Gaze Accuracy," 2016). Precision is calculated as root mean-square (RMS) of successive samples ("Specification of Gaze Precision and Gaze Accuracy," 2016). Precision is dependent on distance from the eye tracker. For example, for a distance of 65 cm from the screen the precision of Tobii X300 device is about 0.07 degrees of visual angel according to ("Specification of Gaze Precision and Gaze Accuracy," 2016).

Figure 2.2 shows how accuracy and precision play important role in measuring the gaze location on the stimuli. The required level of accuracy and precision depends on the nature of the eye tracking study. Small uncertainties, for instance, can be critical when analyzing gaze data in reading studies or studies with a small stimulus. During data collection, accuracy and precision are used as indicators of the eye tracker data validity. A system with stronger accuracy and precision will provide more valid data as it is able to correctly describe the location of a person's gaze on a screen.

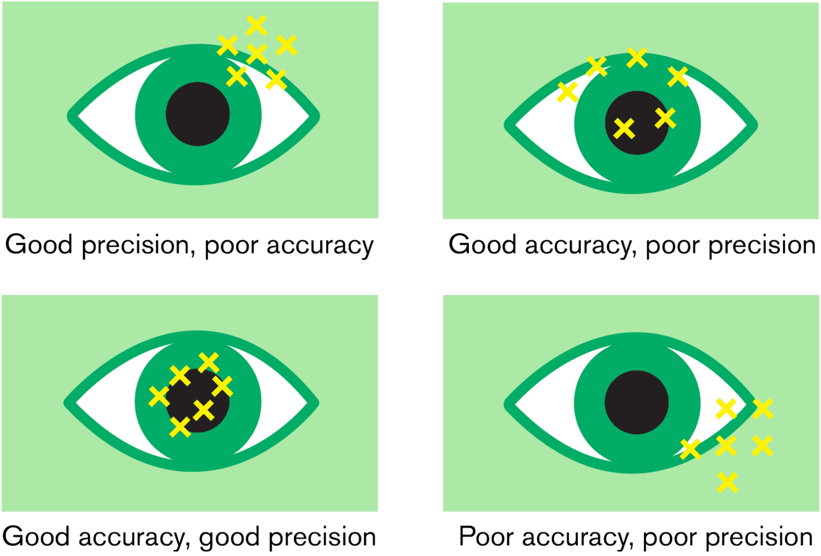

Good precision, poor accuracy
Good accuracy, poor precision
Good accuracy, good precision
Poor accuracy, poor precision

Figure 2. 2 Accuracy and Precision in Measuring Gaze Location on the Stimuli ("Eye tracker accuracy and precision," n.d.)

As mentioned earlier, the main objective of this dissertation is to develop machine learning models that can automatically predict user's level of cognitive load as well as user's age characteristics, using eye movements of user completing a cognitive task. In the next section I will provide an overview of cognitive load theory and different methods used to measure cognitive load. Further, I will explain the relationship between cognitive load measurement and different types of eye movements used in previous research.

2.2 Cognitive Load in Information Systems and HCI Research

Cognitive load (also referred to as mental effort) is defined as the relationship between the cognitive demands placed on a user due to a cognitive task and the user's limited cognitive resources (Wickens, 2002). The higher the cognitive load, the higher the chance is that the user will not complete a given task. High cognitive load affect how users make decisions and hence negatively influence user's judgment and performance. Thus, measuring a user's cognitive effort has been an important problem in Information Systems-Human Computer Interactions (IS-HCI) research from past to present (Buettner et al., 2015; Buettner, 2014; Riedl et al., 2010; Stassen et al., 1990). IS Scholars have traditionally investigated a user's cognitive effort based on the user's performance and subjective measurements (Ayyagari et al., 2011; Gupta et al., 2013; Ragu-Nathan et al., 2008; Tarafdar et al., 2010), however, such measures are intrusive or require a lot of equipment and expertise. Researchers in the field of NeuroIS have proposed determining a user's mental effort based on objective psychological measurements such as Eye Tracking, Skin Conductance Response (SCR) and Functional Magnetic Resonance Imaging (fMRI) (Dimoka, 2012; Riedl, 2010). In HCI, the complexity of the interface and the task complexity can jointly affect the user's attention or behavior due to different levels of cognitive load (Wang et al., 2014). Therefore, cognitive load associated with inessential complexity of the system needs to be minimized to improve user's performance in fulfilling a cognitive task (Oviatt, 2006). Being aware of a user's mental status is an important step in implementing a personalized user experience model. A successful user-centered design leverage from adapting to users' behavior and preferences (Attar, 2016).

The concept of cognitive load was introduced in Cognitive Load Theory (CLT) which was developed by John Sweller in the late 1980s (Sweller, 1988). Sweller identified three types of cognitive load: extraneous, intrinsic, and germane load. Intrinsic load is cognitive load that is inherent in the content or task to be learned, and is determined by the given complexity of the task. Extraneous load is cognitive load that has been introduced by the way information is presented, and could have been avoided by alternative presentation. An increase in extraneous cognitive load corresponds to an increase in additional information processing (Korbach et al., 2017). The better the format of information presented, the lower the amount of extraneous cognitive load (Brunken et al., 2003; Paas et al., 2003). Germane load is the amount of load dedicated to relevant information processing and mental model construction resulting in higher learning performance. An updated model of cognitive load theory (Choi et al., 2011) considers only two of the three components: intrinsic and extraneous load. The deletion of germane load was due to the close relationship between intrinsic and germane cognitive load. In this dissertation, I focus on detecting and predicting the extraneous cognitive load that is due to the task condition or the way information is presented to the user.

2.2.1 How can cognitive load be measured?

Measurement of cognitive load plays a crucial role in HCI research that focus on cognitive-load, and in developing practical implications for efficient universal design (Korbach et al., 2017). According to past research, cognitive load can be measured using two different methods. One is objective measurement, and the other one is subjective measurement.

Objective methods of cognitive-load measurement include the analysis of task performance and as well as the analysis of cognitive activity indicated by eye-tracking data (Brünken et al., 2010; Dimoka, 2012). Each of these objective methods is essential due to the continuous nature of the measurement. Objective measurement of cognitive load allows producing more detailed and accurate data and facilitates measurement of cognitive load continuously during the cognitive process. Subjective methods include ratings of perceived task difficulty, engagement or effort, which are completed by each participants (Korbach et al., 2017). Two examples are the subjective rating scale introduced by (Paas, 1992) and the NASA Task Load Index (NASA TLX) (Hart & Staveland, 1988), and SMEQ (Zijlstra & van Doorn, 1985).The advantage of subjective methods is that subjective rating scales are very easy to implement and can be used in different contexts. However, rating scales are criticized because of methodological problems concerning the criteria of objectivity, validity, and reliability (Brunken et al., 2003; Moreno, 2006; Brünken et al., 2010; Clark & Clark, 2010). In particular, it is difficult to distinguish between different types of cognitive load with a universal subjective rating scale. Another disadvantage is that subjective ratings are in general requested subsequent to the cognitive activities, which have to be evaluated by the subjects. Therefore, rating scales provide no continuous information about the actual cognitive load during the cognitive process (Korbach et al., 2017).

To enhance the user experience of a web page, cognitive load imposed by the interface should be minimal to free sufficient cognitive resources to process the contents of the website (Sweller, 2011). In other words, interfaces need to be designed in a way that they reduce the demand on user's cognitive capacity (Albers, 2011; Czaja & Lee, 2007).

Therefore, to enhance the user experience, it can be argued that there is a need for accurate, automatic and objective measurement of cognitive load. This measurement should not rely solely on the user's subjective rating of mental effort. Hence, one major focus of this study is on the automatic measurement of cognitive load when completing a task using eye movement data such as fixation, saccade, and pupil dilation. These eye movements are known in the literature to be the proxy for measuring cognitive load.

2.2.2 Cognitive load and eye movements

Eye-tracking is a powerful means of investigating cognitive load during information processing by offering a spatio-temporal record of visual attention (Hill et al., 2011). Eye movement analysis facilitates measures of allocation of attention and cognitive activities spent to process information (Korbach et al., 2017). According to several studies, there is a strong correlation between eye-movements and cognitive load (Just & Carpenter, 1976; Rayner, 1998; Holmqvist et al., 2011; Rosch & Vogel-Walcutt, 2013). Four major eye-movement data that are broadly used in the eye tracking literature that studied cognitive load are pupil dilation, fixation, saccade and blink. A brief review of each of these metrics and how they are used as measurement of cognitive load is given in the following. All of these eye movement metrics are readily obtained through typical eye-tracking technology.

Fixation

Fixation refers to a collection of relatively stable gaze points that are near in both spatial and temporal proximity. During fixation, the eyes hold steady on an object, and thus fixation reflects attention to a stimulus (Holmqvist et al., 2011; Poole & Ball, 2005). A number of studies have associated fixation related metrics to cognitive effort. For example, the number of fixations within an area of interest (AOI) has been used to compare cognitive effort of millennials and baby boomers when viewing a web page (Djamasbi et al., 2011). How frequently people fixate on an object has also been used to assess cognitive effort in business to consumer (B2C) transactional processes, when an option must be selected prior to continuing with the transaction (Hogan et al., 2015). Additionally, the number of fixations has been shown to strongly correlate positively with task performance (Van Orden etal., 2001). Because task performance is also correlated with effort spending (Payne etal., 1993), this result suggests a link between fixation frequency and cognitive effort.

Similarly, fixation duration, or the amount of time a user looks at stimuli, can be used to measure effort. To attend to a stimulus or an object, the user has to expend effort to keep his/her gaze steady on the object (Djamasbi, 2014). Moreover, studies provide evidence that fixation duration increases as information processing becomes more effortful (Van Orden et al., 2001; He & McCarley, 2010; Meghanathan et al., 2014).

Saccade

Saccades refer to small, rapid eye movements when jumping from fixating on one object to another (Goldberg & Kotval, 1999). While visual information is not processed during saccadic eye movements, they still can provide information about viewing behavior (Holmqvist et al., 2011; Jacob & Karn, 2003). For example, people tend to exhibit more saccadic eye movements when reading long words (De Luca et al., 2002). Similarly, saccade amplitude, or the path traveled by a saccade between two consecutive fixations, tends to increase when reading longer words (De Luca et al., 2002). When interacting with an online resource, longer saccadic amplitudes can reflect whether users have become familiar with an interface.

Having a better internal representation of an interface allows users to move their eyes directly to a desired location on the screen, hence producing longer saccadic amplitudes (Goldberg & Kotval, 1999). Consistent with this point of view, difficulty in locating information when browsing a webpage is likely to impact the duration of saccades. According to the theory of visual hierarchy (Faraday, 2000), a stimulus is inspected by scanning it through a sequence of visual entry points. Each entry point acts like an anchor, which allows the user to scan for information around it. According to this perspective, longer duration of saccadic eye movements could indicate increased cognitive effort in finding a suitable entry point into a visual display (Djamasbi, 2014).

Blink

Blinks are the involuntary act of shutting and opening the eyelids. They are known to reflect changes in attention and thus they are likely to reflect an individual's cognitive effort (Poole & Ball, 2005; Van Orden et al., 2001). In particular, fewer blinks have been associated with increased attention (Ledger, 2013). For example, a study shows that surgeons had a lower number of blinks when performing surgery as compared to when they were engaged in casual conversations (Wong et al., 2002).

In addition to the number of blinks, the duration of blinks can also indicate cognitive effort. For example, shorter blink durations were associated with increased visual workload during a traffic simulation task (Ahlstrom & Friedman-Berg, 2006). Similarly, comparing blink data during a hard (math problem solving) and easy task (listening to relaxing music), people exhibited shorter blink durations during the hard task (Andrzejewska & Stolińska, 2016). Because of its observed association with cognitive effort, blink duration has been used to assess mental effort in educational games (Ikehara et al., 2013).

The above studies suggest that people often exhibit fewer or shorter blinks during more challenging tasks because they want to minimize missing visual information. After all, when the eyes are closed during a blink, there is no incoming visual information to process.

Pupillometry

The use of pupil size as an indicator of cognitive processes dates back to the 1800s, and since the 1960s, there have been many studies into the behavior of the pupil in cognitive psychology. Changes in pupil size, which are controlled by the involuntary nervous system, can serve as a reliable proxy of cognitive load or mental effort (Laeng, et al., 2012). Many pupillometric studies have suggested that there is a link between pupil size and cognitive load. For example, when people are asked to memorize numbers, retain them in memory, or perform multiplication, the size of their pupil seems to correlate with the difficulty of the task (Hess & Polt, 1964; Kahneman & Beatty, 1966; Schultheis & Jameson, 2004; Iqbal et al., 2005; Bailey & Iqbal, 2008; Piquado, Isaacowitz, & Wingfield, 2010). Iqbal et al. (2004) and Bailey and Iqbal (2008) presented a framework for detecting task boundaries based on pupil dilations as a measure of cognitive load. In all of these studies, pupil size was shown to be a reliable indicator of task difficulty.

There has also been a great research interest in understanding the cognitive load in a visual search task and its relationship with pupillary responses. Porter et al. (2007) used pupillometry to study cognitive load during visual search. Based on their findings, when individuals perform a visual task, pupil size appears to be a function of the cognitive effort and attention required. Klinger (2010) found a link between mental effort and pupillary responses during map reading and searching tasks. According to his findings, looking up for a given locality caused significant differences in pupil size compared to legend reading.

Similar to pupil dilation, variation in pupil dilation can also carry information about cognitive load (Buettner et al., 2015; Shojaeizadeh et al., 2015; Fehrenbacher & Djamasbi, 2017) . For example, the level of difficulty measured as number of steps required to complete a task has been shown to impact pupil dilation variation (Buettner et al., 2015). Increased cognitive load measured as implicit and explicit time limit also has a significant impact on pupil dilation variation. It is argued that pupil dilation variation is particularly effective in detecting the impact of complex decision tasks on users because these tasks often involve a number of smaller subtasks. These subtasks are likely to require different types of mental activity with varying levels of difficulty. Consequently, complex decision tasks may result in variability in pupil size over the course of the task (Buettner et al., 2015). Another explanation for the suitability of pupil dilation variation in measuring cognitive load is rooted in the Adaptive Decision Making theory which asserts people often switch their information processing strategies to conserve their limited cognitive resources.

This flexibility in adjusting to the decision environment, which involves balancing one's cognitive load, is likely to be detected by the variation in pupil dilation (Fehrenbacher & Djamasbi, 2017; Shojaeizadeh et al., 2017). Table 2.1 provides the list of the eye movement behaviors, their respective parameters and some of the supporting studies, as discussed in the above sections.

Table 2. 1 Eye Movement Behaviors and Parameters for Measuring Cognitive Effort

Behaviour	**Parameter**	
Fixation: Relatively stable gaze points close in proximity and time	Fixation count	(Djamasbi et al., 2011; Hogan et al. 2015; Van Orden et al., 2001)
	Fixation duration	(Djamasbi, 2014; He & McCarley, 2010; Just & Carpenter, 1980; Meghanathan et al., 2014; Van Orden et al., 2001)
Saccade: Rapid eye movements between fixations	Saccade count	(De Luca et al., 2002)
	Saccade duration	(Djamasbi, 2014)
	Saccade amplitude (the distance travelled between two adjacent fixations)	(De Luca et al., 2002; Goldberg et al., 2002)
Blink: Involuntary act of shutting and opening the eyelids	Blink count	(Ledger, 2013; Poole & Ball, 2005; Van Orden et al., 2001; Wong et al., 2002)
	Blink duration	(Ahlstrom & Friedman-Berg, 2006; Ikehara et al., 2013; Andrzejewska & Stolińska, 2016)
Pupillary Response: Changes in pupil dilation	Pupil dilation (size of pupil diameter)	(Beatty, 1982; Kahneman & Beatty, 1966; Piquado et al., 2010; Chen et al., 2011)
	Pupil dilation variation (derivative of pupil diameter or rate of change in pupil size)	(Buettner et al., 2015; Fehrenbacher & Djamasbi, 2017; Shojaeizadeh et al., 2015; Shojaeizadeh et al., 2017))

2.3 Eye Movements Characteristics in Reading

This section surveys the literature for eye movements that are relevant to information processing during reading.

Eye movements have been broadly used by pioneers in the reading psychology literature who have used eye tracking to understand the link between eye movements and reading behavior (Rayner, 1998; Ashby etal., 2005; Rayner 2009; Rayner K. & Pollatsek A., 2012; Campbell &

Bovee, 2014;) This is because movements of the eyes are a natural part of reading, and hence studying eye movements provides important information into the user's covert cognitive processes during reading. According to Rayner (1998) eye movements can be used to understand the ongoing cognitive processes that occur during reading. Further, Rayner et al. (2006) indicated that eye movements can be used to reflect text difficulty in reading. Eye movements have been also used to understand whether or not and how textual information is processed (Iqbal et al., 2004; Salojärvi et al., 2005). For example, Gustavsson (2010) and Campbell et al. (2014) used eye movements to detect whether a person was reading a text or not. Similarly, eye movements were used to detect whether users read or skim textual information (Buscher et al., 2008). During reading, we naturally make sequences of fixations and saccades. Fixations and saccades can vary in duration and frequency, which is the result of how the information are being processed in our brain. For example, during reading, re-fixating on a word or fixating for a longer time on that word could indicate that the reader is uncertain about the semantics of the current sentence and needs to return to a previously read word to comprehend the sentence. During these saccades and fixations, the pupil dilates and contracts as a physiological response. Emotional and cognitive events along with other factors from the environment such as brightness can cause the pupil to constrict or expand. Accurate assessment of these meaningful signals is crucial for research in HCI as an indicator of cognitive load (Attar, 2016).

When reading, fixation duration is around 200-300 milliseconds, with a range of 100-500 milliseconds and saccades range in duration between 10-20 msec for short between word saccades, and between 60-80 msec for longer saccades from end of one line to the beginning of another line (Rayner, 1998).

The majority of saccades during reading English are made from left to right, however, in skilled readers, about 10-15% of the saccades are regressive, they are backward saccades to the previously read words or lines (Rayner, 1998). In general, saccades during reading are divided into two categories: 1) Progressive saccades in the direction of reading text, 2) Regressive saccades, or backward saccade, to the opposite direction of reading (Rayner, 1998). Short within-word regressions can be due to problems in processing the currently fixated word. Longer regressions (more than 10 letter spaces back along the line or to another line) are because of the difficulties in comprehension, or may be because the text is particularly difficult and the reader cannot understand the text (Rayner, 1998). Regression depends on the difficulty of the text. As the texts become more difficult, saccade size decreases, fixation duration increases, and regression increases.

According to previous research, eye movements provide several important advantages as a measure of reading behavior. First, observing eye movements provides the ability to examine text processing demands at a global level (across an entire text), the sentence level (individual sentences), or the local level (individual words or phrases) (Campbell & Bovee, 2014).
This is because for example, changes in global difficulty lead to changes in several measures of eye movements, such as total reading time, number of forward fixations, and the number of regressions. Changes in local level difficulty also affects several measures, such as reading times for individual words, the probability of fixating words, and likelihood of making regressions to specific words (Campbell & Bovee, 2014). It is important to note that while eye movement analyses during reading provide such details, overall reading times or sentence-by-sentence reading times do not provide such detailed measures of reading behavior.
A second advantage of using eye movement in assessing reading performance and behavior, is that eye movements are a natural part of reading (Campbell & Bovee, 2014); therefore, no additional task demands are placed on a reader. Furthermore, eye tracking provides different types of eye movements (e.g., fixation duration, saccade amplitude, and regression count), and hence facilitates processing various elements of reading process. Eye movements also reflect individual differences in readers, since they vary according to reader's ability (Ashby et al., 2005), prior knowledge about a topic (Kaakinen, Hyönä, & Keenan, 2003), and age of the reader (Rayner et al., 2006). Due to the above mentioned characteristics, eye movement serves as an ideal assessment of reading behavior (Campbell & Bovee, 2014).

One way to summarize the eye movements is to present averages of various eye movement metrics over a large segment of text such as a passage, a paragraph, or a set of sentences. These measures, such as the mean fixation duration, the mean duration of forward and backward saccades, have been shown to globally reflect the difficulty of the reading process. For example, reading a more difficult passage results in longer mean fixation durations, shorter average forward saccade duration, and more regressions (Rayner & Pollatsek, 2012).

A number of researchers have used eye tracking to examine the differences between the two most populated age groups in the US, meaning Baby Boomers and Generation Y, during an online web experience. Detecting user's age characteristics is an important component of this dissertation. To personalize the user experience for each age group, one needs to know the differences between the two groups in online web experience.

2.4 Personalization and Age: Differences between Older Adults and Generation Y in Online User Experience

In this section the differences between older adults specifically Baby Boomers and Generation Y in online reading and web experience is explained, based on the literature.

Baby Boomers, born between 1946 and 1964 (age in 2018, 54 to 72) are the second largest generation in the U.S. With 74.9 million people in 2015, they make up 26 percent of the total U.S. population ("Baby Boomers and Credit generational," n.d.). People 65+ represented 14.5% of the population in the year 2014; they are expected to grow to be 21.7% of the population by 2040. The internet provides a number of benefits for older adults. It is used as a means of communication via E-mail, chat rooms, discussion groups, and direct messaging. The Internet also contains a wealth of medical information that can be particularly useful for older adults when health becomes a greater issue and concern.

Generation Y (born in 1980s and 1990s) are the largest population in US with 82 million reported in 2015. Generation Y are an economically powerful generation, spending $200 billion annually (Djamasbi et al., 2010). Gen Y are known as incredibly sophisticated, digitally and technologically, as they have been exposed and grew up with technology since their early childhood (Djamasbi et al., 2010). Therefore, to better accommodate Generation Y user experience, corporations need to understand their needs and preferences and adapt to them accordingly, to stay competitive in the marketplace.

In the following section I provide an overview of the literature that examined the older adult (e.g., Baby Boomers) web experience. Then I discuss the differences in online web experience of these two generations, and will further discuss the importance of considering age as an important construct in implementing a personalized UX model. The majority of these studies have used eye tracking technology as an effective tool for examining the older or younger adult's web experience.

2.4.1 Older Adult Web Experience

Prior research demonstrate that older adults are slower in cognitive processing than younger people (Fisk et al., 2012) and hence may expend more cognitive effort when processing online information. Slowdowns of cognitive processing in older adults can be explained by the cognitive theory of aging, which elucidates the age related differences between older and younger adults (Kemtes & Kemper, 1997; Salthouse, 1996). According to this theory, declines in performance of older adults in accomplishing a cognitive task (e.g., in reading) is due to general functioning or specific difficulties in older adults, which results in predictable age related differences between older and younger adults (Kemtes & Kemper, 1997; Salthouse, 1996).

In addition to cognitive processing, aging can result in progressive decline in visual, auditory, and motor skills, which significantly influence the web experience of older people, and hence results in the differences between older adult web experiences as compared to generation Y (Chadwick-Dias et al., 2004; Czaja & Lee, 2007; Money, et al., 2010). For example, cognitive declines due to aging results in longer processing and response times for older adults as compared to their younger counterparts (Arch et al., 2009; Czaja & Lee, 2007; Priest et al., 2007).

In addition, researchers who have investigated the older adult web usability have reported that even though older adults use the internet often, they experience difficulty when using it (Chadwick-Dias et al., 2003; Boechler et al., 2006; Brandtzæg, Lüders, & Skjetne, 2010; Hertzum & Hornbæk, 2010; Romano Bergstrom et al., 2013). Further, according to the findings of (Chadwick-Dias et al., 2003), Baby Boomers often have different usability issues as compared to young adults, and this is due to numerous reasons such as social, cognitive, psychological, and physical factors as well as overall differences in life experience.

2.4.2 Eye Tracking and Older Adult Web Experience

Older adults are likely to have more problems than younger ones with conventional usability methods (Fisk et al., 2012). Therefore, an especially valuable benefit of eye-tracking is its application with older adults, and hence, a lot of researchers have used eye tracking as a means of investigating older adult's web experience and usability.

Psychological eye-tracking literature has reported that older adults show age-related slowing which reflects in their eye movements. For example, Abel & Douglas (2007) demonstrated that older adults, as compared to younger adults, are slower to react to stimuli and their performance is more variable, while Kemper et al. (2004) and Kliegl et al. (2004) demonstrated that older adults read more slowly and make more fixations and regressions than younger readers.

Rayner et al., (2006) examined the eye movement characteristics of older people. They asked young and old adults to read sentences containing target words that varied either in frequency (low-frequency vs. high-frequency target words) or in predictability (low-predictable, medium-predictable, or high predictable target words) to determine whether frequency and predictability interact with age when these target words are read. They learned that older adults make more and longer fixations (a mean of 260 ms compared to 246 ms for younger readers) and more regressions. Overall, these studies suggest that older adults are performing at a reasonable level and they are successful in reading but it seems that they make more effort in reading as demonstrated by longer fixation durations (Hill et al., 2011). Nevertheless, Kliegl et al., (2004) showed that the differences between older and younger adults in reading do not necessarily have a significant impact on reading performance (at least where reading is "easy").

Despite the advantages that eye-tracking offers for research with older adults, the field is currently small and has potential for improvement. First, only a small number of web usability studies have been carried out with older adults. There is scope for further and more detailed research in this area. Some of the studies show results in the form of heatmaps, rather than giving detailed spatio-temporal eye-movement metrics such as fixation or saccade. In addition, the eye trackers used in most of the studies had relatively low sample rates, ranging from 30 Hz (Fukuda & Bubb, 2003) to 60 Hz (Josephson & Holmes, 2004). There is also less reports on higher order cognitive tasks such as reading textual information, which is an important part of web experience. Despite the small number of reports, outcomes have been relatively consistent. Researchers conclude that older adults are slower and hence experience lower usability.

2.4.3 A Comparison between Older and Younger Adults in Online Web Experience

One of the early study involves an eye-tracking examination of railway timetables (Fukuda & Bubb, 2003), with younger (mean: 22.5) and older users (mean: 67). They concluded that older people made longer fixations due to visual difficulties reading small text (point 10). It also took older people a longer time to complete tasks as compared to younger people. A similar finding emerged from two studies looking at expert older users working at an investment company, Fidelity. In their first study, they examined the behavior of expert older adults (50-69), who worked in the office and used the web daily, compared to younger colleagues (20-39) (Tullis, 2007). The results showed that older adults spent on average 42% more time looking at the content of the pages than did the younger adults. They also spent 51% more time looking at the navigation areas as compared to the younger adults. Their results also suggest that older participants distributed their gaze more widely across the pages and read more of the text than younger users did. In another study, conducted at Fidelity, researchers examined preferences for web page presentation, and reported that older adults aged 44-62 tended to fixate for longer on large images and search bars as compared to their younger counterparts (Capozzo et al., 2008). In another study, Zaphiris & Savtich, (2008) compared older (58-87) and younger (19-27) web users browsing health information sites of varying depth of hierarchy, and concluded that older adults looked at more of the page and spent longer considering which link to choose. The researchers found no significant difference in reading speeds, suggesting that older adults were not simply slower than younger ones.

Chadwick-Dias et al. (2003), conducted two usability studies to investigate the differences between older and younger adults in completing a task using a prototype employee/retiree benefits page on Fidelity's website. In their first study they examined whether there were differences in how older adults interact with the web and whether changes in text size would affect performance. Users completed tasks on the website using various text sizes. They learned that older users (55 years or older, mean=69.2) had significantly more difficulty using the web than younger users (55 and younger, mean=35.9), but that did not significantly affect performance in any age group. In their second study new participants performed the same tasks on a version of the site that was redesigned to address the usability problems encountered by older users in the first study, with the purpose of improving the performance of older adults. The results were that performance improved significantly for both older and younger users. They also observed that older users tended to read more text and often read all the text on a screen. Additionally, they found that older adults were particularly cautious and not confident about clicking on links that were nouns, like Accounts. When they changed those links to actions, like Go to Accounts, both older and younger users were faster and more confident. Coyne et al. (2002) found that their older adult participants distribute their gaze more widely across the pages and read more than younger adults in their studies. They also concluded that older adults were more likely to read messages, prompts, and pop-ups than younger study participants.

In a more recent study, Djamasbi et al. (2011) examined the differences between old and young adults in reactions to a set of homepages through a laboratory experiment. Users' reactions were captured using self-report measures and eye tracking. Their results showed that both generations reported similar visual preferences, and both generations preferred pages that had images and little text. However, the two generations also displayed differing online viewing behavior and preferences. For instance, eye tracking data revealed that Baby Boomers had significantly more fixations and that their fixations covered more of the pages than younger generation. In addition, Baby Boomers reported a significantly higher tolerance for having more web components on a page (Djamasbi et al., 2011).

Loos & Romano Bergstrom, (2011), examined the differences between older and younger users in information search behavior. Researchers asked the participants to complete a search task related to health information on three different websites. Researchers then investigated whether age or other factors such as gender, educational background and frequency of internet use have the biggest impact on navigation patterns, the use of the search box, effectiveness, and efficiency and user satisfaction. According to their findings, older users were less likely to make use of the search box than younger users. In addition, younger users were more successful in accomplishing the search task and were much faster than their older counterparts in completing the task. Although there are some differences between older and younger users in fulfilling an information search task, the greatest factor affecting information search behavior is not always age. For example, when comparing the navigation patterns of older people using internet daily with those of younger age group no significant differences were observed between the two groups.

Doube and Beh (2012) studied factors influencing the interaction of cognitive processing with visual and motor skills during website use by older adults. Twenty-eight older adults and 18 younger adults completed an on-line air-ticketing search task. Their results showed that compared with the younger group, older adults took significantly longer and made significantly more errors. Although age was a major contributing factor resulting in longer task times of older adults, less experience and less regular practice were also contributing factors. In addition, the larger number of search errors made by older adults was also correlated with age. Furthermore, although errors in search results were correlated with age alone, greater experience with website use could free cognitive resources for problem-solving and possibly would improve task performance.

It is worth restating that aging slows down the cognitive processing which results in different viewing and reading behavior between old and young adults. Older users read almost all the text that appears on a screen (Chadwick-Dias et al., 2003), are more patient in reading and spend more effort when reading passages (Fukuda & Bubb, 2003), while younger users dislike reading and show less patient reading behavior (Djamasbi et al., 2011).

Research suggests that Generation Y and Baby Boomers differ in their sense of immediacy and patience. This difference is attributed to the two generations' differing experiences during maturation; in particular, Baby Boomers were not exposed to the rate of change and advancement that the Internet provided Generation Y (Locher, 2009). From a technological perspective, Generation Y has matured in a fast-paced world that facilitated instantaneous response and encouraged multitasking (Felthousen, 2008). Because Generation Y has grown up in such an environment, this generation tends to have a shorter attention span and tends to be less patient than Baby Boomers, who, as they were growing up, did not have access to the technologies that were available to the younger generation.

Hence, it can be argued that there is a noticeable distinction between older and younger user's behavior while interacting with user interfaces. This suggests that there is a need in the user experience research to take into account the differences in reading behavior of the users via a user personalization model. User personalization has been an interest of recent research in the field. I know discuss instances of the more distinguished studies and user personalization approaches in the following section.

Fisk et al. (2012) show that presenting simple contents that are in plain language is extremely important especially when it comes to older adults, since they have difficulty drawing inferences from complex text. Turns and Wagner (2004) suggest limiting the reading level to the lowest possible level, such as eighth grade, or testing audience literacy of the domain to ensure that the content is written to the appropriate degree of simplicity. Other researchers ask content developers and information designers to present technical information in a non-technical way that is easy to read and understand (Coyne, 2002; Craik & Salthouse, 2000). Theofanos and Redish (2003) recommend implementing short, clear, straightforward sentences. This supports ease of skimming but also supports information retention and reduces memory load. Cognitive researchers e.g., Fisk et al. (2012) have found that older adults can be more easily distracted by extraneous information and that as people get older they have more and more difficulty making inferences. Therefore, they suggest that information should be organized in ways that show how the pieces are related.

Chadwick et al., (2003) found that older people are more cautious about clicking links and are more likely to click links that explicitly tell them what will happen when they click. They also observed that older users do not readily recognize links, even when they are blue and underlined. They also found that older users may not be familiar with terms like URL, Home, or Back. Based on these findings, they provided some design suggestions to help to improve the user experience, especially for older people. They suggest using action word links, making link style and color consistent and obvious throughout the site, using scalable fonts and options to increase text size, provide options for increasing the size of text on the site, and keeping the language simple.

Grahame et al. (2004) compared the abilities of older and younger adults in performing a visual search task on a Web page. The task involved determining the presence of a blue, underlined hyperlink, the most common means of indicating potential targets in sequential Web search and general browsing. They manipulated link size, location, number of distracting links, the amount of clutter, and presence of the target hyperlink. The results of their eye tracking experiment produced several suggestions for improving the user experience of web search for older adults. For example, they suggest that hyperlinks should be salient and in locations expected by the common user to facilitate improved search. Specifically they suggest increasing relative size, changing used and unused link colors so that they contrast with the background, grouping links, and using common standards (such as underlined text to mark a hyperlink) for assisting people in finding and recognizing the information. Additionally, decreasing clutter and the number of links will improve a person's ability to sort through a web site, and this is especially true if that person is an older adult.

Important online information is often conveyed via text-based communication. Thus, examining the reading behavior of older users and comparing it to those of younger users allow designers to better meet the need of both user groups. To personalize the UX by age and cognitive load first we need to learn about the user's age group and whether or not they experience cognitive load, unobtrusively, via a machine learning eye tracking model. In the next section I present a survey on the applications of machine learning in eye tracking research.

2.5 Application of Machine Learning in Eye Tracking Research

Machine learning (ML) and classification approaches have previously been used in eye-tracking research to automatically analyze eye-movement data. The objective is to find some computational structure that describes a link between the low-level raw data and high-level behavioral units. Machine learning focuses on developing models that can automatically learn from data. Because machine learning models learn to perform tasks by generalizing from examples, they are often more cost effective than manual programming (Domingos, 2012). The goal of machine learning is to develop a system that can learn from a given set of data, so as to make predictions about a yet unseen set of data. One of the most mature and widely used types of machine learning is classification. Herein our focus is on supervised classification to predict categorical responses.

This section provides a review of relevant eye tracking studies that use machine learning to predict a variety of different behaviors such as information retrieval, performance, intention, distraction, reading and domain expertise. The findings of these studies, which show the effectiveness of combining eye tracking data with machine learning approach, provide support for developing an eye tracking, machine learning system for predicting user's characteristics and information processing behavior.

Given the ability of classification to generalize from examples, it has been used in a number of eye tracking studies to predict behavior. A supervised learning approach was applied to a set of eye movement metrics (fixation count, total fixation duration, mean fixation duration, and regression duration) to predict information retrieval behavior on mock search results pages (Salojärvi et al., 2003). A similar approach was applied in the area of Content-Based Image Retrieval (CBIR) to predict from the eye movement data (total and average length of fixations and fixation count) whether the retrieved images were relevant to the search terms used (Klami et al., 2008). Machine learning was also used to predict from eye movement data (fixation count, mean and standard deviation of fixation duration, mean and standard deviation of saccade length and saccade direction) whether a user is searching for a word, an answer to a question, or looking up the most interesting title in a given list (Simola et al., 2008). They achieved about 60% prediction accuracy when inferring in which of three states a user can be during information search tasks.

Using eye movement data (fixation duration, path distance or saccade amplitude, fixation count, fixation rate), a classification approach was also used to predict how well people would solve a puzzle (Eivazi & Bednarik, 2011). They applied a Support Vector Machine (SVM) based approach for learning and classifying cognitive states during problem-solving, and achieved an accuracy of approximately 53% on the classification problem. In addition to predicting task performance, classification has been used to predict user intention from their eye movement data (saccade length, saccade duration, saccade velocity, and saccade acceleration). The authors developed a classification system to predict whether study participants intended to give a command to a gaze-based interface (Bednarik et al., 2012). Another study used machine learning classification from eye movement data of people collaborating on building concept maps (gaze count and gaze duration) to distinguish expert participants from novice participants

(Liu et al., 2009). Other instances of a machine learning approach for eye-tracking data analysis have been reported by (Bednarik et al., 2005) and (Kinnunen et al., 2010). The authors have applied a state-of-the-art biometric person authentication system based on traditional signal processing methods. Using eye movement velocity and pupil size, or a histogram of the velocity and gaze direction, the authors achieved identification rates of 60% or equal error rates of 29%, respectively. Table 2.2 summarizes the above and more applications where machine learning along with eye-tracking technology was used.

Table 2.2 List of Machine Learning and Eye-tracking Research Studies in the Literature

Reference	**Application**	**# of Participants or data set**	**Eye-movement Features**	**Machine Learning Method(s)**	**Performance**
Salojarvi et al. (2003)	Predicting information retrieval behavior on mock search results pages	60	Fixation count, total fixation duration, mean fixation duration, and regression duration	Linear Discriminative Analysis (LDA)	80.5%
Marshal, (2007)	Classifying users who perform an arithmetic problem-solving task vs. a doing-nothing task.	NA	Pupil size and point-of-gaze	Neural Network	70%
Klami et al., (2008)	Inferring the relevance of images based on eye movement data (predicting from the eye movement data whether the retrieved images were relevant to the search terms used)	27	Total and average duration of fixations and fixation count	LDA	91.2%

Simola et al., (2008)	Predicting whether a user is searching for a word, an answer to a question, or looking up the most interesting title in a given list.	10	Fixation count, mean and standard deviation of fixation duration, mean and standard deviation of saccade amplitude and saccade direction	Discriminative hidden Markov model	60.2%
Liu et al., (2009)	Classifying expert participants from novice participants on building concept	64	Fixation count and fixation duration	Profile hidden-Markov model	96%
Richstone et al., (2010)	Classifying surgeons into expert and non-expert cohorts based on their eye-data gathered during simulated and live surgical environments.	22 highly unbalanced sample	Pupil dilation	Linear discriminate analysis and nonlinear neural network	Around 90%
Eivazi et al., (2011)	Predicted how well people would solve a puzzle	14	Fixation duration, Saccade amplitude, Fixation count, Fixation rate	Support Vector Machine	53%

Bednarik et al., (2012)	Task-independent prediction of interaction intents (predicted whether study participants intended to give a command to a gaze-based interface)	12	Saccade amplitude, saccade duration, saccade velocity, and saccade acceleration	Support Vector Machine	76%
Kardan & Conati (2012)	Classifying students' performance in a problem-solving task using eye-movements	50	Fixation rate, number of fixations and fixation duration, saccade amplitude, relative saccades direction and absolute saccade direction	Decision Tree based, Support Vector Machine, Linear Ridge Regression, Logistic Regression Neural Network	71%
Henderson et al., (2013)	Participants were engaged in four tasks over 196 scenes and 140 texts: scene search, scene memorization, reading, and pseudo reading.	12	Mean and standard deviation of fixation duration, mean and standard deviation of saccade amplitude, number of fixations per trial.	Naive Bayes	Over 80%
Schneider et al., (2013)	Predicting students' learning scores using eye-movement data	75	52 features: (7 Fixations), (42 Saccades). And (3 pupil size).	Support Vector Machine Naïve Bayes Logistic Regression	93%
Borji et al., (2014)	Subjects' gaze patterns were recorded while performing one of a number of image-viewing tasks (like estimating the ages of people shown in the picture or their material circumstances).	21	Fixation map and scan path	Neural Network	Around 60%

Najar et al., (2014) [60]	Classifying novices and advanced students when learning from examples	22	Fixation duration, total fixation duration, fixation count, visit duration	Decision Tree	Between 61% to 89%
Steichen et al., (2014) [61]	Classifying: users attending different cognitive tasks, tasks complexity, visual Working Memory, verbal Working Memory, visualization type, perceptual speed	35	Fixation rate, Number of fixations, Fixation duration, Saccade amplitude, Relative saccade angles, absolute saccade angles	Decision Tree, Support Vector Machine, Neural Network Logistic Regression	Between 55-70%
Krol & Krol (2017) [62]	Classifying two strategic decision tasks that differ in terms of the information about the counterpart's behaviour that players are provided with and to choose her own strategy accordingly.	92	Pupil dilation and gaze dispersion	Neural Network	67%

2.6 Research Questions

While previous research offers machine learning as a potential solution for detection of human cognitive and visual behavior from eye tracking data, little is known about an automatic user profiling using various features of the eye movements. In this research I attempt to fill the research gap in building an automatic and unobtrusive user profiling model which is based on cognitive load and age, two important factors affecting user's experience according to the above literature review. I address the following research questions in building the user profiling model.

RQ1: Can a machine learning model predict the cognitive load from the eye movement data of a user doing time-limited problem solving task?

RQ2: Which eye movement features best describe the extraneous cognitive load due to a task condition during a problem solving task?

RQ3: Can a machine learning model predict the age group of a user, in a reading comprehension task, via their eye movement?

RQ4: Can a machine learning model predict the cognitive load of a user, in a reading comprehension task, via their eye movement?

RQ5: Which eye movement features best describe the extraneous cognitive load due to task condition, and age group of the user during a reading comprehension task?

To address RQ1 and RQ2 I designed study one (Chapter 3), where I developed machine learning model that can predict cognitive load level of a user doing a math problem solving task.

Studies two and three are designed to address RQ3 to RQ5. In Chapter 4 I focus on assessing cognitive load from pupillary responses during reading. In Chapter 5, I assess user's age characteristic using eye movement data during reading. In Chapter 6, I develop a user profiling model based on user's age group and the level of cognitive load.

3 Study One: Detecting Cognitive Load in a Problem Solving Task

The objective of this study is to address RQ1 and RQ2 presented in Section 2.6, that is to predict cognitive load due to solving time-limited math problems automatically, using eye movement metrics. As described in chapter 2, section 2.2.2, eye movements are used in research as means of measuring cognitive load (Beatty, 1982; Iqbal et al., 2004; Buettner et al., 2015). Pupil dilation, for example is known to be a reliable measure of cognitive load, in a problem solving task. The following section provide the details about two studies (pre and main) that were conducted related to this line of research.

3.1 Pre Study: Does Pupillary Data Differ During Fixations and Saccades? Does it Carry Information about Task Demand?

3.1.1 Introduction

Recent technological advances have made it possible to capture a user's experience of a system through the analysis of his or her eye movement data. IS researchers have used eye movement data to examine users' attention, awareness, search behavior, preferences, and behaviors (Cyr et al., 2009; Djamasbi, 2014).

Pupillometry can serve as a reliable measure of cognitive load. However, exploring pupil data is a relatively new phenomenon in IS-HCI research. In this study, I argue that a user's pupil data may be different during saccades and fixations. Moreover, I explore the relationship between task demand and pupil data. Fixations and saccades represent two different types of eye movement events. The former is used to collect visual information to send to our brain for processing, while the latter is used to scan our visual field for the next fixation event (Djamasbi, 2014). Pupil data refers to changes in pupil dilation/constriction as well as variation in such changes. Because pupil dilation is an involuntary reaction that has been shown to represent cognitive activity (Buettner et al., 2015), I tested to see whether pupil dilation during fixations and saccadic eye movements are different. While prior studies have used pupil data to capture cognitive load, to my best knowledge no study has looked at pupil data within fixations and saccades separately. As explained above, fixations and saccades reflect very different types of eye movements, so it is likely that pupil data differ during fixations and saccades. Thus, I expect to see differences in pupil data during these two types of eye movement events.

I also tested to see whether pupil data during fixation and saccades carry information about task demand. To do so, I examined the relationship between subjective perception of task demand and objective measure of pupil data.

The results of this study were published in the *Proceedings of 14th annual Pre-ICIS workshop on HCI Research in MIS*, Fort Worth, Texas, in December 2015.

3.1.2 Methodology

Eighteen graduate students, participated in the study, completed 10 GRE math questions in 5 minutes. Tobii X300 eye tracking device and Tobii Studio version 3.2.3 were used to collect the eye movement data, while the user was completing the task. IV-T filter with 30 deg/sec saccadic velocity threshold was used to filter the raw gaze data into fixation and saccades. In addition, subjective experience of task demand was measured via NASA Task Load Index (TLX), with 5 dimensions of Mental Demand, Physical Demand, Time Demand, Performance, Effort, and Frustration (Lin & Imamiya, 2006). Pupillary data were exported from the eye tracking software for analyses. Pupil dilation data during saccade events were separated from pupil dilation data during fixation event. Two paired t-tests were used to investigate the differences between pupillary data (pupil dilation and pupil variation) during saccade and fixation. Pupil dilation variation was calculated as the rate of change of pupil dilation as proposed by Iqbal et al. (2005). In addition, the relationship between subjective experience of task demand (TLX items) and pupil dilation and pupil variation during saccades and fixations was examined via regression analysis.

3.1.3 Results

The results of t-tests (Table 3.1) showed that the mean value of pupil dilation during fixation (3.079) was almost the same as the mean value of pupil dilation during saccades (3.078). The results of t-tests (Table 3.2) showed that pupil dilation variation during fixation (0.166) was significantly (p=0.016) smaller than pupil dilation variation during saccade (0.169).

Table 3. 1 Paired t-tests comparing users' pupil dilation (mm) during fixations and saccades

	Pupil Dilation	
	Mean	SD
Saccades	3.078	0.461
Fixation	3.079	0.462
	df= 17, t Stat= 0.93, p=0.36	

Table 3. 2 Paired t-tests comparing users' pupil dilation variation during fixations and saccades

	Pupil Dilation Variation	
	Mean	SD
Saccades	0.169	0.061
Fixation	0.166	0.061
	df= 17, t Stat= 2.68, p=0.02	

In addition, I tested the relationship between subjective experience of task demand (TLX items) and pupil dilation and pupil dilation variation during saccades and fixations via regression analysis. The results (Table 3.3 and Table 3.4) showed a significant relationship between task demand and pupil variation (in saccades: R^2 =0.25, p=0.035, B=51.08, in fixations: R^2=0.24, p=0.041, B=51.64), however, the results did not show any significant relationship between task demand and pupil dilation. The effect size for the significant results was rather large and slightly larger for fixation (f^2 = 0.33) than for saccades (f^2 = 0.31). The unstandardized coefficient (B) was also slightly larger for fixation as compared to saccades. In summary, these analyses showed that pupil dilation was not significantly different between fixations and saccades. However, there was a significant difference in pupil dilation variation between these two eye movement events. These results also showed that pupil dilation variation had a strong significant correlation with the Time Demand dimension of TLX, while I did not find the same relationship between TLX and pupil dilation data. These results suggest that pupil data during saccades and fixations can be different, and that may be useful to consider for in some studies. The results also suggest that pupil dilation variation may be more sensitive in terms of revealing differences between fixation and saccadic eye movements. Because these results indicate that pupil data may carry information about a user's subjective experience of task environment, they provide a new direction for using pupillometry in studying user experience.

Table 3. 3 Regression Analysis (Y: Pupil Variation during Saccade, X: TLX Variables)

	Mental Demand	**Physical Demand**	**Time Demand**	**Performance**	**Effort**	**Frustration**
R^2	0.15	0.03	**0.25**	0.10	0.07	0.05
P-value	0.11	0.48	**0.03**	0.20	0.28	0.37
B	36.82	14.09	**51.08**	27.39	22.97	-14.96

Table 3. 4 Regression Analysis (Y: Pupil Variation during Fixation, X: TLX Variables)

	Mental Demand	**Physical Demand**	**Time Demand**	**Performance**	**Effort**	**Frustration**
R^2	0.15	0.03	**0.24**	0.09	0.07	0.05
P-value	0.12	0.51	**0.04**	0.23	0.30	0.40
B	36.89	13.51	**51.64**	26.37	22.43	-14.42

3.2 Main Study: Automatic Detection of Task Condition in a Problem Solving Task

As mentioned above, the goal of this study is to detect extraneous cognitive load due to a task condition (e.g., time limit) using eye movements while completing a math problem solving task. Cognitive load is a major factor that influences how computerized systems are used to make decisions. High cognitive load contributes to difficulty in use, as well as a lack of adoption. Responding to the need of users to ease demand on cognitive resources thereby provides an opportunity for improving the effective use of decision aids at a personalized level. A first step is to gain a more informed understanding of the cognitive load experienced by users. Such information should be collected in an automated, accurate, and seamless manner. In this study, I argue that eye movements can provide invaluable data for unobtrusive and automatic detection of cognitive load. I test this assertion by developing an eye tracking machine learning system to detect the level of cognitive load experienced during a math problem-solving task both in the presence, and absence, of a time limit threshold. The results support my hypotheses, revealing that eye movements provide valuable information about cognitive load. The machine learning system implemented in this study can reliably predict increased mental effort from the eye movement data.

From the results of this study a manuscript is being prepared, and targeted to be submitted to the *Decision Support Systems* Journal.

3.2.1 Introduction

Despite the growing importance of computerized decision making and problem solving in a globally connected world, studies suggest that computerized tools may not always be used to their full potential. For example, while computers can augment an individual's information processing capacity, people seem to use them in a way to reduce their effort rather than to improve their decision accuracy (Todd & Benbasat, 1991, 1992, 1994). According to Adaptive Decision Making theory, this behavior is not due to inherent laziness or indifference (Payne et al., 1993), whereas accurate, rational decisions are the intention, due to limited cognitive capacity a natural consequence is the conservation of cognitive resources (Payne et al., 1993; Simon, 1955). Consequently, cognitive effort plays a major role in users' technology usage behavior (Davis, 1989; Todd & Benbasat, 1991, 1992, 1994). Because people place a high value on effort reduction, a comprehensive understanding of user experience of cognitive load provides excellent opportunities for designing computerized decision tools that are used more effectively.

In particular, if decision tools can detect changes in cognitive load, it is reasonable to conclude that they can more successfully respond to user need in an adaptive way. For example, an adaptive decision tool that can detect user experience of high cognitive load might then provide feedback or suggestions for the user to help ease cognitive effort or more effectively use limited cognitive resources (Todd & Benbasat, 1999; Barkhi et al., 2005; Hess et al., 2005; Shah et al., 2011). The absence of such cognitive information, however, makes it difficult, if not impossible, to envision such an adaptive decision tool.

In this study, I develop and test a machine learning system that can reliably, automatically, and unobtrusively detect the experience of cognitive load through analyzing eye movement data. Detecting cognitive load using eye tracking has several important advantages. Because eye movements reflect how people visually inspect stimuli, and because vision is our most dominant sense (Dähne, et al., 2014; Andrzejewska & Stolińska, 2016), eye tracking provides a natural method for examining information processing behavior. Eye trackers collect eye movements continuously, thereby providing a comprehensive picture of behavior. Moreover, modern remote eye trackers are integrated into monitors, or can be easily attached to such visual displays. Hence, they can collect eye movement behavior unobtrusively, without requiring users to wear special gear (Holmqvist et al., 2011; Poole & Ball, 2005; Djamasbi, 2014).

The integration of machine learning technology into eye-tracking devices holds promise not only for a dynamic and flexible mechanism for detecting user experience of cognitive load, but also one that is easily scalable. The advent of machine learning approaches carries the promise of discovering meaningful insights even on data sets of massive size. Because machine learning systems can generalize from a given set of data (Domingos, 2012), advanced machine learning eye tracking systems are bound to improve over time as their respective eye tracking data set grows every time they are used. Moreover, as eye tracking technology matures, high quality remote eye trackers become increasingly affordable (Djamasbi, 2014). This in turn, not only makes developing advanced machine learning eye tracking systems possible but also cost effective.

Based on Adaptive Decision Making theory (Payne et al., 1993), eye tracking research (Holmqvist et al., 2011), and machine learning literature using eye tracking datasets (Eivazi & Bednarik, 2011), I argue that a machine learning approach is a reliable and effective way to use gaze data to predict the cognitive demand of a user. This assertion was tested via a laboratory experiment that manipulates level of cognitive load using two different task conditions. In the following sections I provide a brief review of relevant theory and literature, thereby establishing the framework for my research. I subsequently form two hypotheses and discuss the methodology that was used to test these hypotheses. I will also discuss the method that was used to develop an eye tracking machine learning system, and will report the results of the tests for examining the effectiveness of this system in predicting the level of cognitive demand experienced by a user.

3.2.2 Hypothesis

Task conditions have a major impact on the level of demand placed on cognitive resources.

To conserve the inherently limited cognitive resources required for processing information, decision makers skillfully adjust their efforts in the way they go about solving a problem so that they can meet the demands of the task at hand. A number of studies, summarized in Table 2.1 in Chapter 2, suggest that eye movements can carry information about cognitive load. Hence, it is likely that effort used to meet task condition is reflected through the way information is processed. Thus, I hypothesize that eye movement data can provide information about user experience of task condition:

Hypothesis 1: Eye movement data will carry information about task condition experienced by a user.

Modern remote eye tracking devices allow us to collect information about user gaze unobtrusively and seamlessly (Holmqvist et al., 2011). The inherently rich and vast amount of eye movement signals collected for a user have been shown to provide suitable information for developing predictive machine learning systems. As reviewed in section 2.6, studies have demonstrated that machine learning can automatically learn from data to predict users' domain knowledge, intention to give a command, distraction during driving, performance in a puzzle game, and patterns of information search and retrieval (Salojärvi et al., 2003; Klami et al., 2008; Simola et al., 2008; Liu et al., 2009; Eivazi & Bednarik, 2011; Bednarik et al., 2012). Moreover, as discussed earlier and summarized in Table 2.1, studies suggest that eye movement data can carry information about cognitive load. Hence, I hypothesize that it is possible to develop a machine learning system using eye movement data that can predict task demand:

Hypothesis 2: The developed eye movement classification system can predict user experience of task condition.

3.2.3 Eye Tracking Experiment

To test the hypotheses, the first step was to build the eye tracking machine learning system. To do so, I conducted a laboratory experiment to collect eye movement data for a cognitively complex problem solving task under two different task conditions. In this section I explain the methodology for the eye tracking experiment.

3.2.4 Task

The problem-solving task used in this study asked participants to provide correct answers to a set of ten mathematical questions. This set of questions were manually selected from a pool of problem-solving practice tests for the Graduate Record Examination (GRE), which is a standardized test required for admission to most graduate degree programs in the United States.
The full set of these practice questions were retrieved from www.majortests.com. These math questions were then used to develop an online multiple choice math test.

3.2.5 Experimental Design & Participants

I used a time constraint to manipulate the level of demand placed on cognitive resources (Ferrari, 2001). It is well-known that a time constraint increases the use of cognitive resource by making problem-solving tasks more demanding (Payne et al., 1993). I created two treatments by manipulating the time available for completing the task. In the control treatment no time limit was enforced, while in the experimental group the time available for completing the task was set to five minutes. This created two treatments with different task conditions: control treatment (lower cognitive load) and experimental treatment (higher cognitive load). Participants were randomly assigned to either a control or experimental group. Participants in both groups completed the same problem-solving task, however, in the experimental group participants had to complete the task within five minutes, while in the control group they could take as long as they wished to complete the task.

Because GRE math problems were used for the problem-solving task, I recruited participants from the pool of graduate students in various technical disciplines (e.g., computer science, electrical and computer engineering, robotics engineering, etc.) at WPI. Because students are accustomed to taking timed tests, the task and setting created an appropriate and realistic environment for our participants.

The eye movements of 48 participants (21 female, and 27 male, ages ranging between 24 and 31) were collected during this study. The eye movement data for half of the participants was collected under a time constraint, while the remaining 24 were collected from participants that had no time limit. I used Tobii X300 remote eye tracker with a sampling rate of 300 Hz mounted on a 21-inch monitor at a resolution of 1920 x 1200 to collect the gaze data. To track eye movements, each participant completed a brief eye-calibration process. While seated, participants were asked to observe a moving dot on the eye-tracking monitor. This calibration process took less than one minute to complete.

3.2.6 Measurements

I used the I-VT filter with 30°/sec saccadic velocity threshold, provided in the Tobii Studio software version 3.2.3, to identify fixations and saccades in the gaze stream. Saccade amplitude (the distance traveled between two adjacent fixations), measured in degrees, as well as pupil dilation (size of pupil diameter) was also calculated by the Tobii Studio software. Pupil Dilation Variation (PDV) or rate of change of pupil dilation was calculated by taking the temporal derivative of pupil dilation (Iqbal et al., 2005; Van den Brink et al., 2016). Blinks were calculated as complete eye closure lasting between 100-500 milliseconds (Aarts et al., 2012).

3.2.7 Developing the Eye Tracking Machine Learning System

The eye tracking machine learning system developed for this study performs a classification problem. Classification refers to the process of identifying the correct category for a new piece of information based on prior observations. In this case, classification refers to identifying whether eye movements were collected under lower or higher level of cognitive load. To achieve this goal, the classifier must first be trained with a set of (eye movements, task condition) data. That is, during this training phase the system has access to both the collected eye movement data as well as the task condition under which the data was collected. With a successful training, the system will be able to take as input a new set of eye-movement data only, without information about task condition, and reliably predict the task condition under which the data was collected.

To develop the eye tracking machine learning system to classify tasks into higher and lower cognitive demands, I conducted the following three steps. First, grounded in the eye tracking literature discussed in section 2.2.2, I determined a set of variables, or feature set, that were likely to be most effective in predicting cognitive load. Second, I prepared the collected data set for training and testing the system by removing seven outliers from the data. Finally, I selected a classification algorithm for developing and testing the eye tracking machine learning system. In the following sections, I discuss how each step was carried out.

Step 1: Feature selection

The first step of the development process required the identification of relevant data attributes for designing our machine learning classification system. A list of relevant eye-movement features supported in the literature and summarized in Table 2.1 were used. Each parameter was measured over the duration of the task completed by each participant in the study. Machine learning feature sets are often developed using statistical properties of fundamental parameters. Hence, basic statistical properties, such as mean and standard deviation, were calculated for each of the parameters in Table 2.1. According to my previous study, pupil data during the saccadic and fixation events has been shown to differ (Shojaeizadeh et al., 2015), thus I considered pupil data for fixations and saccades separately.

In addition to calculating the average duration values for saccades, fixations, and blinks, I also considered their normalized duration metrics. Normalization was carried out by dividing the total duration of the fixation, saccade, or blink parameter by the total task completion time. Additionally, certain eye movement behaviors were combined to develop ratios that could provide additional insight. For example, the ratio of saccades to fixations reveal the amount of time spent searching for information, versus the amount of time spent on processing the information visually (Djamasbi, 2014). This in turn can provide insight about cognitive effort (Goldberg & Kotval, 1999). Together, the feature set of the eye tracking machine learning system consisted of thirty different eye metrics. This feature set is displayed in Table 3.5.

Table 3. 5 Feature Set: List of Eye Movement Metrics for the Machine Learning System

Eye Movements	Eye Metrics (Features)
Fixation	Average fixation duration (millisecond) Standard deviation of fixation duration Normalized fixation number (fixation number/task completion time) Normalized total fixation duration (total fixation duration/task completion time)
Saccade	Average saccade duration (millisecond) Standard deviation of saccade duration Average saccade amplitude (degree) Standard deviation of saccade amplitude Normalized saccade number (saccade number/task completion time) Normalized saccade duration (total saccade duration/task completion time)
Blink	Average blink duration (millisecond) Standard deviation of blink duration Normalized blink number Normalized blink duration (total blink duration/task completion time)
Pupil Dilation	Average pupil dilation (PD) during fixation (millimetre) Standard deviation of PD during fixation Average pupil dilation variation (PDV) during fixation Standard deviation of PDV during fixation Average PD during saccade (millimetre) Standard deviation PD during saccade Average PDV during saccade Standard deviation of PDV during saccade
Combined Eye Movements	Eye Metrics (Features)
Ratios	Average (PD during saccade/ PD during fixation) Standard deviation of (PD during saccade/PD during fixation) Average (saccade duration/fixation duration) Standard deviation (saccade duration/fixation duration) Average (PDV during saccade/PDV during fixation) Standard deviation (PDV during saccade/PDV during fixation) Normalized saccade duration/normalized fixation duration Normalized saccade number/normalized fixation number

Step 2: Selecting an Algorithm

Machine learning algorithms are typically selected based on the complexity of the problem at hand. The purpose of the present study is to detect the level of effort expenditure based on user eye movements during the decision-making process in a math problem solving task. According to Adaptive Decision Making theory (Payne et al., 1993), people expend effort to balance the conflict between maximizing accuracy and minimizing effort using various strategies, and moreover this effort expenditure is highly contingent upon task conditions. In particular, during complex tasks under time limit people are likely to switch between multiple strategies to meet the task demand (e.g., they may increase their processing speed, use less information, and/or switch to a less demanding strategy such as heuristics) (Payne, Bettman, & Johnson, 1988). This flexibility in decision behavior suggests the need for an algorithm that is suited for processing complex models.

I used the Random Forest (RF) classification problem by creating several individual models, or trees, using bootstrapping (Hastie, Tibshirani, & Friedman, 2009). Individual trees are developed by randomly selecting sub-samples from the original dataset. Each individual tree is a type of classifier that uses the divide-and-conquer methodology combined with bootstrapping. Individual trees are considered weak learners in the random forest framework. The algorithm generates a strong learner by combining the individual weak learners into a single overall tree that can produce more accurate results than any of the weak learners (Hastie et al., 2009).

Step 3. Developing the Proposed Classifier

Figure 3.1 displays the bootstrapping algorithm that was used to develop the random forest classifier. The bootstrapping methodology caused each sample to appear exactly *200* times in the computation. Each data point was taken with equal probability, hence some of the samples may have appeared several times in the bootstrap set and others not at all. Next, the data generated by bootstrapping was divided into “training” and “test” sets (80% and 20%, respectively) (Hastie et al., 2009). The training dataset was used to train the classifier, that is, the subsamples in the training set were used to create individual trees for the random forest classifier. To emphasize, the random forest classifier was computed using the 80% of the data marked as the training data (Figure 3.1). Thus the random forest classifier was *totally unaware of the remaining 20% of data that has been reserved for testing.*

```
Step 1. Initialization

1.1 Set number of replications $i$ = 200
```

Step 2. Training and Test

2.1 Generate at random training sets out of the feature matrix dataset and use these for training the untrained classifier. Training set generation is done "with replacement".
2.2. The resulting trained classifiers are tested on the corresponding test data.
2.3. This procedure is then repeated *i* times.

Step 3. Classifier Accuracy

3.1. Compute the classification error at each replicate.
3.2. Calculate the bootstrapping generalized error by averaging over the errors of all *i* classifiers.

Figure 3. 1 Bootstrapping Procedure

After the training was completed, the test dataset was used *to assess the accuracy* of the random forest classifier in predicting task condition. That is, I assessed the performance of the random forest classifier created with the training data (80%) with the remaining 20% of data put aside for testing. By doing so, I was now able to ask the random forest machine learning algorithm questions about task condition *for which the answer was known to me, the experimenter, but unknown to the random forest machine learning system.* This split of the data allowed me to accurately test the predictive performance of the random forest for real world scenarios.

One common way to estimate the predictive effectiveness of a classifier is to measure its performance (in our case the level of error in answering questions about the task condition) on the unseen test data. Resampling methods are commonly used to estimate the generalization error of classifiers (Efron, 1979; Beran, 1992; Rao et al., 2008). I used a bootstrapping methodology to test the accuracy of the trained classification system. As shown in Figure 3.2, each tree (RF$_i$) was trained with a bootstrapping sample (training data set) and tested with the remaining data in the original set (test data). The accuracy of the classifier was then measured by comparing the output of each individual tree with the task condition of its test data. If there was a match the error variable for that particular subtree was set to 0, or 1 otherwise. The average error value for the subtrees represented the generalized error for the random forest classifier. I used 200 Bootstrap replications as suggested by (Efron & Tibshirani, 1994). A very large bootstrap replication is not suggested as it results in a computational burden.

The machine learning algorithms discussed in this section were implemented in R version 3.4.2 on Windows 7, with Core i5 CPU and 3.30 GHz speed machine. I used R libraries including ISLR (James et al., 2018), tree (Brian & Ripley, 2018), random forest (Cutler, 2014), e1071 (Meyer et al., 2017), and caret (Max et al., 2018).

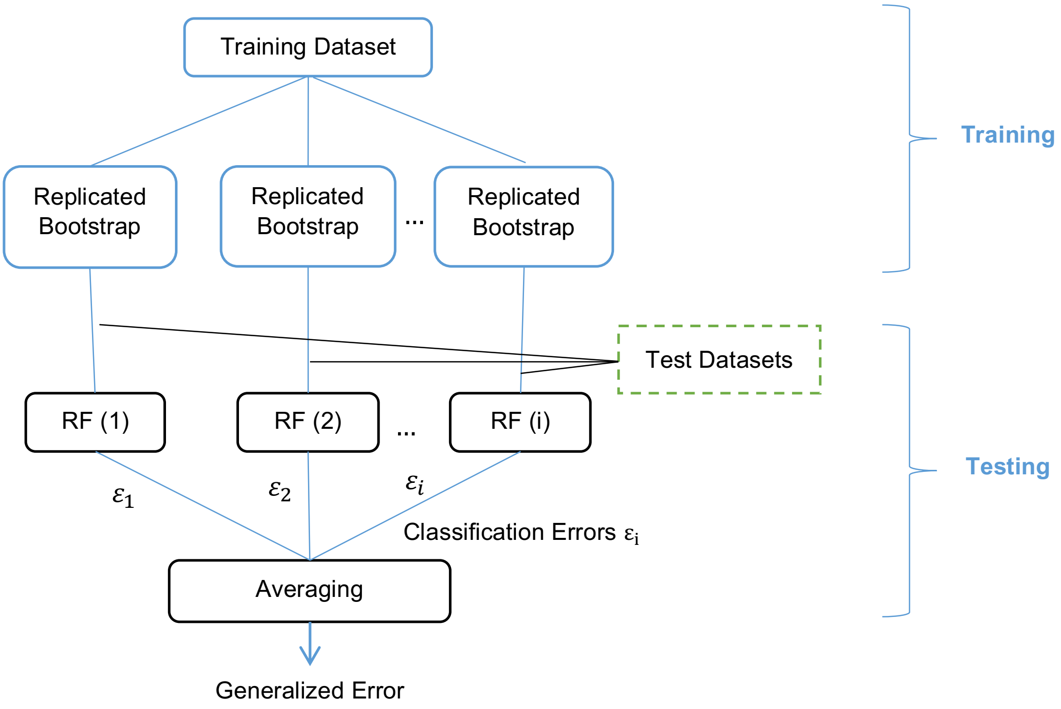


Figure 3. 2 Bootstrap Procedure for Random Forest Classifier

3.2.8 Results

To prepare the data for analyses, as suggested in prior research I examined the quality of eye movement recording and removed the data sets for those participants that had less than 80% gaze sample (Kruger et al., 2013). The gaze sample refers to percentage of the times that eyes were correctly detected by the eye tracker for each participant. For example, 100% means that one or both eyes were detected by the device throughout the recording, whereas 50% means that one eye or both eyes were found for half of the recording duration. While screen based eye tracking experiments typically require users to look at the screen while completing a task, some people may look away to think about a problem or look down, e.g., at the keyboard or mouse.

I manipulated task condition using time pressure, and checked the impact of time limit on performance by using t-test to compare the performance of time pressure vs no time pressure groups. Performance was measured as the total number of correct responses to the math questions. As indicated in Table 3.6, performance of the experimental group is significantly lower than the performance of the controlled group (0.56 ± 0.15 vs. 0.67 ± 0.20, p-value < 0.05). This means that by applying time pressure we could in fact operationalize task condition or different levels of cognitive load in a problem solving task.

Table 3. 6 Performance Results for the Experimental vs Control Groups

	Experimental (time limit)	**Control (no time limit)**
Mean	0.56	0.67
STD	0.15	0.20
p-value	<0.05	

I proposed two hypotheses. The first, (H1), is that eye movement data is likely to carry information about task demand. The second, (H2), is that this information is distinct enough for building an effective eye tracking machine learning system that can predict whether a user experiences higher or lower task demand.

As mentioned earlier I used the random forest algorithm to implement the machine learning model. One of the strengths of random forests is their ability to automatically establish the effectiveness of predictors in the feature set with respect to classification accuracy. Random forest ranks the importance of each metric based on its ability to predict the outcome. This is done by permuting each metric and computing the prediction accuracy of the out-of-bag portion of the data before, and after, the permutation (Breiman, 2001). The random forest output is displayed in Figure 3.3, highlighting the metrics ordered by variable importance. The variable importance is expressed using the Gini index, a measure of node purity or homogeneity of nodes in sub-trees (Hastie et al., 2009). Every time a particular variable is used for splitting a node in the tree, the Gini index for the child node is calculated and compared to the Gini index of the original (parent) node. Variables and cut locations that result in nodes with higher purity have a lower Gini index. Accordingly, this implies that these nodes and cut locations are better at predicting the desired response (James et al., 2014).

The next step was to select the features that were sufficiently discriminative for the classification task. I carried out a forward stepwise feature selection (Hastie et al., 2009), systematically investigating the cognitive load prediction accuracy of the random forest by iteratively adding features based upon variable importance. This process resulted in a minimized error after adding the first ten features; additional features provided only marginal increases in the performance of prediction of cognitive effort. Accordingly, to avoid overfitting I selected only the first ten out of thirty features (Hastie et al., 2009). These ten features are listed based on their order of importance in Table 3.7.

As apparent in Table 3.7 and Figure 3.3, half of the top ten factors predicting task demand are related to pupil data: Avg. (PD during Saccade / PD during Fixation), STD (PDV during saccade/ PDV during Fixation), STD of PDV during Fixation, STD of PD during Fixation, STD of (PD during Saccade / PD during Fixation). These results support research linking pupil data and cognitive effort (Kahneman & Beatty, 1966; Beatty, 1982; Piquado et al., 2010; Buettner et al., 2015; Fehrenbacher & Djamasbi, 2017).

The ratio of pupil dilation and variation during saccades and fixations reflect the distribution of cognitive effort during information search and information processing. The distribution of effort between search and information processing, as suggested by our results, may provide valuable information about user experience of task demand.

Thirty percent of the top predictive factors were related to saccade parameters (STD of Saccade Duration, STD of Saccade Amplitude, and Normalized Saccade Duration) and twenty percent to blink patterns (STD of Blink Duration, Avg. Blink Duration). These results suggest that saccade and blink eye movements have a major influence in effective classification of the eye movement data based on task demand in a problem solving task. Hence, these results provide further evidence in support of literature indicating saccades and blinks are associated with cognitive effort.

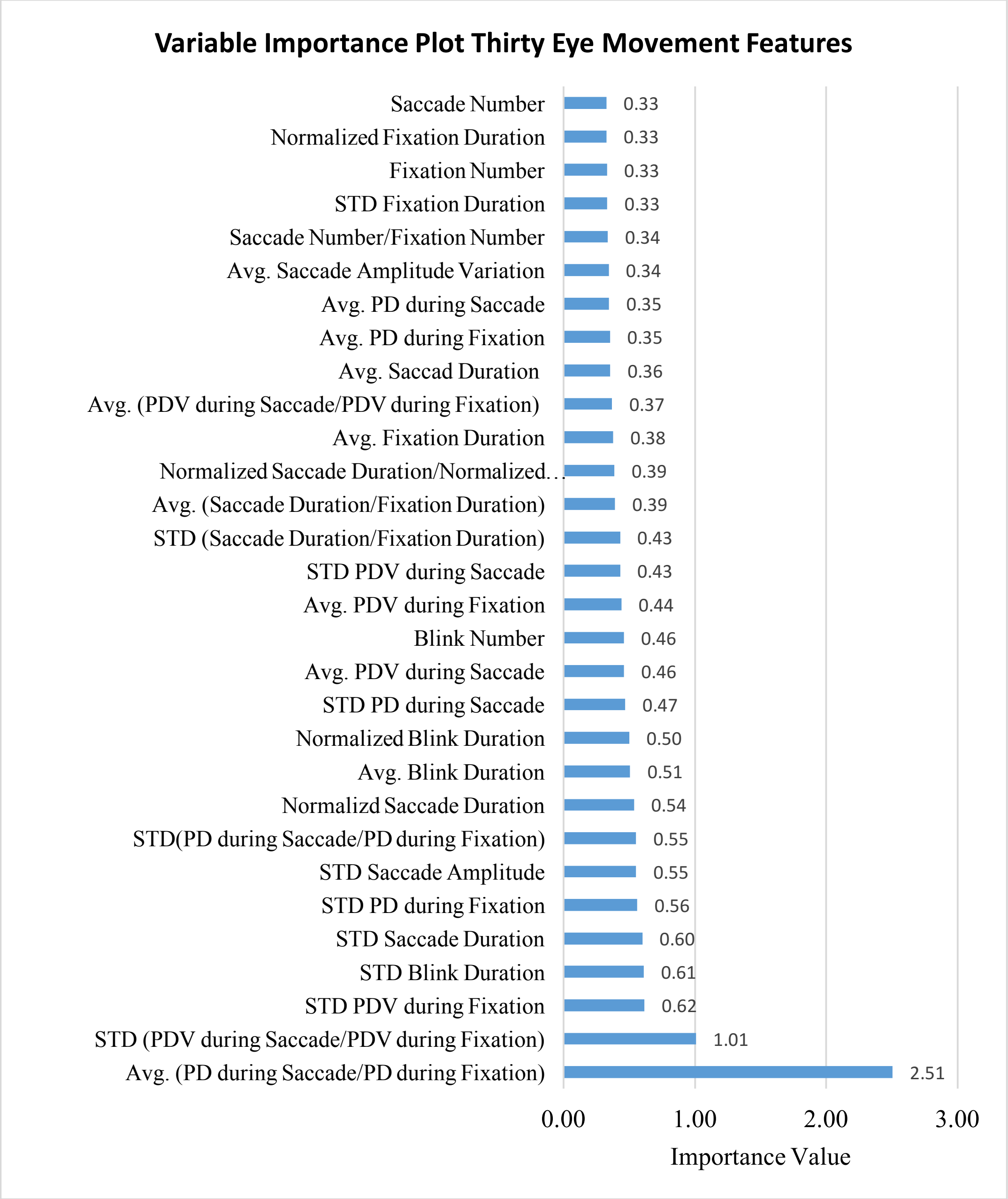


Figure 3. 3 Variable Importance Plot for 30 Eye Movement Features

Table 3. 7 List of features selected by *Variable Importance*

Avg.(PD during Saccade / PD during Fixation)	*STD.(PDV during Saccade / PDV during Fixation)*
STD. PDV during Fixation	*STD. Blink Duration*
STD. Saccade Duration	*STD. PD during Fixation*

STD. Saccade Amplitude	*STD.(PD during Saccade / PD during Fixation)*
Normalized Saccade Duration	*Avg. Blink Duration*

Interestingly enough, the results did not indicate fixation parameters, such as fixation duration and count, to be major contributors to classifying cognitive load. This contrasts with previous research that shows a positive link between fixation duration and cognitive effort – the very nature of viewing a stimulus requires effort in keeping the gaze steady for the information to be visually processed (Djamasbi, 2014). While fixation serves as a reliable and direct indicator of attention and thus information processing, these results indicate that more effective in classifying task condition are the saccade and pupil dilation.

Perhaps most interesting among these results is that pupil dilation ratio values involving saccades and fixations play a major role in classifying higher/lower task demand (Table 3.7). In particular, the variable importance for Avg. (PD during Saccade / PD during Fixation) ratio was noticeably larger than all other metrics. The importance of the average PD ratios during saccades and fixations were more than twice as large as the standard deviation of PDV ratios during saccades and fixations and over four times as large as the rest of the factors. These results both support extant literature summarized in Table 2.1, and also extend previous findings by showing that only pupil, saccade, and blink related data were major predictors in classifying task demand. Further, average PD ratios during saccades and fixations appear to be far more important than the rest of the feature set.

The random forest algorithm can be used to develop different sets of forests that have varying numbers of trees. To find the number of trees that correspond to a stable classifier, I constructed random forests with the number of tree values in the range [1,100]. Figure 3.4 shows the classification performance for 200 replications of bootstrapping, and as a function of the number of trees. The optimal number of trees for our classifier is determined via a standard technique having to do with individual tree error rates, namely, the out of bag error rates (Hastie et al., 2009). When the error rates stabilize and reach a minimum value, the corresponding number of trees constitute the optimal number of trees. From Figure 3.4, the accuracy rate initially increases as the number of trees increase; however, once the number of trees reaches approximately fifteen, the accuracy of the model stabilizes and corresponds to an eye movement classifier with 69.6% accuracy. These results show that the proposed model can predict task demand not only reliably but also quickly. These results support H1 because they show that eye movement data, and in particular the ratios of pupil dilation, and pupil variation, during fixation and saccades, carries information about task demand. They also support H2 because they show that the developed eye tracking classifier can predict user task condition with approximately 70% accuracy.

Of course, one might wonder how such results could be improved. The stability of the results after applying fifteen trees indicates that additional computational effort will likely not improve the results beyond those already achieved for our fixed model and fixed data set. As far as the model is concerned, one could imagine the application of a more sophisticated or customized model giving superior results. On the other hand, overfitting is always a concern, and random forests were intentionally selected for their accuracy and broad applicability. As far as the data is concerned, additional and more detailed measurements would likely increase performance. It is precisely the goal to pursue such improved data generation in my future work.

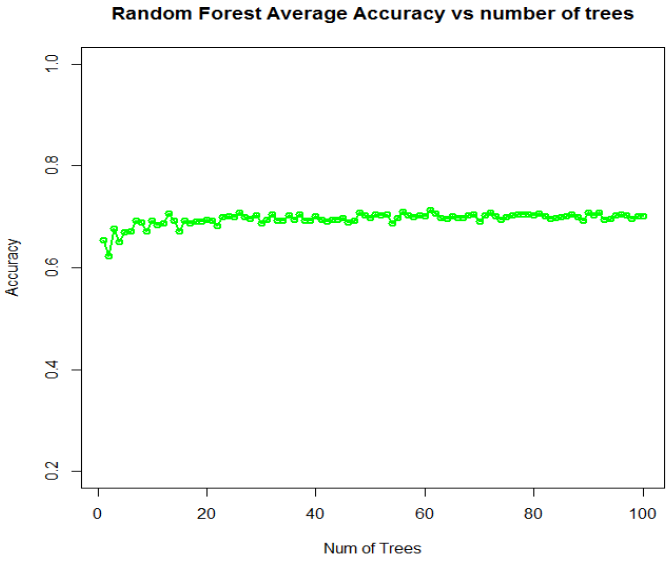


Figure 3. 4 Random Forest Accuracy vs. Number of Trees: Comparison of Random Forest Classifiers Performance by Increasing Number of Trees

To further describe the performance of classification I calculated the confusion matrix and ROC curve for the classification problem. The confusion matrix represents the true positive, true negative, false positive and false negative of the classification task. ROC curve shows a trade-off between true positive rate (sensitivity) and false positive rate (specificity) and is a measure of test accuracy (Zweig & Campbell, 1993). Both confusion matrix and ROC curve for bootstrap#1 with 15 number of trees were calculated and are presented in Table 3.8, and Fig 3.5. According to the confusion matrix, the accuracy of bootstrap # 1 in prediction is 75%, which is calculated as total number of true positive and true negative divided by total number of test samples (20), in Table 3.8.

Table 3. 8 Confusion Matrix of Bootstrap#1 with 15 trees

N=20	**Predicted NO**	**Predicted YES**
Actual NO	**7**	3
Actual YES	2	**8**

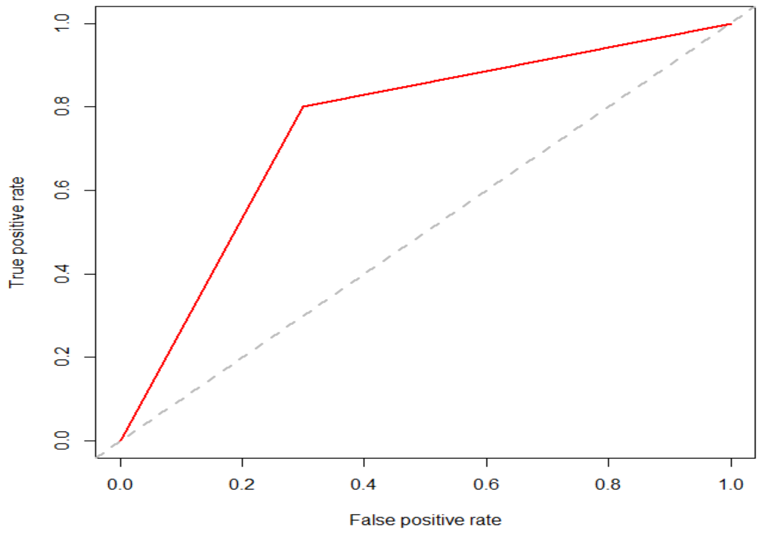


Figure 3. 5 ROC Curve for Bootstrap # 1 and 15 Number of Trees

3.2.9 Additional Analyses

Besides Random Forest, I also investigated the performance of Support Vector Machine linear and nonlinear classifiers on the aforementioned feature set. As shown in Table 3.9, the linear or nonlinear SVM classifiers cannot outperform the proposed RF model. Therefore, these results suggest that RF as compared to SVM is a more suitable method for classification of eye-movement data.

Table 3. 9 Support Vector Machine Classification Performance

SVM	Linear SVM	Nonlinear SVM with polynomial degree of 2	Nonlinear SVM with polynomial degree of 3	Nonlinear SVM with radial basis kernel
Accuracy	56%	43%	41%	48%

As mentioned earlier half of the features among selected features are related to pupil data (see Table 3.7). Here to further investigate each category of features based on eye-movement metrics (e.g., pupil dilation, blinks, fixation, and saccade), I created 6 different categories (Table 3.10). Therefore, 6 different RF models were trained with these different feature sets to investigate the classification performance accordingly. Table 3.10 presents the performance results. Interestingly the highest accuracy (79%) was achieved from PD/PDV related features (category #5 in Table 3.10). Second column of the table show the features listed based on their importance order according to RF Variable Importance values. It is important to note that similar to when all the 30 features were used, the most effective features in the classification is *Avg. (PD during Saccade / PD during Fixation).* These results show that ratio of PD and PDV during fixation and saccades are the most important metrics in the problem solving classification task.

Table 3. 10 RF Classification Performance Using Different Categories of Eye Features

Feature Categories	**Features**	**RF Performance**
1. Fixation Features	1. Avg. Fixation Duration, 2. Normalized Fixation Duration, 3. Normalized Fixation Number, 4. STD of Fixation Duration	40%
2. Saccade Features	1. STD of Saccade Duration, 2. Normalized Saccade Duration, 3. STD of Saccade Amplitude, 4. Avg. Saccade Duration, 5. Avg. Saccade Amplitude, 6. Normalized Saccade Number	51%

3. All PD and PDV Features	1. Avg. (PD-Saccade/ PD-Fixation), 2. STD (PDV-Saccade/PDV-Fixation), 3. STD PDV-Fixation, 4. STD (PD-Saccade/PD-Fixation), 5. Avg. PDV-Fixation, 6. STD PD-Fixation, 7. Avg. PDV-Saccade, 8. Avg. (PDV-Saccade/PDV-Fixation), 9. STD PD-Saccade, 10. STD PDV-Saccade, 11. Avg. PD-Saccade, 12. Avg. PD-Fixation	69.54%
4. Ratio of Saccade Features to Fixation Features	1. STD (Saccade Duration/Fixation Duration), 2. Avg. (Saccade Duration/Fixation Duration), 3. Normalized Saccade Duration/Normalized Fixation Duration, 4. Normalized Saccade Number/Normalized Fixation Number	43%
5. PD and PDV (Only Ratios)	1. Avg. (PD-Saccade/ PD-Fixation), 2. STD (PDV-Saccade/PDV-Fixation), 3.STD (PD-Saccade/PD-Fixation), 4. Avg. (PDV-Saccade/PDV-Fixation)	**79%**
6. Blink Features	1. STD Blink Duration, 2. Normalized Blink Duration, 3. Avg. Blink Duration, 4. Blink Number	52%

3.2.10 Discussion

The two objectives of this study were to examine whether eye movement data can predict cognitive load, and if so can such data be used to develop an advanced system that can predict task demand automatically and reliably. To do so, I conducted a laboratory experiment to collect eye moment data. I selected a set of eye movement metrics, or features, for developing a predictive model for task demand. I used the random forest framework to develop a machine learning system that can predict task demand based on eye movement data.

The results of this experiment supported my hypotheses, demonstrating that eye movements can predict task demand during problem solving and that the random forest classifier could reliably learn from the provided data to predict unseen data. Interestingly, these results showed that pupil data was the *most important contributor* in the predictive model. Many studies have shown that pupil data is a reliable predictor of cognitive load (e.g., see Table 2.1). A novel contribution of this study is that it not only supports this previous finding by showing that half of our top ten predictive factors were related to pupillometry, but also refines previous findings by showing that pupil data was the most prominent predictive factor among the set of thirty eye movement features. The ratio of pupil dilation in saccades and fixations were far more important than other features in predicting cognitive load, perhaps even more important than the absolute pupil dilation reported in previous studies.

It is well established that visual information is processed during fixation, as opposed to saccade. Upon focusing on an object, the eye can only see a small portion vividly and colorfully, namely a limited area around the fixation center. Visual acuity drastically degrades with an increasing distance from the center of the fixation. To compensate for this limitation, saccades are used to rapidly collect high quality visual information. Because saccades change the center of our attention, they represent information search. Since pupil dilation is linked to cognitive activity, pupil dilation during saccade suggests cognitive activity related to information search and pupil dilation during fixation indicates cognitive activity related to information processing. Thus, the results suggest that the ratio of cognitive activity during information search and information processing can provide invaluable insight for classifying task demand.

Another key insight of this study is that, among the top ten discriminating features selected by the machine learning model, *none* are related solely to fixation (see Table 3.7 and Figure 3.3). In this study, metrics related to saccades and blinks were more important than metrics related to fixations. In particular, saccade duration and amplitude were among the top ten predictive factors of task demand. Because saccades indicate effort in locating relevant information, these results suggest how long people took to locate a fixation and how far their eyes had to travel to locate that information provided more insight about task demand than data about their fixation.

Similarly, the results demonstrate that average blink duration and variation were more effective than fixation-related information in predicting (classifying) task condition. Blink duration has been associated with task complexity (Ahlstrom & Friedman-Berg, 2006; Andrzejewska & Stolińska, 2016; Ikehara et al., 2013). This is substantiated in the results. Average blink duration, and variation in blink duration, are likely indicating adjustment to task load, which according to Adaptive Decision Making theory is what people do when making complex decisions (Payne et al., 1993).

Adaptive decision making theory asserts that task conditions can strongly influence cognitive effort (Payne et al., 1993). The results of this study show that this argument can also be observed at a physiological level through pupil data. This in turn supports a recent exploratory study (Fehrenbacher & Djamasbi, 2017), suggesting that Adaptive Decision Making theory can provide a suitable framework for explaining the relationship between cognitive load and pupillometry during problem solving and decision making. As such, the results extend this previous study and provide a rationale and theoretical direction to use eye movement data to develop automatic predictive models. Because the level of effort expenditure affects user strategy in problem solving (Payne et al., 1993), detecting effort unobtrusively and continuously can help designers to have a more complete picture of user experience. Because systems that require too much effort are less likely to be adopted and less likely to be used effectively (Gregor & Benbasat, 1999; Todd & Benbasat, 1994), this study suggests a new avenue for designers to develop more supportive systems.

Advances in technology make it increasingly possible to embed eye-tracking technology in computing devices at affordable prices. The resulting data holds a wealth of information to improve the understanding of user behavior and decision making. The advent of robust machine learning approaches provides an attractive opportunity to capitalize on this information. Hence, designing machine learning predictive models using eye tracking is likely to continue as a productive line of research and development. Using eye movements to detect reactions to task demands is an important first step in designing information systems that can more effectively respond to user needs. Such adaptive tools are likely to be particularly effective in solving complex problems for novice users.

4 Study Two: Assessing Cognitive Load using Eye-Movements during a Reading Comprehension Task

Cognitive load is a major factor affecting user experience. Therefore, to enhance user experience, a better understanding of cognitive load is needed, which helps designing interfaces that can better accommodate user's need according to the level of cognitive load. Research suggests that pupillometry may serve as an excellent unobtrusive measure to study user information processing behavior when under high cognitive load (section 2.2.2). While eye tracking is gaining popularity in IS-HCI research, pupillometry is relatively less explored in IS-HCI eye tracking studies.

The purpose of this study was two-fold. One was to examine the relationship between cognitive load and pupillary responses for a task that required people to either read a text passage from an actual website or read the simplified version of the same text passage (Part I). The simplified text passage was constructed in a way to assure reduced cognitive load, that is, to facilitate communication of textual information in a way that it can be read and understood easily and quickly (see Figure 4.1).

The second goal of the study was to examine the relationship between cognitive load and pupillary responses for a decision making task, when the same participants were asked to answer two questions related to the passages, one literal and one inferential (Part II).

4.1 Part I: Text Simplification and Pupillometry Analyses during Reading

The following sections provide details about the first part of the study. The results of this study is published in the *Proceedings of the 11th HCI International Conference*, in 2017, Vancouver, Canada (Shojaeizadeh et al., 2017).

4.1.1 Introduction

Examining eye movement data is gaining popularity in IS research. As mentioned earlier, research suggests pupillometry may serve as an excellent unobtrusive measure to study user information processing behavior (see Section 2.2 for a review of related literature). Pupil dilation can be measured continuously during processing of a task. This would enable pupillary data to be a potentially robust measure of cognitive load. Further, pupil dilation can be measured continuously during processing of a task (Beatty, 1982; Iqbal et al., 2004), therefore, it could be used as a robust measure for cognitive load during a reading task.

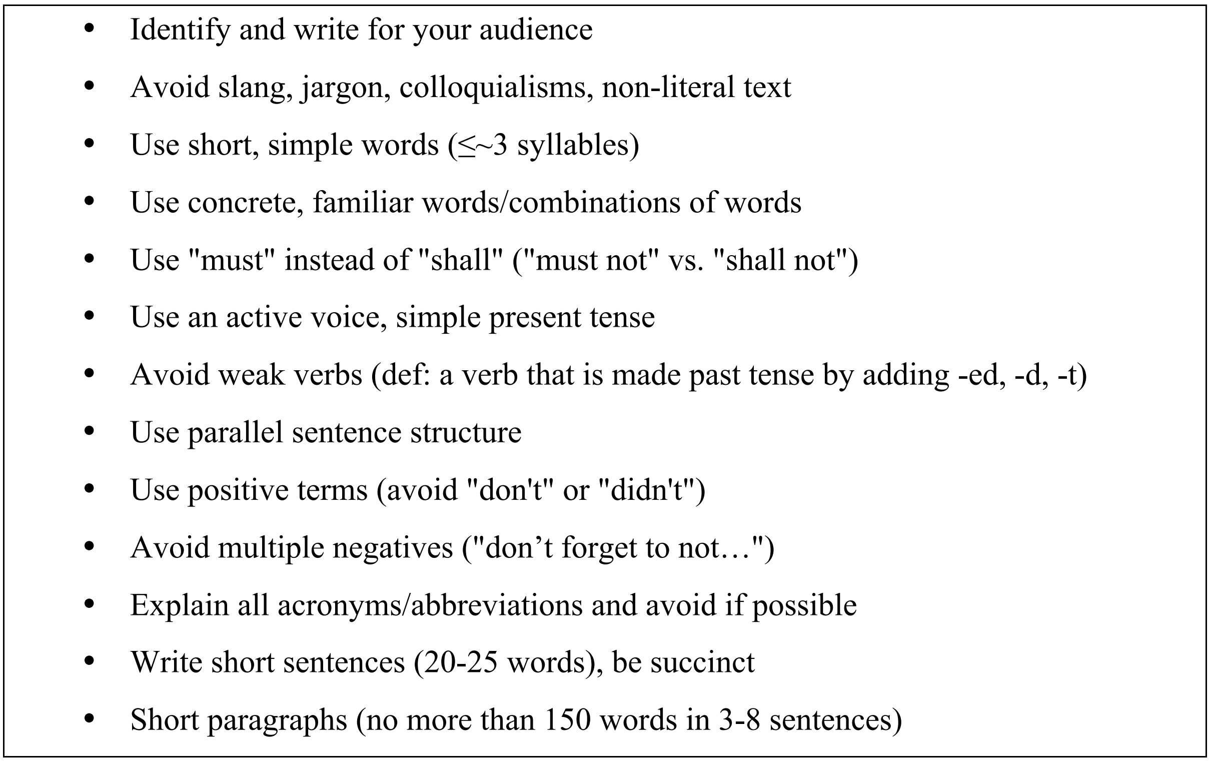
In a previous study Djamasbi et al. (2016-b) showed that by applying a set of plain language standards (PLS) obtained from (Djamasbi et al., 2016-a) to passages obtained from internet they can be simplified so that there is less cognitive demand on the readers. Simplification also resulted in increased user's performance in answering questions related to passages. Additionally, I showed that participants of the simplified version of the passage had shorter average fixations and exhibited a more efficient visual search behavior as compared to those who read the original version of the passage. In this study, I extended this research by investigating time series analysis of eye-movement (pupil dilation) as a proxy for measuring cognitive load during reading these passages.

To this end, I conducted an exploratory analysis to understand how text simplification affected pupil dilation over time and whether or not this effect was consistent over different time intervals during reading. Inspired from my previous research findings (in Section 3.1), I separated PD data during fixation from PD data during saccade and investigated the effect of text simplification on these two variables separately over time. Time series analysis was conducted on the same eye tracking data set reported in the previous study that showed text simplification was effective in reducing cognitive load (that is, it improved performance significantly) (Djamasbi et al., 2016).

The results show that text simplification had a significant impact on pupil dilation and that it affected pupil dilation differently at distinctive reading intervals. Additionally, the results show that examining pupil dilation during fixations and saccades separately can provide new insights for understating cognitive load.

4.1.2 Eye Tracking Experiment

A comprehensive set of plain language standards (Djamasbi et al., 2016) were used to convert an original text passage about sports (18th grade reading level) to a simpler version (10th grade reading level). Please see Figure 4.1 for a list of plain language rules used to simplify the passage. Each participant was randomly assigned to one of the two versions of the text passage, which was displayed on a computer screen. Participants were recruited from a pool of college students. Out of the 54 collected datasets, 26 were from participants assigned to read the original version of the text and the rest were from those assigned to the simplified version of the same passage. After reading the passage, participants were asked to answer two questions about the passage. One of the questions about the passage was literal and the other was inferential. To avoid order effect, the order in which the questions were displayed on the screen was randomized.



- Identify and write for your audience
- Avoid slang, jargon, colloquialisms, non-literal text
- Use short, simple words (≤~3 syllables)
- Use concrete, familiar words/combinations of words
- Use "must" instead of "shall" ("must not" vs. "shall not")
- Use an active voice, simple present tense
- Avoid weak verbs (def: a verb that is made past tense by adding -ed, -d, -t)
- Use parallel sentence structure
- Use positive terms (avoid "don't" or "didn't")
- Avoid multiple negatives ("don’t forget to not…")
- Explain all acronyms/abbreviations and avoid if possible
- Write short sentences (20-25 words), be succinct
- Short paragraphs (no more than 150 words in 3-8 sentences)

- Use transition words in paragraphs (pointing words, echo links, explicit connectives)
- Check/use correct grammar and spelling
- Use "you" and other pronouns to speak to the reader
- Organize document chronologically
- Use lists
- Use tables to make complex material easier to understand
- Do not use ALL CAPS for emphasis
- Do not use underlining for emphasis
- Use bold and italics for emphasis

Figure 4. 1 Plain Language Guidelines

Tobii X300 eye tracking device was used to collect eye movement data. The eye tracker was calibrated for each participant before starting the task. Tobii software version 3.2.3, and I-VT filter with 30°/sec saccadic velocity threshold was used to process raw gaze data into fixations and saccades.

To capture eye-movements for each participants while answering the questions (for Part II of the study) two video segments were created: one capturing user eye movement activity when completing the inferential question and one when completing the literal question.

4.1.3 Time Domain Analysis of Pupil Dilation

Eye-movement data obtained from the eye tracking software were individually saved in .csv format for further processing. Because the task was not time limited, the duration of reading differed among participants, which resulted in dissimilar number of data points for each participant. To facilitate the comparison of time series analysis, studies often equalize the number of data points by designing the task in a way to have pre-specified time windows (Beatty, 1982; Einhäuser et al., 2008). While this approach is useful and relevant for many experiments, I was interested in examining reading behavior in a setting that allowed users to take as much time as they needed to read and understand the text. To compensate for the unequal number of data points in such a setting, I used cubic spline interpolation method (McKinley & Levine, 2002) to construct equal size arrays of PD data for all participants. Interpolation is the estimation of intermediate values between precise data points (Reinsch, 1971). This process created an equal number of pupil data points for each of the participants in each of the two experimental groups (original and simplified conditions). To study the changes in pupil dilation over time, average PD values for all participants were calculated for each cell in the arrays of data. I examined both overall pupil dilation as well as pupil dilation during fixations and saccades. I also looked at changes in PD over three distinct reading periods: beginning, middle, and end. To do so, interpolated data points were divided into three equal intervals.

4.1.4 Results of Pupillary Analyses

The comparison between PD when reading original and simplified passages is indicated in Figure 4.2. Pupil dilation was not separated during saccade and fixation in this plot. Figure 4.2 displays the trend of overall PD. As shown in Figure 4.2 PD trend is similar between reading the original and simplified version of the text except for the beginning and the end part of the graph. I used a t-test to see whether the overall averages for these two trends were different. The results of the two-sample t-test for the means of the overall PD during reading the original or simplified passages showed that there were no significant differences between the two trends (t-stat = 0.93, p = 0.35). In other words, no significant differences were detected in pupil dilation during reading original vs. simplified versions of the same text.

Next, I refined the above analysis by investigating differences in the means of PD during three equally sized different time intervals (beginning, middle and end). A two-way factorial ANOVA was conducted to compare the effects of two independent variables: (1) text simplification, and (2) time interval. Text simplification included 2 levels (1. Original and 2. Simplified) and time interval included 3 levels of equal size (1. Beginning, 2. Middle and 3. End). The results, shown in Table 4.2, indicated that text simplification did not have a main effect on pupil dilation (F(1,938) = 0.88, p = 0.35). The results, however, show that time interval did have a main effect on PD (F(2,938) =17.44, p < 0.001). The interaction effect was not significant (F(2,944) = 2.44, p = 0.09). By comparing the pairwise interactions between time intervals and text simplification, as shown in Table 4.3 and Figure 4.3, it can be seen that there is a significant difference in PD between simplified and original versions in the last time interval. There are no significant differences between simplified and original versions in the beginning and the middle reading intervals (p-beginning = 0.32, p-middle = 0.63).

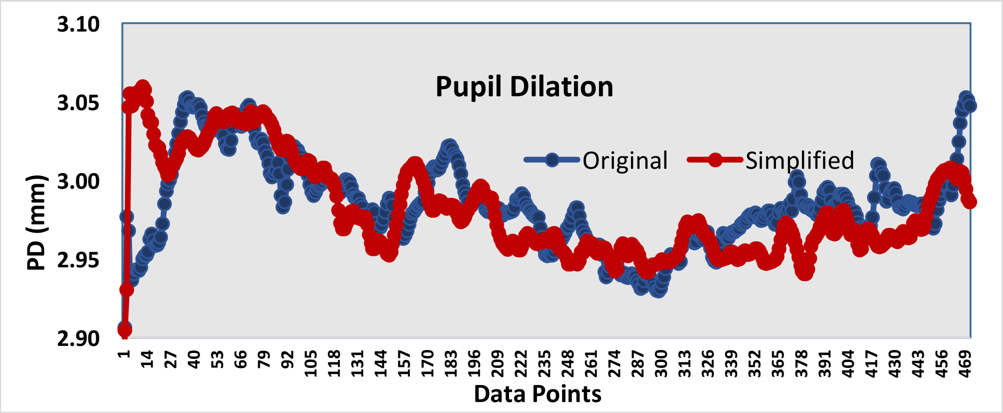


Figure 4. 2 Time series trend of pupil dilation, red: simplified, blue: original passage

Table 4. 1 Descriptive Statistics and T-test Results for PD when Reading Original and Simplified Passages

	Mean	SD	t Stat	df	p-value
Original	2.982	0.0298	0.926	905	0.35
Simplified	2.978	0.036			

Table 4. 2 ANOVA Results Comparing the Means of PD during Different Intervals and among Original and Simplified Passages

	F	P-value
Task Demand	0.88	0.35
Intervals	17.44	= 0.001
Interaction	2.44	= 0.09

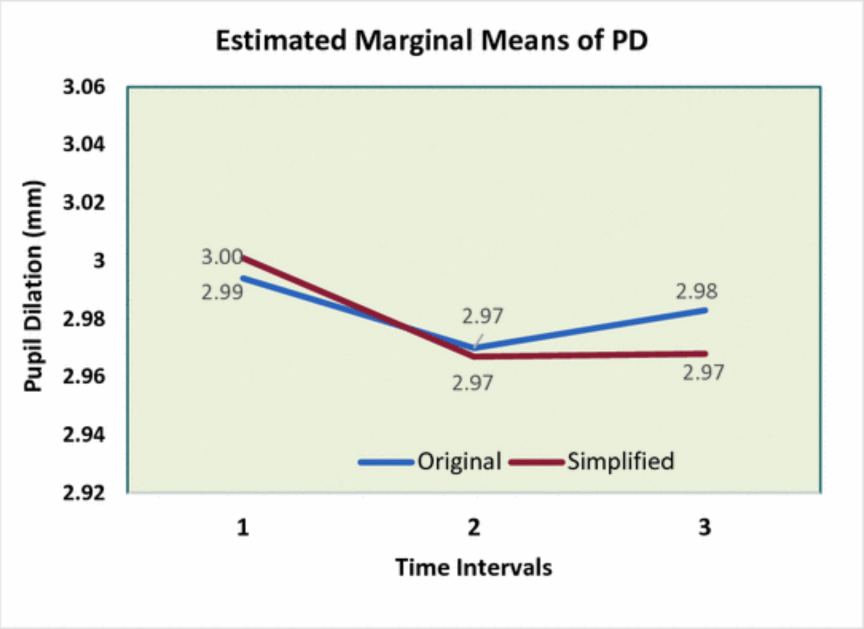


Figure 4. 3 The Main Effect of Text Simplification and Time Intervals Segmentation and Their Interaction Effect on the Dependent Variable PD

Table 4. 3 Pairwise Comparison between Different Time Intervals for Overall PD

Time Intervals	Mean ± SD (PD-Original)	Mean ± SD (PD-Simplified)	P-value
Beginning	2.99 ± 09	3.00 ± 0.11	0.32
Middle	2.97 ± 02	2.97 ± 0.01	0.63
End	2.98 ± 01	2.97 ± 0.01	0.03

Next, similar analyses were performed, to examine PD during fixation (PD-Fixation) and saccades (PD-Saccade) separately. Figures 4.4 and 4.5 show the time series trend of PD-Fixation and PD-Saccade when reading original versus simplified passages. These graphs show more nuanced differences between PD trends in the original vs. simplified conditions.

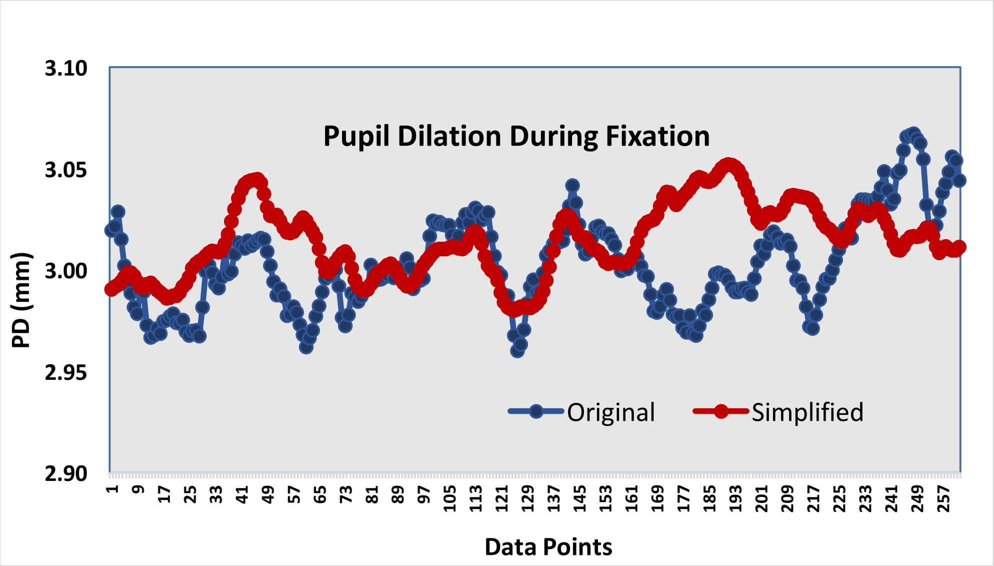


Figure 4. 4 Time Series Plot for Pupil Dilation during Fixation (blue = original, red = simplified)

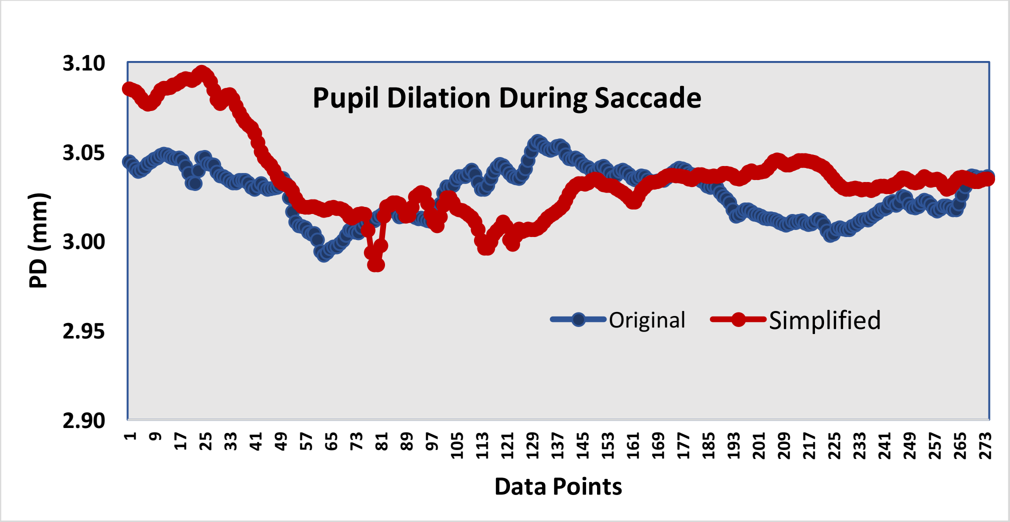


Figure 4. 5 Time Series Plot for Pupil Dilation during Saccade (blue = original, red = simplified)

A two-way ANOVA was performed to investigate the effects of (1) text simplification and (2) fixation/saccade separation on PD. The results of ANOVA in Table 4.4 show that PD is significantly affected by text simplification, for both fixation and saccade measures (F(1,1068) = 68.71, $p < 0.05$) and that PD values are significantly different during saccades and fixations for both task conditions (original vs. simplified) (F(1, 1068) = 331.64, $p < 0.05$).

There is no significant interaction effect between text simplification and fixation/saccade separation as indicated in Table 4 (F (1, 1072) = 1.63, p = 0.20). The graphical representation of this analysis is displayed in Figure 4.6 which shows that no matter what type of passage the participant was reading (original or simplified) the average PD during fixation (blue line) was smaller than average PD during saccade (red line), and this difference remains almost the same either when reading the original passage or the simplified passage. Both PD-Fixations and PD-Saccades had larger average values in the simplified text condition. These results show that separating PD-Fixation and PD-Saccade provides additional information that is useful when performing time-series analysis of pupil dilation.

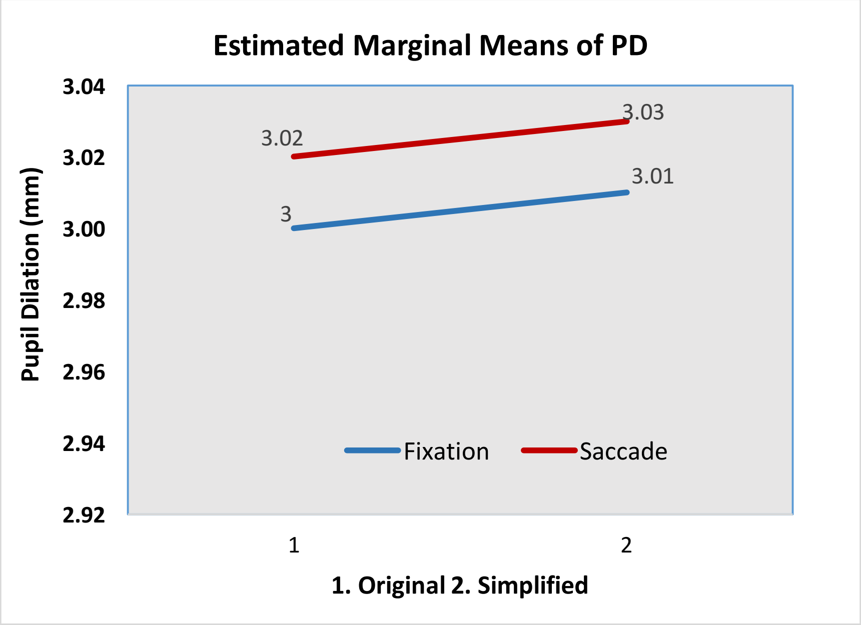


Figure 4. 6 The main effect of text simplification and fixation/saccade separation and their interaction effect on the dependent variable PD

Table 4. 4 ANOVA Results Comparing the Means of PD between Fixation and Saccade during Original and Simplified Passages

	F	P-value
Task Demand	68.71	< 0.001
Pupillary Segmentation	331.64	< 0.001
Interaction	1.63	0.20

Having separated PD-Fixation from PD-Saccade, next I investigated this data using the three reading time intervals. A two-way ANOVA test was used for PD-Fixation and PD-Saccade to compare the overall effects of two independent variables (1) text simplification, and (2) time interval separation.

Table 4.5 displays the overall results of the ANOVA tests. The results show that PD values are significantly different both during fixations and saccades when people read original vs. simplified text passages ($F(1,518) = 54.23$, $p < 0.05$ for PD-fixation and $F(1, 542) = 35.885$, $p < 0.05$ for PD-saccade). The results show that PD values during fixations are also significantly different over the three time intervals ($F(1,518) = 65.17$, $p < 0.05$ for PD-fixation). Additionally, the differences in PD-fixations between the original and simplified conditions are significantly different in the three time intervals ($F(2,524) = 10.13$ and $p < 0.05$). The same is true for PD values during saccades ($F(1,542) = 17.05$, $p < 0.05$). The results show significant interaction effect between text simplification conditions and time intervals ($F (2,548) = 75.59$, $p < 0.05$).

Table 4. 5 ANOVA Results Comparing the Means of PD-Fixation and PD-Saccade within Different Time Intervals

PD- Fixation	F	P-value
Task Demand	54.23	P < 0.001
Intervals	65.17	p < 0.001
Interaction	10.13	p < 0.001

PD-Saccade	F	P-value
Task Demand	35.88	p < 0.001
Intervals	17.05	p < 0.001
Interaction	75.59	p < 0.001

These differences are further shown in Table 4.6, which displays the pairwise comparison between PD saccade/fixation for original and simplified passages among different time intervals. In other words, PD-fixation is significantly different between original and simplified passages in the beginning ($p < 0.05$), and end ($p < 0.05$) of the reading duration but not in the middle of the reading duration ($p > 0.05$). Identically, PD-saccade is also significantly different between original and simplified passages in the beginning ($p < 0.05$) and the end ($p < 0.05$). However, the difference in PD-saccade between original and simplified passages is also significant in the middle of reading ($p < 0.05$). Figures 4.7and 4.8 display graphical interpretations of these results. As it can be seen in these figures, average PD-Fixation is larger in the simplified group compared to the original group at the beginning and the end intervals. In the middle interval Average PD-Fixation values are the same in both groups. While we observe a similar trend for PD-Saccade in the beginning and end intervals, in the middle interval average PD-Saccade for the original passage is significantly larger than average PD-Saccade for the simplified passage.

Table 4.6 Pairwise Comparison Between Different Time Intervals for PD-Fixation and PD-Saccade

PD-Fixation

Time Intervals	Mean ± SD (Original)	Mean ± SD (Simplified)	P-value
Beginning	2.99 ± 0.01	3.01 ± 0.02	p < 0.001
Middle	3.00 ± 0.02	3.00 ± 0.01	p =0.53
End	3.01 ± 0.02	3.03 ± 0.01	p < 0.001

PD-Saccade

Time Intervals	Mean ± SD (Original)	Mean ± SD (Simplified)	P-value
Beginning	3.02 ± 0.2	3.05 ± 0.03	p < 0.001
Middle	3.04 ± 0.01	3.02 ±0.01	p < 0.001
End	3.02 ± 0.01	3.04 ± 0.00	p < 0.001

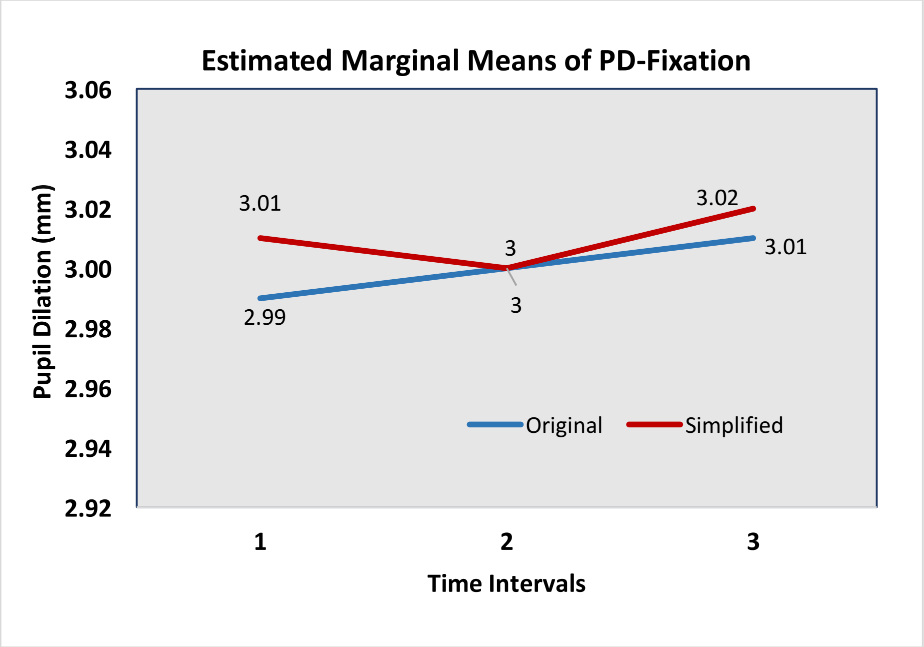


Figure 4. 7 The Main Effect of Text Simplification and Time Intervals and their Interaction Effect on the Dependent Variable PD during Fixation

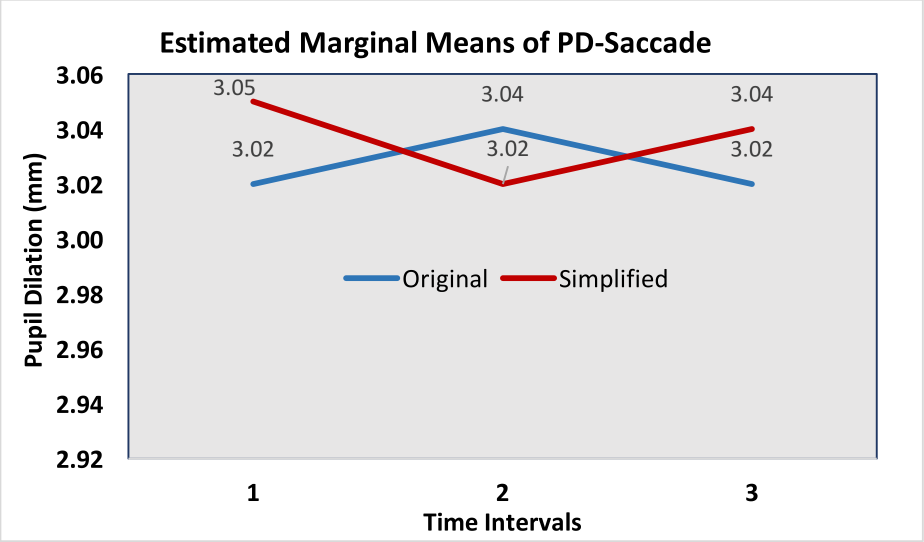


Figure 4. 8 The Main Effect of Text Simplification and Time Intervals and their Interaction Effect on the Dependent Variable PD during Saccade

These findings are consistent with previous literature (Beatty, 1982; Chen et al., 2011; Iqbal et al., 2004; Klingner, 2010) that identify pupil dilation as a reliable measure of cognitive load in cognitive tasks. In addition; these results indicate that separating pupil dilation during fixation and saccades can provide more nuanced information that is not available when considering only the overall PD. Furthermore, these results open new research questions in the field of pupillometry related to HCI research, which will be discussed in the next section of this paper.

4.1.5 Discussion and Conclusion

Time series analysis of eye-tracking data is important because it provides a continuous measure of eye-movement data, which allows us to examine moment by moment analysis of eye-movement data. In this study, I conducted time-series analysis of pupil dilation, which is considered a reliable measure of cognitive effort.

I investigated whether reducing cognitive load of readers by simplifying text passages can affect their pupil dilation during reading and whether this effect remained steady over different time intervals of reading. The simplified text passage used in this study was developed using a set of plain language rules. The original passage, which was an actual news passage about sports, was simplified from 18th grade reading level to 10th grade reading level through systematic application of the plain language rules described in Djamasbi et al., (2016-a).

The results of t-test comparing overall PD values between the original and simplified groups showed that text simplification did not significantly affect pupil dilation over time. However, when dividing the data points into three equally size intervals, the results showed that PD values were significantly different between the original and simplified groups in the last part of reading. Because pupil dilation is associated with increased cognitive load, the results displayed in Figure 4.3 suggest that participants were experiencing more cognitive load at the beginning of the task (compared to the two other time intervals) when they were familiarizing themselves with the text. The results also show that cognitive load was similar for the two text conditions (original vs. simplified) at the beginning and middle time intervals but it was significantly lower for people in the simplified text condition at the end interval. These results provide evidence that examining PD in various time intervals can provide additional information for understanding cognitive load.

Next, I examined PD for fixations and saccades separately. The results showed that the impact of text simplification on PD was significant when separating PD-Fixation from PD-Saccade. These findings support the argument that examining PD during fixations and saccades separately is useful in refining the explanatory power of pupillometry. These results are consistent with the argument that the observed differences are due to differences in the nature of fixation and saccadic eye-movements (Shojaeizadeh et al., 2015). Further, the results showed that pupil dilation was slightly larger during saccades as compared to pupil dilation during fixations (Figure 4.6). The results also show that PD measured during both fixations and saccades, was larger in simplified version of the text. Note that PD during fixations refers to visual information processing. Because during saccades we cannot process visual information (our eyes move too fast to be able to take foveal snapshots), PD during saccades may indicate cognitive processing beyond what is typically associated with attention measured as foveal processing of visual information. Given this interpretation, the results in Figure 4.6 suggest higher cognitive activity during saccades compared to fixations for both participants reading the original or the simplified versions of the text. It also suggests higher cognitive activity in the group that was reading simplified text. Given that the performance for the same set of data indicated that people provided significantly more accurate answers to questions about the text in the simplified group (Djamasbi et al., 2016), higher cognitive activity in the simplified group in this case may indicate higher level of engagement with the task.

Next, I examined PD during fixations and saccades over the three reading intervals: beginning, middle, and end. The results, displayed in Figure 4.7 and Figure 4.8, reveal different effects. During fixations, PD increases consistently over the three time periods when people read the original version of the text. However, when people read the simplified version, PD in the middle of the reading is significantly smaller than the two other intervals. During the saccades, PD values in the middle interval are higher than the two other intervals for the original version of the text while they are lower than the two other intervals for the simplified version of the text. These results indicate the presence of different types of activities in the middle of reading as represented by PD during saccades and fixations. While future experiments are needed to fully explain these differences, these results provide evidence for the usefulness of examining PD during fixations and saccades separately during various intervals.

Overall these findings are consistent with previous literature that have employed pupil dilation as a reliable measure of cognitive load. Additionally, the results indicate that investigating pupil dilation in different time intervals is useful in providing a better understanding of cognitive load in reading and that separating the analysis of pupil dilation during fixations and saccades can provide additional useful information about cognitive load. These results provide a rationale for new research questions in the field of pupillometry related to HCI research. For example, why do PD-Fixation and PD-Saccade show similar behavior at the beginning and end of the reading when comparing reaction to original and simplified passages, but show different behavior in the middle of reading? Is this a consistent behavior even when we test different passages or with a different population of readers?

4.2 Part II: Task Condition and Pupillometry Analyses in a Cognitive Decision Making Task

In the first part of this study the relationship between cognitive load and pupillary responses were studied during a reading task, which included reading either an original version of a passage (higher task demand) or a simplified version of that (lower task demand). The results revealed that overall PD values between the original and simplified groups of text simplification (different task condition) did not significantly affect pupil dilation over time. However, when dividing the data points into three equally size intervals, the results showed that PD values were significantly different between the original and simplified groups in the last part of reading. In the second part of this study the objective is to test whether the same relationship between pupillary responses and task demand exist during a decision making process such as answering questions about the passages. Another objective of this study is to determine if text simplification has also affected the way participants make a decision, and whether this effect could be measured from pupillary responses.

The results of this study have been published in the *Proceedings of the AMCIS Conference* 2017 (Shojaeizadeh et al., 2017-b).

The Adaptive decision making theory asserts that task condition affects information processing behavior. According to this theory, I argue that users' pupillary responses will be different under different task conditions.

4.2.1 Introduction

The Adaptive Decision Making theory asserts that task condition affects information processing behavior (Payne et al., 1993). Further, according to this theory people choose an information processing behavior based on the demand placed on their cognitive resources. In addition to the pupil dilation (see section 4.1), there is indication that variability in pupil dilation or rate of change of pupil size may also reveal users' cognitive load due to task demand (Buettner et al., 2015; Chen et al., 2011; Fehrenbacher & Djamasbi, 2017). Therefore, grounded in this theory, and the literature that supports the relation between pupillary data and cognitive load, I argue that pupillary responses are likely to carry information about task condition in a decision making task. Similar to section 4.1, I tested this assertion via an eye tracking laboratory experiment. In the following section I will briefly reintroduce the literature presented in section 2.2.2, that support the hypotheses built to test the research questions in this part of the study.

4.2.2 Hypotheses

Research shows pupillary data can serve as a reliable measure of cognitive effort. For example, Beatty and Kahneman (1966) observed an increase in pupil size as people completed harder tasks. Similarly, Chen et al. (2011) observed a positive relationship between pupil size and task difficulty (i.e., recalling the number of player positions in a basketball game). Klingner et al. (2011) measured pupil dilation during a mental multiplication; and found that easy-multiplication problems triggered the smallest pupil dilations and hard problems the largest. Other recent IS scholars also suggest that pupil dilation is a reliable proxy of cognitive load (Klinger et al. 2008, King 2009, Piquado et al. 2010, Zhan et al. 2016). Some recent IS studies have examined the relationship between cognitive load and pupil dilation variation (PDV). PD is defined as the size of pupil diameter and PDV is defined as rate of change in pupil dilation measured as standard deviation of pupil dilation (Shojaeizadeh et al. 2015; Buettner et al. 2015). For example, Buettner et al. (2015) showed that PDV has a positive relationship with performance, and argued that PDV is an appropriate measure of cognitive load in IS research. A recent study provided evidence that task condition has an impact on pupillary response, and suggested the Adaptive Decision Making theory (Payne et al. 1993) may serve as a suitable theoretical framework for IS decision making eye tracking studies (Fehrenbacher et al., 2017). According to this theory people choose an information processing behavior based on the demand placed on their cognitive resources. Grounded in this theory, I argue that different task demands impact cognitive resources in different ways. This, in turn, is likely to impact how people manage their cognitive loads and thus is likely to impact their pupillary responses. Therefore, I hypothesize:

(H1) Pupil dilation (PD) will be different in different task conditions.

(H2) Pupil dilation variation (PDV) will be different in different task conditions.

4.2.3 Methodology

The methodology (experimental set up, data analyses) in this study is similar to section 4.1.2. In this part of the study eye tracking data was obtained for only the duration of answering questions about the passages, from the same participants of part 4.1.

The task required participants to read a text passage and answer two questions about the passage they just read. Participants were randomly assigned to one of the two text conditions (original or simplified). One of the questions about the passage was literal and the other was inferential. To avoid order effect, the order in which the questions were displayed on the screen was randomized. For each participant two video segments were created: one capturing user eye movement activity when completing the inferential question and one when completing the literal question.

4.2.4 Results

To test the hypothesis, I investigated pupillary responses to 1) task demand (original/simplified) and 2) to question type (inferential/ literal). I performed two mixed model ANOVAs one for PD and one for PDV. The results (Tables 4.7 and 4.8) indicate that PD values for participants in the simplified text condition was significantly different from the PD values for those in the original text condition (F(1,477) =16.92 and p-value <0.05). Additionally; the results show that when answering literal questions compared to when answering inferential questions, PD values were significantly different (F(1,477)=65.71, p-value <0.05). Further, the results show that PD values were impacted significantly by the interaction between the type of text (simplified vs. original) and type of question (inferential vs. literal) (F(1,477) =22.29, p-value <0.05). The pairwise comparisons show that in the simplified text condition PD during responding to the inferential question was significantly different from PD during answering the literal question (PD- inferential=2.992 vs. PD-literal=3.007, F (1,477) =47.72, p-value <0.05). I did not observe differences in the original text condition (PD-inferential=2.994 vs. PD-literal=2.998, F (1,477) =1.18, p-value >0.05) (Table 4.8).

PDV values for people in the simplified text condition were also significantly different from the PDV values for people in the original text condition (F (1,524) =85.17, p-value <0.05). Similarly, PDV values were also significantly different between literal and inferential questions (F (1,524) =11.12, p-value <0.05). The results, however, did not show a significant interaction effect (F (1,524) =0.43, p-value >0.05). In other words, for both types of passages the users read (original or simplified) their PDV was significantly affected by type of the questions they answered (Table 4.8). The results of pairwise comparisons in Table 4 .8 show that PDV in the original text condition was significantly different between inferential and literal questions (PDV- inferential= 0.0060±0.0002 vs. PDV-literal=0.0066 ± 0.0002, F (1,524) =58.80, p-value <0.05). PDV in the simplified text conditions was also significantly different between inferential and literal questions (PD-inferential=0.0041±0.0001 vs. PD-literal=0.0049±0.0002, F (1,524) =31.97, p-value < 0.05).

Table 4. 7 Results of Mixed Model ANOVA Comparing the Means of PD among Literal and Inferential Questions

	F(1,477)	p-value
Task Demand	16.92	< 0.005
Question Type	65.71	< 0.005
Interaction	22.29	< 0.005

Table 4. 8 Descriptive Statistics and Pairwise Comparison between Means of PD in Different Task Conditions

Task Condition	Question type	Mean ± SD	F(1,477)	p-value
Original Text	Inferential	2.994 ±.004	1.18	0.28
	Literal	2.998 ±.004		
Simplified Text	Inferential	2.992 ±.003	47.72	< 0.005
	Literal	3.007 ±.004		

These results together support the hypothesis that different task demands significantly affect PD and PDV during a cognitive decision making task. The results also support the assertion that question type also affects PD and PDV values. The results show an interaction effect between text type and question type on PDV, but not on PD.

Table 4. 9 Results of Mixed Model ANOVA Comparing the Means of PDV among Literal and Inferential Questions

	F (1,524)	p-value
Task Demand	85.17	< 0.005
Question Type	11.12	< 0.005
Interaction	0.43	0.52

Table 4. 10 Descriptive Statistics and Pairwise Comparison between Means of PDV of Different Task Conditions

Task Condition	Question type	Mean ± SD	F(1,524)	p-value
Original Text	Inferential	.0060± .0002	58.80	< 0.005
	Literal	.0066 ± .0002		
Simplified Text	Inferential	.0041 ± .0001	31.97	< 0.005
	Literal	.0049 ± .0002		

4.2.5 Exploratory Analysis

As shown in part I of this study, examining pupil data in various time intervals is useful, and provides further information on the relation between pupillary responses and task condition. In addition, Fehrenbacher & Djamasbi (2017) showed that the way people distribute their effort over the decision time to complete a cognitive task varies within different time intervals. Therefore, the objective of this exploratory study is to test whether we observe the same eye movement pattern or user behavior when it comes to making decisions in answering questions about an original passage or its simplified version.

I divided the interpolated time into three equal size portions of beginning, middle and end, and investigated the differences in the means of PD and PDV during these time intervals. As mentioned earlier in part I, because the task was not timed task duration was different among the participants, which resulted in different PD and PDV data points for each participant. To standardize the number of PD and PDV values we used a cubic spline interpolation (McKinley & Levine, 2002). Therefore, the number of data points in each time interval is equivalent in the both task conditions (literal vs. inferential questions).

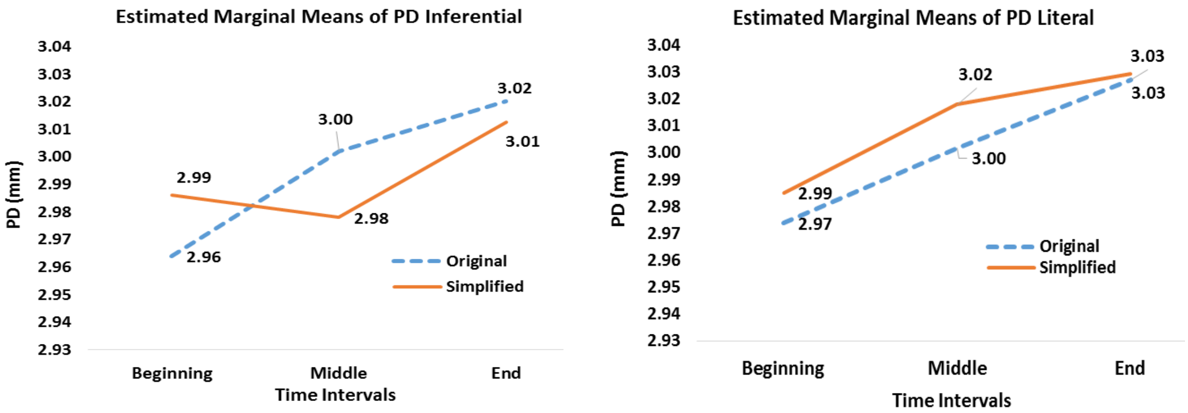
Next, I performed two separate two-way ANOVAs. The first ANOVA tested whether PD during answering the inferential question (PD-inferential) was significantly different between different text conditions during different time intervals (PD-inferential); and the second ANOVA tested the same effects on PD but this time during answering the literal questions (PD-literal).

The results displayed in Table 4.10 show that, when answering inferential and literal questions, PD values for participants in the simplified text condition were not significantly different from PD values for participants in the original text condition during both inferential and literal questions. In other words, the results do not show a main effect for text type on PD-inferential (F(1,950)=0.13, p-value >0.05) as well as on PD-literal (F(1,1042)=3.42, p-value >0.05). The results however show that PD values were significantly different during the three time intervals (F (2,950) = 23.14, p-value <0.05). Furthermore, the results show that PD values were significantly affected by the interaction between the text type condition and time interval divisions (F(2,950)=7.59, p-value <0.05) during the inferential question. Results also show that PD values were not significantly affected by the interaction between the text type condition and time interval divisions during the literal question (F(2,1042)=0.58, p-value >0.05). Figure 4.9-a shows the pairwise comparison between PD of the original text condition and PD of the simplified text condition during the three different time intervals and during inferential question. Results show that PD value during the simplified text condition is significantly different from PD value during the original text condition during the beginning and middle time intervals but not during the end time interval. An upward trend for PD values can be observed in both text conditions suggesting more intense cognitive activity at the end, when users were making decisions. This figure also shows opposite trends, in different text conditions, in the beginning and the middle time intervals.

Table 4. 6 Results of ANOVA for PD within different time intervals during inferential and literal questions

	Inferential		Literal	
	F	p-value	F	p-value
Task Demand	0.13	0.72	3.42	0.07
Intervals	23.14	<0.005	28.30	<0.005
Interaction	7.59	<0.005	0.58	0.56

Figure 4.9-b shows the pairwise comparisons between PD-literal values in the original text condition, and PD-literal values in the simplified text condition during different time intervals. Results show that PD was not significantly different during any of the three intervals between the simplified and the original text conditions ($P_{Beginning}$, $P_{Middle,}$ P_{End} > 0.05).


(a) Inferential Question

(b) Literal Question

Figure 4. 9 Average Values of PD during Three Different Time Intervals (Beginning, Middle, and End) and Two Task Conditions (Original vs. Simplified), and Two Different Question Types (Inferential vs. Literal)

Next, I investigated the differences in the means of PDV during the three time intervals: beginning, middle, and end. As before, we conducted two separate two-way ANOVAs, one for PDV values captured during answering an inferential question (PD-inferential), and one for those collected during answering a literal question (PD-literal).

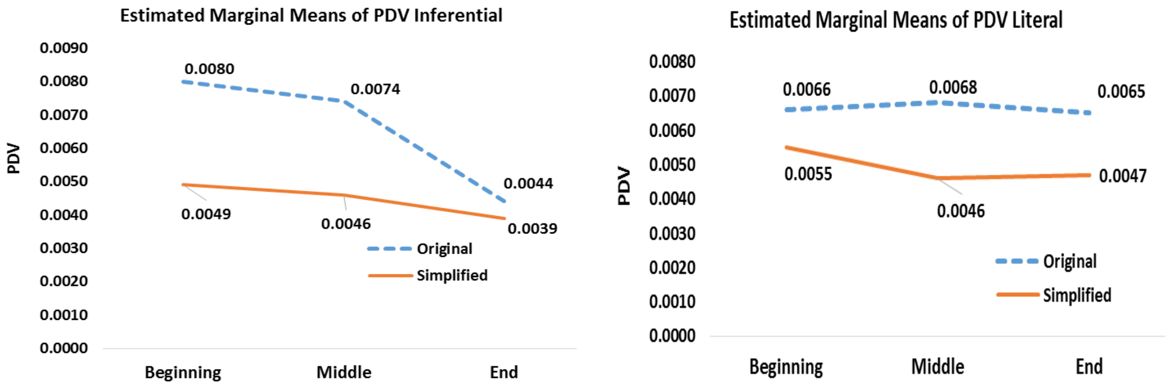
The results of ANOVA and pairwise comparisons are indicated in Tables 4.11 and Figure 4.10 for PDV during inferential and literal questions. As shown by the results, PDV is significantly different between simplified and original task conditions. PDV values are also significanlty different during different time intervals when answering an inferential question. There is also a significant interaction effect between task condition and time intervals during inferential questions.

The results of ANOVA for literal questions were different from those obtained for the inferential questions. The results in Table 4.11 show that PDV values were significantly different between two task conditions, but they were not significantly different during the three time intervals. There was also no significant interaction effect (F(2,1044)=1.20, p-value >0.05) between task condition and time intervals. This suggests the significant effect of task condition on PDV did not depend on the time interval when answering literal questions. Results of the pairwise comparison in Figure 4.10-a indicate that, only during the last time interval.

Table 4. 7 Results of ANOVA for PDV within Different Time Intervals during Inferential and Literal Questions

	Inferential		Literal	
	F	p-value	F	p-value
Task Demand	59.03	<0.005	32.91	<0.005
Intervals	26.25	<0.005	0.83	0.44
Interaction	8.89	<0.005	1.20	0.30

PDV is not significantly different between original and simplified text conditions, but PDV is significantly different during the first and second time intervals (the beginning interval and the middle interval). A downward trend for PDV can be observed in both text conditions during answering inferential questions. The downward trend is more pronounced in the original text condition. At the end, when participants were making decisions, PDV values were quite similar. A downward trend for PDV can be observed in both text conditions during answering literal questions (Figure 4.10-b). There was a slight increase in PDV values in the middle of the original text condition and a slight decrease in PDV in the middle of the simplified text condition.



(a) Inferential Question

(b) Literal Question

Figure 4. 10 Average values of PDV during Three Time Intervals (Beginning, Middle and End), Two Different Task Conditions (Original vs. Simplified), and Two Different Question Types (Inferential vs. Literal).

4.2.6 Conclusion and Discussion

The main objective of this study was to test whether differences in task demand could be detected via pupil data during a reading task. Because pupil dilation and variation have been associated with cognitive activity, I hypothesized that pupillary responses were likely to reflect task demand in this study (simplified vs. original text conditions; and answering inferential vs. literal questions). The results of ANOVA tests showed there was an overall significant difference in PD and PDV between the two text conditions and the two types of questions. While the results showed an interaction effect between text conditions and question type for PD, no interaction effect was observed for PDV. These results support the hypotheses. I also conducted exploratory ANOVA tests to refine the analyses. I divided the task time into three intervals, and investigated PD and PDV values in each interval. Prior research indicated that studying PD and PDV during different time intervals could provide interesting insights for future pupillometry studies (Fehrenbacher & Djamasbi, 2017). Supporting the results of previous research, the results showed larger pupil dilation at the end of the decision period. These results also supported the finding of this prior research, which showed higher PD values were associated with lower PDV values.

These results have important implications. First they show that PD and PDV can carry information about task condition. It is important to note that both the task and task conditions in the previous study were different from the task and task conditions in this study. Hence, the results provide evidence for the robustness of pupillometry in IS research. The results also show that examining pupil data in various time intervals during a decision task can provide valuable insight about cognitive demand in a decision making task. For example, the results showed an upward trend for PD values during the decision time. The upward trend in pupil dilation for both inferential and literal questions at the beginning of decision periods suggests that participants were experiencing increased cognitive load at the beginning of the decision task (especially in the original text condition). The results also showed that during inferential question the difference in PDV values between text condition depended on the time interval of the decision task, they were signficatly different at the beginning and in the middle of the task but not signficantly different at the end. This suggests that participants in the two different text conditions may have used different information processing strategies at the begining and in the middle of the task but used the same strategy at the end. Such findings suggest that PD and PDV can provide continuous measurement of cognitive load in a decision task. Future research is needed to examine these possibilities. Future studies can benefit from these findings which suggest calculating PDV as well as PD within different time intervals is likely to provide a more comprehensive understanding of cognitve load in cognitive tasks. In this study I looked at three different time intervals, however, future studies can explore whether breaking the total time into smaller intervals can improve understanding of user cognitve load.

As in any laboratory experiment, the results of this study are also limited to the setting and the task used. I used a single text passage. More text passages with varying level of complexity are needed to replicate the results. The participants in the study were college students. Different participant groups with various demographics can help to test whether the results extend to other populations.

5 Study Three: Eye Movements and Reading Behaviour of Younger and Older Users

In this section I conduct exploratory analyses to investigate the differences between generation Y and baby boomers in reading textual information online. The results of this research have been accepted for publication in the *Proceedings of the HCI International Conference* 2018.

5.1 Introduction

As discussed in the literature in chapter 2, Baby Boomers, born between 1946 and 1964 (age in 2017, 53 to 71) are the second largest generation in the U.S. Thus examining the reading behavior of older users and comparing it to those of younger users allows designers to better meet the needs of both user populations. Additionally, recent research calls for designing advanced systems that can respond to user needs in real time. To achieve this goal, various studies are needed to identify eye movements that can reliably detect user experience. To address this need, in this study we examined eye movement factors that are likely to reflect the overall reading experience of Baby Boomers and Generation Y. As presented in Chapter 2, fixation and saccade are two major eye movements that represent information processing behavior. Thus fixation and saccade metrics are extracted to investigate information processing during reading.

An eye tracking study was conducted with 20 participants including 10 young generation and 10 baby boomers. The task required each participant to read a text passage about law and to provide answers to a set of questions about the passage, while the participants' eye movements were recorded by a high speed eye-tracking device. The main objective of the study was to investigate a range of eye movement data that prove to be important in reading behavior and to examine whether these eye movements can reliably predict user age group and performance. This investigation not only facilitates a better understanding of the differences in reading behavior between the two generations but also contributes to research that aims at designing advanced systems. Identifying eye movement metrics that reliably predict a user's age group can help in designing adaptive systems that can respond to older and younger users appropriately in real-time.

5.2 Methodology

This section provides a brief review of the laboratory experiment that was conducted to collect eye movement data used in this study. Furthermore, it provides details on the method used to process the eye-movement data captured from a number of participants who completed a cognitive task online.

5.2.1 Experiment Design

The task selected for this study included reading a passage and answering three questions about the passage. The passage was selected from a pool of GRE sample practice passages available on www.majortest.com. The topic of passage was about law and included 553 words. The passage yielded an overall readability score of 16.1 (Flesh-Kincaid grade level = 16.5, Gunning Fog Index = 20.1, Coleman-Liau Index = 9.7, SMOG Index = 15.8, Automated Readability Index = 18.1) which corresponded to a rather difficult reading level. As in prior research, the readability score was measured using the online tool: https://readable.io/text/. Participants were recruited among college students and staff from a northeastern university at US. Of 20 participants, ten were among young generation (age range of 18-30) and the other ten were baby boomers (age range 53-70). Each participant received a small incentive for their participation in the study. The eye tracker was calibrated for each participant before starting the task. This process requires participants to observe a moving dot on the eye-tracking screen. Tobii software version 3.4.5, and I-VT filter with 30°/sec saccadic velocity threshold was used to process raw gaze data into fixations and saccades.

5.2.2 Data Preprocessing

Studies suggests that older users are more “patient” than younger users when they view online material. They are likely to expend more cognitive effort when scanning a web page and tend to scan more areas on the web page (Chadwick-Dias et al., 2003; Djamasbi et al., 2011). This difference in behavior is likely to be observed via saccadic eye movements when processing textual information. Willingness to expend more cognitive effort is likely to reveal itself in saccadic eye movements, which represent effort to move the eyes from one area of interest and refocus it on another area of interest. The list of saccadic eye movement metrics that we used in our study are displayed in Table 5.1.

Table 5. 1 List of Eye Tracking Metrics

1	Regressive Saccade Count
2	Progressive Saccade Count
3	Average Saccade Duration
4	Average Progressive Saccade Amplitude*
5	Average Regressive Saccade Amplitude*

**Saccade Amplitude* (measured in degree) refers to the visual angle that a gaze travel during a saccade

Eye movement data obtained from the eye tracking software included the x and y-coordinates of the participant's eye location on the screen (pixel), whether the eye movement was a fixation or saccade, and the duration of fixation or saccadic event in milliseconds. Additionally, the software provided the visual angle (measured in degree) that a gaze travel during a saccade (saccade amplitude). Table 5.2 displays the algorithm that we developed to calculate regressive and progressive saccades using x and y- coordinates of two consecutive fixation points. According to Rayner (2009) regressive saccades are backward saccades to a word or a line which were occurred earlier in the text, and hence they can be computed based on the positional information of consecutive fixations (Rayner, 1998; Rayner et al., 2006). I calculated regressive saccades as those in the opposite directions of reading (to the negative of x- and y-direction with respect to the top left corner of the screen delineated as x=0 and y=0).

Table 5. 2 Regressive and Progressive Saccade Tracking Procedure

```
Locate the origin of the gaze x-y coordinate from eye-tracking system[1].
If
(absolute changes in Y values of the most recent consecutive gaze points (k-1 and
k) is less than a predefined threshold, $TH_{inline}$[2],

        $|Y_{gaze}(k) - Y_{gaze}(k-1)| =< TH_{inline}$     (1)

Then check for the changes in X values of those gaze points,
$X_{gaze}(k)$ & $X_{gaze}(k-1)$,

If $X_{gaze}(k) - X_{gaze}(k-1) < -TH_{inline\_Regress}$,[3]

    (It indicates regressive saccade)

Else If $X_{gaze}(k) - X_{gaze}(k-1) > 0$

    (it indicates progressive saccade)
)
Otherwise[4]
(Check if the reader is looking at the point upper than its previous gaze or lower.
If $Y_{gaze}(k) - Y_{gaze}(k-1) < -TH_{inline}$,

(It indicates regressive saccade)

If $Y_{gaze}(k) - Y_{gaze}(k-1) > 0$

(It indicates progressive saccade)
)
End
```

In this study, I was interested in examining overall page scanning behavior. Thus, I excluded shorter regressive saccades that are typically only three characters long (Rayner, 2009).

[1] (The origin (0.,0) is on top left corner of the screen in Tobii X300, which means reading a text from left to right would return gaze points with increasing x values, and reading from top of the text down toward next lines would return gaze points with increasing y values)

[2] $\mathrm{TH}_{\mathrm{inline}}$ is the maximum pixel difference between each lines of the text on interface, which checks whether the reader is in the same line or went to a new line.

[3] $\mathrm{TH}_{\mathrm{inline\_Regress}}$ is number of pixels that include 3 letter character. This threshold is adopted from Reyner et al. 2009).

[4] the reader is reading from a different line: $|\mathrm{Y}_{\mathrm{gaze}}(\mathrm{k}) - \mathrm{Y}_{\mathrm{gaze}}(\mathrm{k}-1)| > \mathrm{TH}_{\mathrm{inline}}$

5.2.3 Regression Analysis

To investigate whether the age group (i.e., Generation Y, Baby boomer) of the user can be detected during reading a passage online through saccadic eye-movements data, regression analysis was performed using the eye movements given in Table 5.1 as independent variables. Equation 1 shows the regression model used in this study.

$$f(x) = \sum_{i=1}^{5} a_i \, x_i + \mathrm{b}, \qquad (1)$$

Where f(x) is a binary dependent variable: $f(x) = \begin{cases} 1, & baby\ boomer \\ 0, & young\ generation \end{cases}$

x_i represents each of the eye metrics shown in table 1, and a_i are the coefficients corresponding to each metric, and b is the intercept.

Saccadic eye movements are representative of reading difficulty (Rayner, 1998), hence, I expected to detect a correlation between performance and saccadic metrics. To investigate this possibility, we used the following regression model.

$$g(y) = \sum_{i=1}^{5} c_i \, y_i + \mathrm{d}, \qquad (2)$$

Where g(y) is refers to performance measured as the number of correct answers to three multiple choice questions. As in equation 1, y_i represents each of the eye metrics shown in table 2, and c_i are the coefficients corresponding to each metric.

5.3 Results

Mean and standard deviation of variables of interest are displayed in Table 5.3. As the values in Table 5.3 indicate, younger users on average had longer (in duration) and more saccadic eye movements. This behavior is consistent with previous research that suggests younger users, compared to older users, exhibit less patient viewing behavior (Djamasbi et al., 2011). The results also showed that older people had larger saccade amplitude, which indicates that to process the provided information their eyes traveled longer distances to scan the text. This eye movement behavior, consistent with previous research (Djamasbi et al., 2011), suggests a greater degree in willingness to expend cognitive effort to read textual information.

Table 5. 3Mean and Standard Deviation for the Eye Movement Variables for Each Age Group

Eye movement Features	**Younger users**	**Older users**
Regressive Saccade Count	124.3	106.9
Progressive Saccade Count	480.7	455.7
Avg. Saccade Duration (msec)	28.51 (±3.25)	25.50 (±3.23)
Avg. Progressive Saccade Amplitude (degree)	3.65 (± 0.34)	4.20 ((± 0.34)
Avg. Regressive Saccade Amplitude (degree)	4.94 (± 1.90)	5.06 (±1.26)

Table 5.4 presents the results of regression analysis as modeled by equation 1. As the results show the two groups did not differ significantly in regressive saccades. However, the progressive and regressive saccade amplitudes, as well as saccade duration and progressive saccade counts were significantly correlated with the age group of the users. The results also show a stronger effect for the relationship between progressive and regressive saccade amplitudes and age of the users (as attested by the stronger p value and larger beta value). These results suggest that saccadic eye movements may serve as a reliable predictor of users' age group.

Table 5. 4 Results of Regression Analysis for Different Age Groups as Dependent Variable and Eye Movements as Independent Variables

$R^2 = 0.87$, Adj $R^2 = 0.83$			
Eye Movement Metric	**t-stat**	**P-value**	**Beta**
Regressive Saccade Count	1.27	0.22	-0.1
Progressive Saccade Count	**2.47**	**0.02**	**0.19**
Avg. Saccade Duration	**2.14**	**0.04**	**0.12**
Avg. Progressive Saccade Amplitude	**8.38**	**7.9E-7**	**-0.59**
Avg. Regressive Saccade Amplitude	**5.88**	**3.9E-5**	**-0.40**

As mentioned earlier after reading the passage each participant was asked to provide answers to three questions about the passage. To further explore the differences between young generation and baby boomers we looked at the difference in performance of these two groups using two sample t-test. The results revealed no significant difference between the two age groups in performance. The results of the t-test support the results reported in Table 5.3, showing no significant differences in regressive saccades between the two groups. The observed behavior support previous research that showed while older adults were slower in cognitive processing, they performed relatively similar to younger adults.

I also investigated the relationship between eye movements and performance. In other words, I examined whether eye movements can be used to predict the reading comprehension performance of users. To do so, I ran a regression on performance as the dependent variable and saccadic eye movement variables (Table 5.1) as independent variables. In the regression analysis, I used performance as a categorical variable with different values of (0, 1, 2, 3), where zero corresponds to no correct answers at all, and three corresponds to answering all the questions right. Since the performance of the two groups was not significantly different, we did not separate the two age groups. The results of regression analysis are shown in Table 5.5.

Table 5. 5 Results of Regression Analysis for Performance as Dependent Variable and Eye Movements as Independent Variables

$R^2 = 0.61$, Adj $R^2 = 0.15$			
Eye Movement Metric	**t-stat**	**P-value**	**Beta**
Regressive Saccade Count	0.93	0.37	0.32
Progressive Saccade Count	0.36	0.73	0.13
Avg. Saccade Duration	0.02	0.99	-0.00
Avg. Progressive Saccade Amplitude	0.42	0.68	0.13
Avg. Regressive Saccade Amplitude	1.96	0.07	0.60

As the results in Table 5.5 show none of the saccadic eye metrics were predictive of the task performance in the reading task.

5.4 Fixation Analysis

In the previous section I examined saccadic eye movements that may predict user age group and/or reading comprehension. In this section I looked at possible differences between the two user groups in regards to fixations. Note that consistent with prior research, fixations with durations shorter than 100 ms were filtered out from the fixation data (Rayner, 2009).

Figure 5.1 shows the heat map of aggregated gaze duration between the two groups of users, (a) for young generation and (b) for baby boomers. Green corresponds to minimum gaze duration, and red corresponds to maximum gaze duration (10.58 s in this heat map), which is the aggregation of gaze duration over all the participants who read the passage. The heat maps of total gaze duration do not seem to reveal significant differences between the reading behaviors of the two groups of users.

In addition to qualitative analysis using heat maps I also conducted a regression between different age groups as dependent variable and fixation eye metrics as independent variables. The result of regression analysis is given in Table 5.6. As the results show fixation metrics, such as average fixation duration and average fixation count, were not significantly correlated with the age group of the users.

Table 5. 6 Results of Regression Analysis for Different Age Groups as Dependent Variable and Eye Movements as Independent Variables

$R^2 = 0.033$, Adj $R^2 = 0.08$

Eye Movement Metric	**t-stat**	**P-value**	**Beta**
Fixation Count	0.44	0.66	0.05
Average Fixation Duration	-0.66	0.52	-0.08

5.5 Discussion and Conclusion

In this research, we examined the differences between young and old adults in online reading experience by comparing their eye movement behavior. Past research indicates that older adults are likely to expend more effort when processing information (Chadwick-Dias et al., 2003; Djamasbi et al., 2011; Kemper et al., 2004; Rayner et al., 2006).

Building on the previous research I examined whether differences between the two user groups reading textual information can be detected using their eye movements. Because I was examining overall reading behavior (over the entire text passage) I expected to see differences in saccadic eye movements. The results show that saccadic eye movements (both regressive and progressive), as well as saccade duration and saccade counts in reading was a significant predictor of the user age group. The results extend previous literature in reading (Rayner, 1998, 2009). First, in this study I focus on overall passage reading rather than sentence or word by word processes. Second, the results suggest that saccadic metrics may serve as a strong predictor of users' age group. Third, the results indicate that average regressive saccade amplitude may serve as a predictor of reading comprehension. These findings have important implications for capturing online and/or screen reading experience of textual information. For example, it can be used to examine the impact of text simplification on reading experience (Djamasbi et al., 2016).

Fixation metrics such as fixation duration or fixation count were not significant in identifying the differences between the two age groups in our study. I also did not observe any major differences between the aggregated fixation duration of the participants on the passage, according to Fig 5.1. This may be because in this study I focused on passage level reading experience and not on the sentence or word level analysis. For example, I did not consider the effect of word predictability or word frequency in fixation duration during reading. Additionally, the sample size was small; by expanding the sample size I may also see significant differences in fixation metrics between old and young users.

The analysis examining the relation between performance and eye movement revealed that reading comprehension performance was not correlated with eye movements. It is likely that with a larger sample size the relationship between regressive saccade and performance would become stronger.

Overall these findings are consistent with previous literature that have indicated that there are differences between young and old adults in online reading and web experience. For example, Rayner et al. (2006) investigated the differences between older and younger adults in reading and learned that older adults make more fixations, longer fixations and more regressions. I also saw significant differences between regressive saccadic eye movements among younger and older adults during reading (Table 5.4). Overall, the results showed that user population (younger vs. older) when they read textual information can be predicted using the eye movement data. These results add to the previous research by investigating the eye movements that are representative of cognitive processing during online reading, and by focusing on the passage level reading rather than sentence level reading, and by comparing the reading behavior of younger and older adults on a relatively long and difficult text passage.

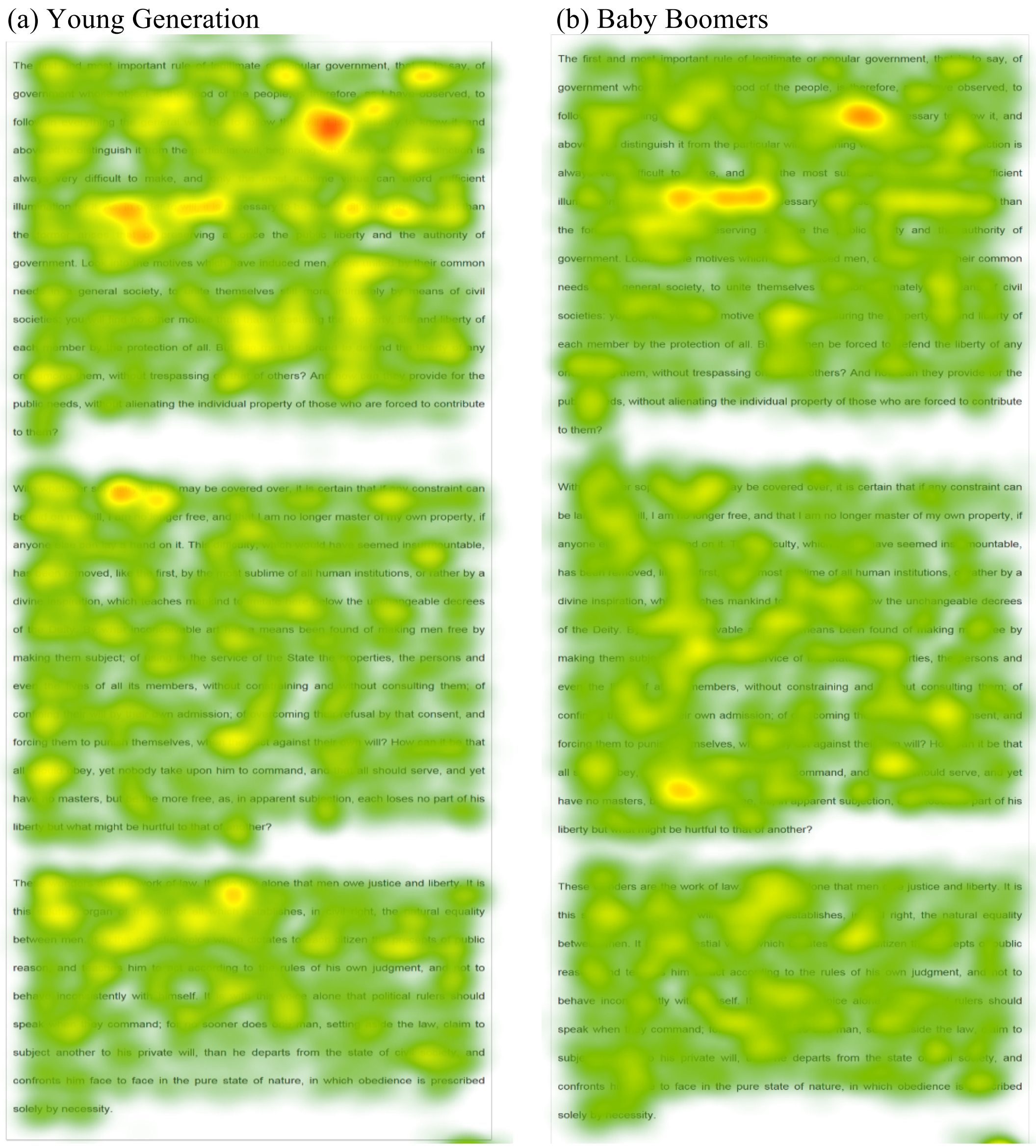


Figure 5. 1 Heat map of aggregated gaze duration – a comparison between (a) young generation and (b) baby boomers

As in any other research our study is not without limitations. Such limitations, however, provide opportunity for directing future research efforts. For example, future studies, including some of our own planned experiments, are needed to test text passages with different content other than law to see whether similar results are obtained. Expanding population to a larger number can also help enhancing the generalizability of our results.

6 Study Four: Developing a Predictive Model for Reading Based on Age and Cognitive Load

6.1 Introduction

Reading interfaces are important in HCI research as they allow the users to interact with the computer interface in their normal language (Attar, 2016). Any improvement in reading interfaces can lead to better performance and comprehension in a wide variety of important tasks.

The goal of this study is to develop two machine learning models that can be used to design a user profiling model based on age and cognitive load of the user during a reading comprehension task. The proposed approach consists of user characterization and user's cognitive load measurement, while the user reads text. Figure 6.1 illustrates the proposed theoretical framework. Eye tracking is used to collect user's eye-movement while reading passages. The eye-movement data is preprocessed and used to train two machine learning models, one for identifying the user's age characteristics (e.g., older vs. younger user) and one for detection of level of cognitive load (due to task condition) that user is experiencing while reading texts.

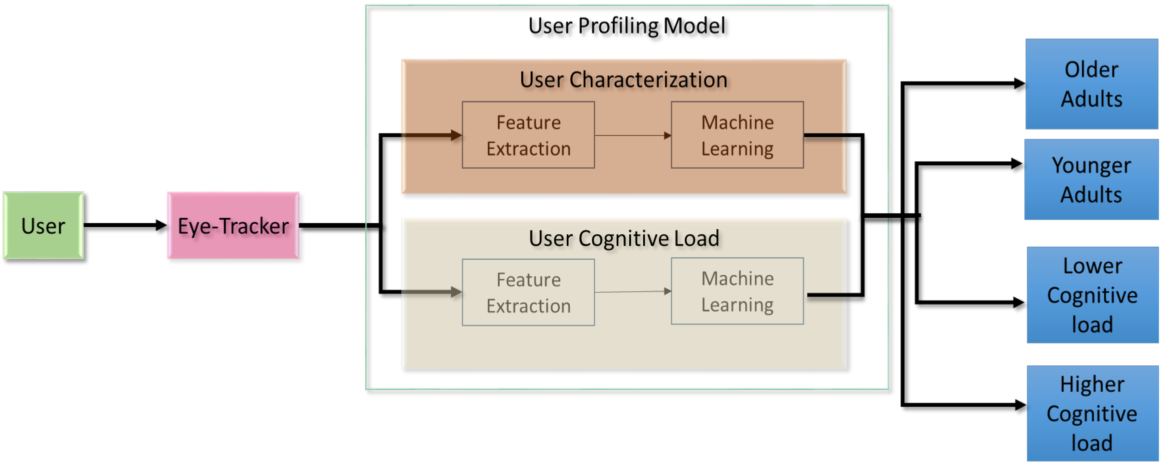


Figure 6. 1 User Profiling Model Block Diagram

The discussed literature in chapter 2 suggests that 1) eye movements can be used to understand the cognitive processing during reading (section 2.3), and 2) there are differences between younger and older adults' online reading and web experience (section 2.4). Taking the differences into account, the first research question is whether the age group of a user (younger/older) can be distinguished using a machine learning model of user's eye-movements when they read textual information, and second question is whether the level of cognitive load can be also detected using another machine learning model using eye movement metrics.

Eye tracking is used to collect user's eye movements while reading passages. The raw eye-movement data is preprocessed and used to detect the user's task load (e.g., higher or lower extraneous cognitive load) as well as user's age characteristics (younger vs. older user). Detecting user's level of cognitive load, and user's age characteristic is important because it can help research in adaptive user interfaces that can provide a personalized experience for individual users.

The machine learning models I developed in this study can automatically detect the following scenarios:

Scenario A. User is older, and has a higher level extraneous cognitive load (or task load)

Scenario B. User is younger, and has a lower level extraneous cognitive load (or task load)

Scenario C. User is younger, and has a higher level extraneous cognitive load (or task load)

Scenario D. User is older, and has a lower level extraneous cognitive load (or task load)

The output of my model can be used in future studies that design adaptive textual interface to better address older and younger user needs.

6.2 Methodology

I designed an eye tracking study to collect eye movement data during reading. A passage selection process was performed to select 2 passages with different topics for the study. The goal was to adopt a subject-rating method to sort passages based on their level of difficulty. The following subsection provides details of the adopted passage selection process.

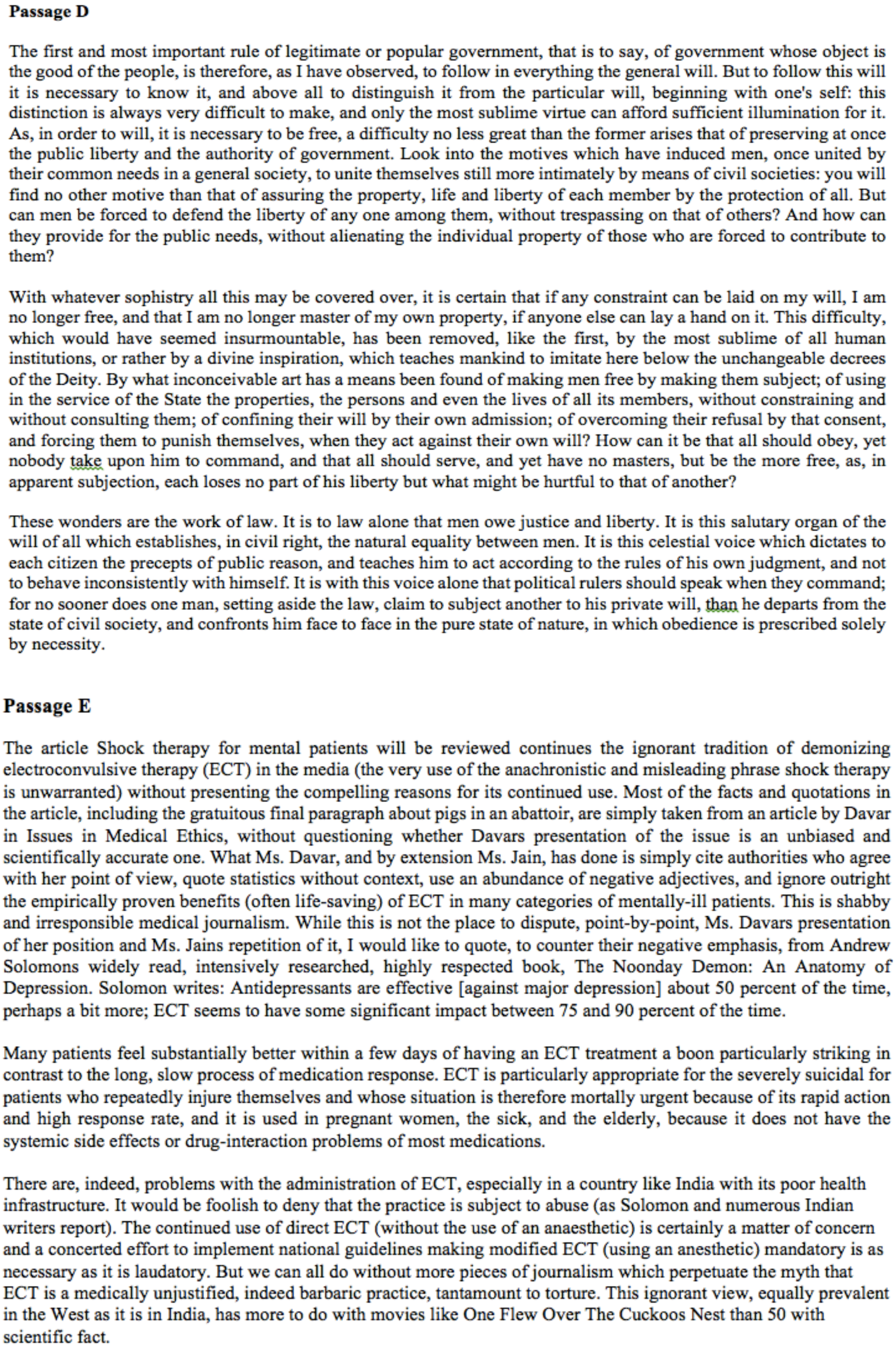
6.2.1 Passage Selection Process

Six passages with different topics were chosen from a GRE sample test practices online.[5] To select the best passages for this study, I went through a systematic procedure. First I measured the readability score of each passage using the free online tool: https://readable.io/text/. Then I picked six passages with highest level of difficulty (highest readability score), and a moderate length (see figure A.1). The goal was to choose two passages among the six passages that not only had highest readability score, but were also rated by participants as the most difficult passages to read and understand. To obtain feedback on the passages difficulty level, I designed a moderated lab study, where 18 students (10 male, 8 female, age range between 22 to 31) were recruited to participate in the study. Each participant was assigned to read a set of four passages (randomly selected among six passages). The task included reading the passages, and rating the difficulty of the passage using the Subjective Mental Effort Questionnaire (SMEQ), which was used to subjectively measure how hard the participant found the passage to comprehend. SMEQ is a single item questionnaire with a rating scale from 0 to 150, which includes nine verbal extending from "Not at all hard to do" (just above zero) to "Tremendously hard to do " (above 110) in their scale (Sauro & Lewis, 2012). The study then followed with an interview session to gather more information about the overall reading experience of the passages. In particular, the participants were asked to rate the task difficulty from the scale of 1 to be easiest to do and 7 to be the hardest to do. The results of SMEQ scores and rating scores obtained during interviews are reported in Table 6.1.

The criteria for choosing the most difficult two passages for the study was based on average SMEQ rating, and the interview ratings. Passages D and E which received the highest scores for SMEQ (66.11, and 70.22), highest rating (5.06 and 4.67) were selected among the six passages to proceed with. Passage D and E yielded overall readability scores of 16.1 and 16.4 correspondingly, which both corresponded to a rather difficult reading level. Passage D was

[5] www.majortests.com

about law and passage E was about Electric Shock Therapy (see Fig. 6.2). The rest of passages are provided in Appendix A.



Passage D

The first and most important rule of legitimate or popular government, that is to say, of government whose object is the good of the people, is therefore, as I have observed, to follow in everything the general will. But to follow this will it is necessary to know it, and above all to distinguish it from the particular will, beginning with one's self: this distinction is always very difficult to make, and only the most sublime virtue can afford sufficient illumination for it. As, in order to will, it is necessary to be free, a difficulty no less great than the former arises that of preserving at once the public liberty and the authority of government. Look into the motives which have induced men, once united by their common needs in a general society, to unite themselves still more intimately by means of civil societies: you will find no other motive than that of assuring the property, life and liberty of each member by the protection of all. But can men be forced to defend the liberty of any one among them, without trespassing on that of others? And how can they provide for the public needs, without alienating the individual property of those who are forced to contribute to them?

With whatever sophistry all this may be covered over, it is certain that if any constraint can be laid on my will, I am no longer free, and that I am no longer master of my own property, if anyone else can lay a hand on it. This difficulty, which would have seemed insurmountable, has been removed, like the first, by the most sublime of all human institutions, or rather by a divine inspiration, which teaches mankind to imitate here below the unchangeable decrees of the Deity. By what inconceivable art has a means been found of making men free by making them subject; of using in the service of the State the properties, the persons and even the lives of all its members, without constraining and without consulting them; of confining their will by their own admission; of overcoming their refusal by that consent, and forcing them to punish themselves, when they act against their own will? How can it be that all should obey, yet nobody take upon him to command, and that all should serve, and yet have no masters, but be the more free, as, in apparent subjection, each loses no part of his liberty but what might be hurtful to that of another?

These wonders are the work of law. It is to law alone that men owe justice and liberty. It is this salutary organ of the will of all which establishes, in civil right, the natural equality between men. It is this celestial voice which dictates to each citizen the precepts of public reason, and teaches him to act according to the rules of his own judgment, and not to behave inconsistently with himself. It is with this voice alone that political rulers should speak when they command; for no sooner does one man, setting aside the law, claim to subject another to his private will, than he departs from the state of civil society, and confronts him face to face in the pure state of nature, in which obedience is prescribed solely by necessity.

Passage E

The article Shock therapy for mental patients will be reviewed continues the ignorant tradition of demonizing electroconvulsive therapy (ECT) in the media (the very use of the anachronistic and misleading phrase shock therapy is unwarranted) without presenting the compelling reasons for its continued use. Most of the facts and quotations in the article, including the gratuitous final paragraph about pigs in an abattoir, are simply taken from an article by Davar in Issues in Medical Ethics, without questioning whether Davars presentation of the issue is an unbiased and scientifically accurate one. What Ms. Davar, and by extension Ms. Jain, has done is simply cite authorities who agree with her point of view, quote statistics without context, use an abundance of negative adjectives, and ignore outright the empirically proven benefits (often life-saving) of ECT in many categories of mentally-ill patients. This is shabby and irresponsible medical journalism. While this is not the place to dispute, point-by-point, Ms. Davars presentation of her position and Ms. Jains repetition of it, I would like to quote, to counter their negative emphasis, from Andrew Solomons widely read, intensively researched, highly respected book, The Noonday Demon: An Anatomy of Depression. Solomon writes: Antidepressants are effective [against major depression] about 50 percent of the time, perhaps a bit more; ECT seems to have some significant impact between 75 and 90 percent of the time.

Many patients feel substantially better within a few days of having an ECT treatment a boon particularly striking in contrast to the long, slow process of medication response. ECT is particularly appropriate for the severely suicidal for patients who repeatedly injure themselves and whose situation is therefore mortally urgent because of its rapid action and high response rate, and it is used in pregnant women, the sick, and the elderly, because it does not have the systemic side effects or drug-interaction problems of most medications.

There are, indeed, problems with the administration of ECT, especially in a country like India with its poor health infrastructure. It would be foolish to deny that the practice is subject to abuse (as Solomon and numerous Indian writers report). The continued use of direct ECT (without the use of an anaesthetic) is certainly a matter of concern and a concerted effort to implement national guidelines making modified ECT (using an anesthetic) mandatory is as necessary as it is laudatory. But we can all do without more pieces of journalism which perpetuate the myth that ECT is a medically unjustified, indeed barbaric practice, tantamount to torture. This ignorant view, equally prevalent in the West as it is in India, has more to do with movies like One Flew Over The Cuckoos Nest than 50 with scientific fact.

Figure 6. 2 Passages Selected for the Study Based on Participants Subjective Rating

Table 6. 1 Results of Subjective Rating and SMEQ for the Six Passages

Passage	A	B	C	D	E	F
SMEQ Score	36.44	44.78	63.33	**66.11**	**70.22**	53.89
Rating	3.56	4.11	4.50	**5.06**	**4.67**	4.50

As mentioned earlier, one dimension of user profiling is cognitive load (Figure 6.1). Hence, to manipulate the level of cognitive load, the next step was to simplify passages D and E using plain language standards used in previous research (Djamasbi, et al., 2016-a) (Table 4.1). The original passages D and E are named OA and OB, and the simplified versions are named SA and SB in the rest of this section.

Previous research has shown that simplifying passages using the plain language standards displayed in Table 4.1 decreases the level of cognitive load or task demand (Djamasbi, et al., 2016-b). Simplifying using plain language rules reduced passage OA readability score from 16.1 to 10 in SA and passage OB readability score from 16.4 to 11 in SB.

Once the passages were selected and simplified using plain language rules, the next step was to design the eye-tracking study to collect participants' eye-movement data during reading of the original and simplified versions of the two passages selected for this study.

6.2.2 Eye Tracking Experiment Design

A total of 136 participants were recruited to participate in the study, among which 80 were from a younger population (18 to 30 years old, mean age=23.5), and 56 were older adults (50-70 years old, mean age=56), and 84 females and 52 males. Each participant received a small incentive for their participation in the study.

Each participant was assigned to read two passages (one original, one simplified) on the screen. To eliminate the order bias, passages were presented to participants in a Latin square fashion, i.e., out of two original passages and two simplified passages, four pairs (OA-SB, SB-OA, OB-SA, SA-OB) were created. To rate the mental effort spent in reading and comprehending the passage, after reading each passage, participants were asked to answer the SMEQ survey which was presented to the participant immediately after reading each passage. Participants were told they are required to answer questions about passages they read. This was to encourage careful reading of the passages. After administering of SMEQ, participants were asked to provide answers to three questions about each passage. The passage was given to the participants for their references, on the left side of the screen that included the questions. Reading and decision making (e.g., selecting the correct answer to a question) are different cognitive activities. Because the focus of this study was user profiling during reading, only the eye movement during reading task was used for analysis.

6.3 Data Analyses

Eye tracking technology was used to collect the stream of eye movement data while the participants were completing the reading task. The eye tracking software and hardware used in this study is similar to the one used in the previous studies in previous chapters. The software was an updated version of Tobii Studio 3.4.5. I-VT filter with 30°/sec saccadic velocity threshold was used to identify fixations and saccades in the gaze stream. A set of eye-tracking metrics, or feature matrix was obtained and computed from eye-tracking raw recordings of gaze coordinates, gaze duration and pupil dilation during fixation and saccade events. The next section describes the details of how the eye movements feature matrix was computed.

Of 56 older users who participated in the study, four did not calibrate well, or their gaze sampling percentage was less than 80% (Kruger et al., 2013). Of the 80 younger participants two students did not calibrate. Thus, the data from these six participants were removed from the data as suggested by Kruger et al., (2013) .

6.3.1 Building of Eye-tracking Feature Matrix

As mentioned earlier in section 2.3 eye movements such as fixation duration, fixation number, and number and duration of regressive and progressive saccades have been used by a number of researchers in the eye tracking-reading literature to understand the link between eye movements and reading (Rayner, 1998; Ashby et al., 2005; Rayner 2009; Rayner and Pollatsek 2012; Campbell & Bovee, 2014). In this study, I used the eye movement measures reported in the previous research for building the feature matrix. I also added pupillary data to the feature matrix, which is further explained in the following section. The reason for adding the pupillary data is that my earlier studies showed this data was successful in detecting cognitive load.

Each parameter was measured over the duration of the task (reading two text passages) completed by each participant in the study. Machine learning feature sets are often developed using statistical properties of fundamental parameters. Hence, basic statistical properties, such as mean and standard deviation, were calculated for each of the parameters.

Fixation Features

Tobii software provides raw gaze durations and gaze types (fixation/saccade) correspond to gaze event. Three features were extracted for fixation events. The average and standard deviation of fixation durations, and total duration of all fixations were calculated. Since the passages had different length, total fixation duration was divided by words count of each passage to obtain and standardized value for each participant's fixation. Table 6.2 present the fixation features computed to be included in the feature matrix.

Table 6. 2 List of Fixation Metrics

1	Total Fixation Duration
2	Average Fixation duration (millisecond)
3	STD of Fixation duration (millisecond)

Saccadic Features

Average and standard deviation of saccade duration, average and standard deviation of regressive and progressive saccade durations as well as average and standard deviation of regressive and progressive saccade amplitude were calculated. Number of regressive and progressive saccades and their ratio were also calculated. Table 6.3 illustrates all saccadic features calculated to be included in the study's feature matrix.

Table 6. 3 List of Saccadic Metrics

1	Average Saccade Duration
2	STD of Saccade Duration
3	Average Regressive Saccade Duration
4	STD of Regressive Saccade Duration
5	Average Progressive Saccade Duration
6	STD of Progressive Saccade Duration
7	Regressive Saccade Count
8	Progressive Saccade Count
9	Regressive Saccade Count/Progressive Saccade Count
10	Average Regressive Saccade Amplitude
11	STD of Regressive Saccade Amplitude
12	Average Progressive Saccade Amplitude
13	STD of Progressive Saccade Amplitude

Pupillary Features

Tobii Studio software also provides raw pupil dilation data. I also calculated the Pupil Dilation Variation (PDV) or rate of change of pupil dilation by taking the temporal derivative of pupil dilation (see section 3.1). I used pupillary data such as average and standard deviations of pupil dilation and pupil dilation variation as suggested by study I in chapter 3. As shown in section 4.1, when pupillary data including pupil dilation (PD) and pupil dilation variation (PDV) are separated during saccade and fixation, they provide invaluable information about cognitive effort during reading. Therefore, in this study, I separated the pupillary data during fixation and saccade. Time series analyses of pupillary data during a reading task in section 4.1 revealed that partitioning the time interval of reading into smaller intervals of beginning, middle and end, and analyzing pupillary data within each interval provides additional information about cognitive effort during reading. In this study, however, dividing the task time into smaller intervals will result in three different values for each pupillary metrics which is not efficient for the classification task. Therefore, instead of time domain analyses I used frequency domain analyses of pupillary data by computing the power spectral density of pupillary data.

Six frequency domain features were calculated from preprocessed pupil dilation and pupil variation signals. Fast Fourier transform algorithm (Cooley & Tukey, 1965), and R package "spectral" were used to calculate power spectral density (PSD) of pupil dilation and variation data. Spectral or frequency domain analysis in a nutshell is decomposition of a time series into underlying sine and cosine functions of different frequencies, which allows us to determine those frequencies that appear particularly strong or important. The reader can refer to (Bloomfield, 2000) for more information regarding frequency domain analysis. Area under the PSD curve (AUC) for both pupil dilation and pupil dilation variation were calculated and were used as the frequency domain features. Further, AUC was calculated for PSD during fixation and saccade separately. These features were then included in the feature matrix set. (Table 6.4).

Table 6. 4 List of Pupillary Metrics

1	Pupil Dilation PSD
2	Pupil Variation PSD
3	Pupil Dilation-Fixation PSD
4	Pupil Dilation-Saccade PSD
5	Pupil Variation-Fixation PSD
6	Pupil Variation-Saccade PSD
7	Avg. Pupil Dilation-Fixation
8	Avg. Pupil Dilation-Saccade
9	STD of Pupil Dilation-Fixation
10	STD of Pupil Dilation-Saccade

11	Avg. Pupil Variation-Fixation
12	Avg. Pupil Variation-Saccade
13	STD of Pupil Variation-Fixation
14	STD of Pupil Variation-Saccade
15	Avg. Pupil Fixation/Avg. Pupil Saccade
16	Avg. Pupil Variation Fixation/Avg. Pupil Variation Saccade

Combined Ratio Features

The results in chapter 3 showed that the ratio of pupil dilation during saccade to fixation was one of the most important metrics in the classification task (Table 3.7). Hence, in this study pupillary data were also combined to develop ratios that could provide additional insight. For example, the ratios of PD and PDV during saccade and fixations were calculated and added to the feature set (Table 6.4). The ratio of pupil dilation and variation during saccades and fixations reflect the distribution of cognitive effort during information search and information processing.

By putting together, the above features (Table 6.2 to 6.4), an eye-tracking feature matrix including 33 features is obtained. Table 6.5 presents the total set of 33 eye-tracking metrics used in this study.

Table 6. 5 List of Eye Movement Metrics used in the Reading Comprehension Study

Eye Movements	**Eye Metrics (Features)**
Fixation	Total Fixation Duration Avg. Fixation Duration STD Fixation Duration
Saccade	Avg. Saccade Duration STD Saccade Duration Avg. Regressive Saccade Duration STD Regressive Saccade Duration Avg. Progressive Saccade Duration STD Regressive Saccade Duration Avg. Progressive Saccade Amplitude STD Progressive Saccade Amplitude Avg. Regressive Saccade Amplitude STD Regressive Saccade Amplitude Regressive Saccade Count Progressive Saccade Count

	Regressive Saccade Count/Progressive Saccade Count
Pupil Dilation	Pupil Dilation PSD Pupil Variation PSD Pupil Dilation-Fixation PSD Pupil Dilation-Saccade PSD Pupil Variation-Fixation PSD Pupil Variation-Saccade PSD Avg. Pupil Dilation-Fixation Avg. Pupil Dilation Saccade STD of Pupil Dilation-Fixation STD of Pupil Dilation-Saccade Avg. Pupil Variation-Fixation Avg. Pupil Variation-Saccade STD of Pupil Variation-Fixation STD of Pupil Variation-Saccade
Combined Eye Movements	Avg. Pupil Fixation/Avg. Pupil Saccade Avg. Pupil Variation Fixation/Avg. Pupil Variation Saccade Avg. Fixation Duration/Avg. Saccade Duration

The feature metrics computation and preprocessing used for this study were implemented in R version 3.4.2 on Windows 7, with Core i5 CPU and 3.30 GHz speed machine. I used R libraries such as “spectral” (Maintainer & Seilmayer, 2016) for frequency domain analysis of pupillary data, and “pracma” (Hans etal., 2018) to calculate the area under the curve of PSDs.

6.3.2 Developing the Eye Tracking Machine Learning Models

Two eye tracking machine learning models were developed in this study to solve the classification problems: Age (younger/older user), and Cognitive Load (original/simplified text). The eye-tracking feature matrix including 33 eye-movement features shown in Table 6.5, was used to develop the model. Table 6.6 shows the feature matrix dimensions for the two classification tasks.

Table 6. 6 Feature matrix used in Cognitive Load classification vs. Age Classifications

	Cognitive Load Classification	Age Classification

Number of subjects	65 [OA] 65 [OB] 130 [OA_OB]	65 [SA] 65 [SB] 130 [SA_SB]	156 [younger]	104 [older]
Number of features	33		33	

Similar to what was used in Chapter 3 (section 3.2.7), the following three steps were adopted to develop the machine learning models:

- Effective feature set for each classification task were chosen according to random forest variable importance and a systematic feature selection approach.

- Training dataset were built using a bootstrapping replacement process for each classification task.

- Random Forest classifiers were applied on the corresponding test datasets.

For more details of the above steps the reader can refer to Chapter 3, section 3.2.7. In this study, I used 200 number of bootstraps (Efron & Tibshirani, 1994), with a random forest classifier of maximum 100 number of trees.

From Table 6.6 one can see that the classes for Cognitive load classifier is balanced (130 original vs 130 simplified), however the classes are imbalanced for the Age machine learning task (156 young vs. 104 old).

Below are reasons which leads to reduction in performance of machine learning algorithm using an imbalanced data set: Machine learning algorithms struggle with accuracy because of the unequal distribution in dependent variable since the assumption is the data set has balanced class distributions, which causes the performance of the classifier to get biased towards majority class. Classifiers can have good accuracy on the majority class but very poor accuracy on the minority classes due to the influence that the larger majority class has on traditional criteria (Yıldırım, 2016). Researchers have generally used two kinds of solutions for data classifications dealing with imbalanced problems: solving in data level by re-sampling, and solving in algorithm level by using design sophisticated classification approaches, where the prior one is mostly preferred (Yıldırım, 2016). The sampling techniques are mainly divided into two subgroups: under sampling and over sampling. Under sampling method removes examples from the majority class to make the data set balanced. This method tries to balance the distribution of class by randomly removing majority class samples. The drawback of under sampling method is that it can discard potentially useful information that could be important for classifiers. Over sampling is a sampling approach which balances the data set by replicating the examples of minority class. The advantage of this method is that there is no loss of data as in under sampling technique. The disadvantage of this technique is that it may lead to over fitting and can introduce an additional computational cost if the data set is already fairly large but imbalanced (Yıldırım, 2016).

In the present study to address the imbalanced classes for the Age classification task, I used an improved under sampling method by using a bootstrap based under sampling technique. The way this process works is in every bootstrapping, training data from 104 randomly chosen younger subjects is used along with the original 104 older subjects' training set to train the random forest classifier. Due to 200 number of bootstraps, this way we can take advantage of all young subjects' training data in building the classifier while making sure that balanced classes are used to perform the classification task.

The machine learning algorithms used in this study were implemented in R version 3.4.2. I used R libraries such as ISLR (James et al., 2018), tree (Brian & Ripley, 2018), random forest, e1071(Meyer et al., 2017), and caret (Max et al., 2018).

6.4 Results

Two machine learning models were developed to detect user's cognitive load (task condition: original vs. simplified passages) and age (young vs. old) using the methodology described above. In the followings, the performance of each machine learning model is discussed. As mentioned before, bootstrapping random forest classifier with 200 replications and maximum number of trees of 100 was used to develop the machine learning models. For details of the random forest algorithm and bootstrapping process, the reader can refer to Chapter 2 of the thesis.

6.4.1 Cognitive Load (CL) Classification Results

As mentioned earlier each participant was randomly assigned to read two passages, one original version and the other simplified. Four sequences were made from four passages (OA-SB, SB-OA, SA-OB, and OB-SA). Therefore, of 130 total participants each passage was read by 65 participants (Table 6.6).

A) Passages Subjective Rating and Performance Results

I used two-sample t-test to verify whether the subjective SMEQ ratings of original and simplified versions of each passage were significantly different. Table 6.7 shows the average, STD and the results of t-test for SMEQ questions. As the results show participants rated passage OA to be significantly more difficult as compared to passage SA (63.69 vs. 46.62). Similarly, participants rated passage OB to be significantly more difficult as compared to passage SB (49.03 vs. 34.40). These results indicate that simplifying passages were effective in reducing the level of cognitive load in reading.

Table 6. 7 Avg., STD and the Results of t-test for Subjective SMEQ Rating for Passages A and B

	OA	SA	OB	SB
Mean	63.69	46.62	49.03	34.40
STD	25.09	26.65	31.91	25.60
p-value	**0.0003**		**0.005**	

Table 6. 8 Avg., STD and the Results of t-test for Performance for Passage A and B

	OA	SA	OB	SB
Mean	1.00	1.05	0.85	0.83
STD	0.85	0.96	0.83	0.74
p-value	0.77		0.91	

To make sure that participants would be engaged in reading the passage they were instructed that they are provided with some questions about the passage after reading the passage. As mentioned earlier the eye movement data during this part (answering questions about the task) was not included in the analysis. Nevertheless, the performance data can provide further insight about the overall difficulty of text passages. Results of performance in Table 6.8 indicates that both original and simplified version of the text were fairly difficult as evidenced by the mean number of correct answers. This is not surprising as the text was selected from the pool of GRE passages, and the ones with highest difficulty level were chosen for this study. To confirm that participants were engaged in reading, gaze videos of each participant was manually reviewed. This manual analysis showed that participants read the provided text carefully and that they were equally engaged in reading both original and simplified passages, even though the level of difficulty was different.

B) Cognitive Load Classification Results

To build the cognitive load machine learning model, feature selection and classification processes were run three times with the following 3 settings.

- Feature matrix with 33 features from participants who read passages OA and SA (33 by 130 feature matrix related to passage A, with 65/65 original/simplified classes)
- Feature matrix with 33 features from participants whom read passages OB and SB (33 by 130 feature matrix related to passage B, with 65/65 original/simplified classes)
- Feature matrix with 33 features from participants who read passages OA, OB, SA and SB (33 by 260 feature matrix related to both passage A and B, with 130/130 original/simplified classes)

Table 6.9 summarizes the above settings.

Table 6. 9 Cognitive Load Classification Tasks and Settings

	CL Classification Task (1)	CL Classification Task (2)	CL Classification Task (3)
Participants involved	$Class\ 1: OA$ (65) $Class\ 2: SA$ (65)	$Class\ 1: OB$ (65) $Class\ 2: SB$ (65)	$Class\ 1: \begin{cases} OA \\ OB \end{cases}$ (130) $Class\ 2: \begin{cases} SA \\ SB \end{cases}$ (130)
Feature Matrix Dimension	33 by 130	33 by 130	33 by 260

Figures 6.3 to 6.5, and tables 6.10 to 6.13 demonstrates the results correspond to above three classification tasks.

For the first classification task, where dataset from OA and SA readers were used to build the classifier, RF variable importance graph (Figure 6.3) shows that one of the frequency domain pupillary feature (area under the curve of PSD of pupil dilation variation) is the most discriminative feature (importance value=7.67) in classifying eye movements of participants who read OA passage versus those who read SA passage.

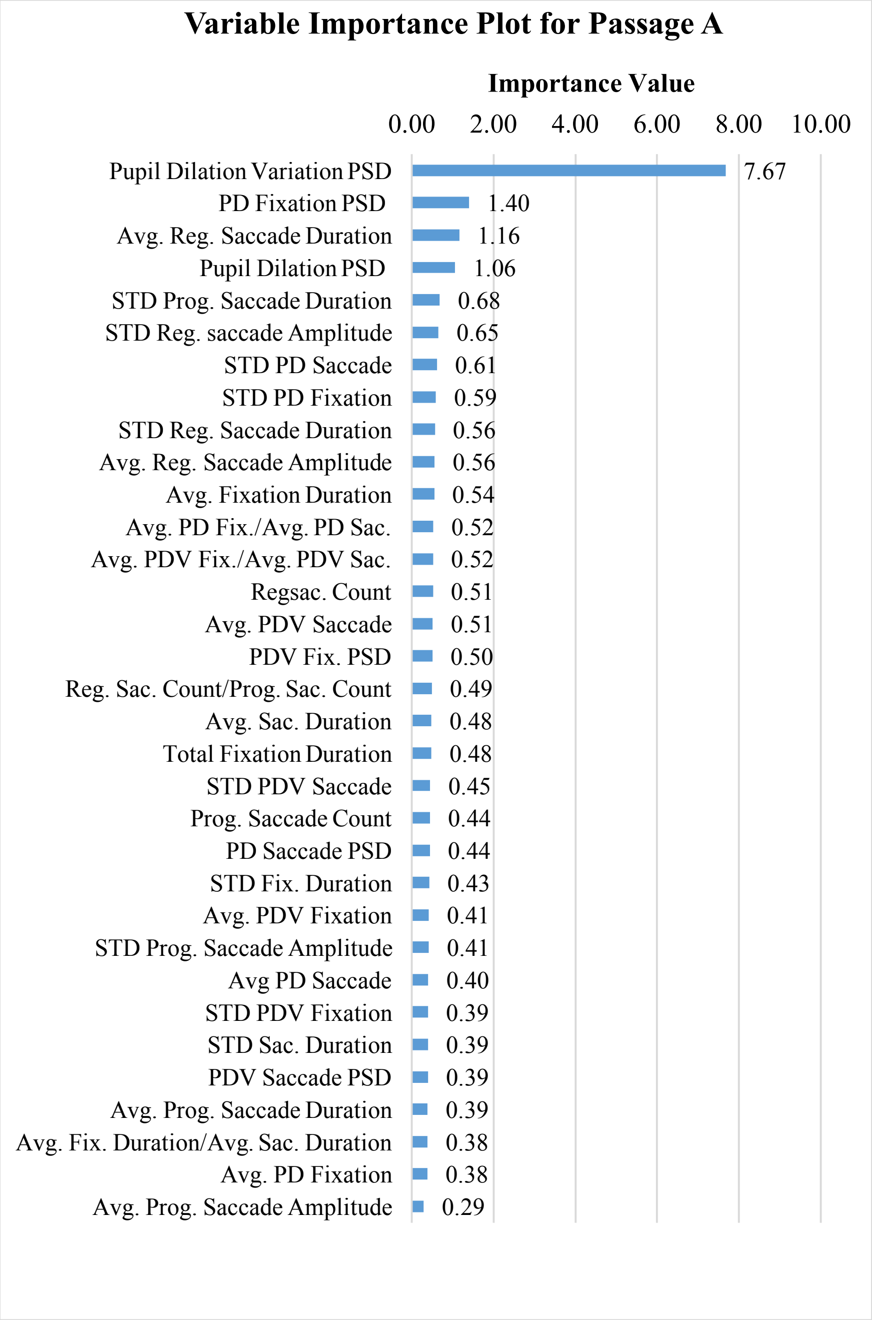


Figure 6. 3 Variable Importance Plot for Passage A (Cognitive Load Classification Task)

Using a similar systematic approach proposed in Chapter 3, four features with the highest variable importance were chosen to build the OA-SA machine learning model (Table 6.10).

Table 6. 10 Selected Features for Cognitive Load Detection Model Using Passage A

1	Pupil Dilation Variation PSD
2	Pupil Dilation – Fixation PSD
3	Avg. Regressive Saccade Duration
4	Pupil Dilation PSD

Figure 6.4 illustrates the performance of the OA-SA classifier built from the selected features (Table 6.10) against the number of trees. It can be observed that the performance of the classifier is about 79% with about 30 number of trees.

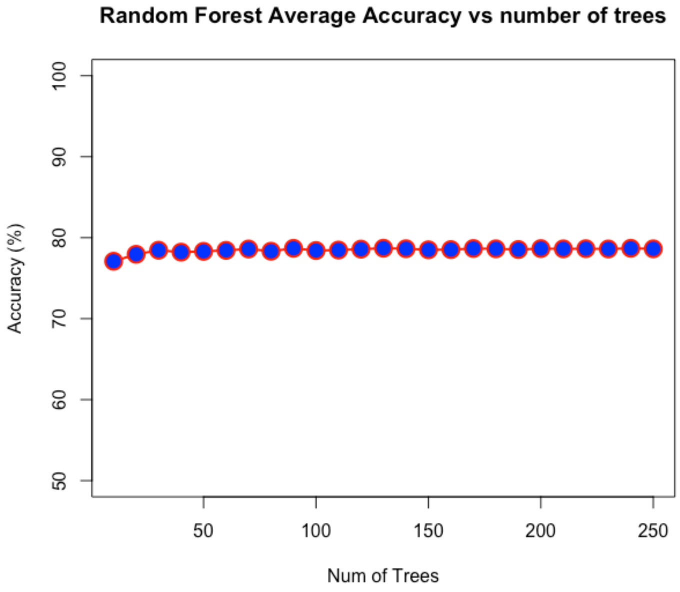


Figure 6. 4 Random forest average accuracy vs number of trees for cognitive load detection model using passage A

RF variable importance graph (Figure 6.5) in the second classification task, where dataset from readers of OB and SB were fed to the classifier, shows that average progressive saccade duration is the most discriminative feature (importance value of 8.10) in classifying the eye movement data of participants who read OB and SB passages.

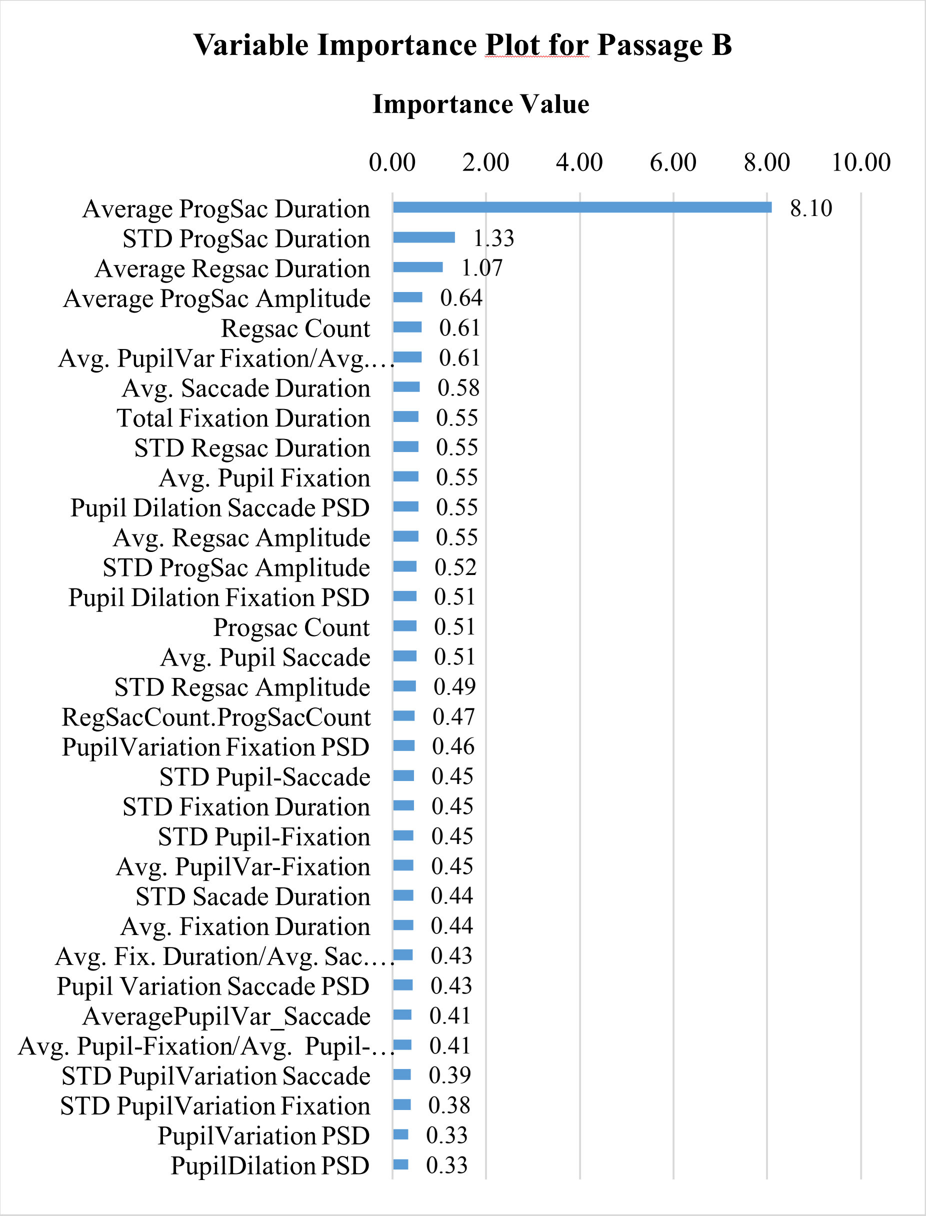


Figure 6. 5 Variable Importance Plot for Passage B (Cognitive Load Classification Task 2)

Three features with the highest variable importance were systematically chosen to build the second (OB-SB) machine learning model (Chapter 3). Table 6.11 shows the 3 features selected for the classification task.

Table 6. 11 Selected Features for Cognitive Load Detection Model Using Passage B

1	Avg. Progressive Saccade Duration

2	STD Progressive Saccade Duration
3	Avg. Regressive Saccade Duration

Figure 6.6 shows the performance of the OB-SB classifier built from the selected features (Table 6.11) vs. the number of tree. One can observe that the performance of the classifier is about 79% with about 20 number of trees.

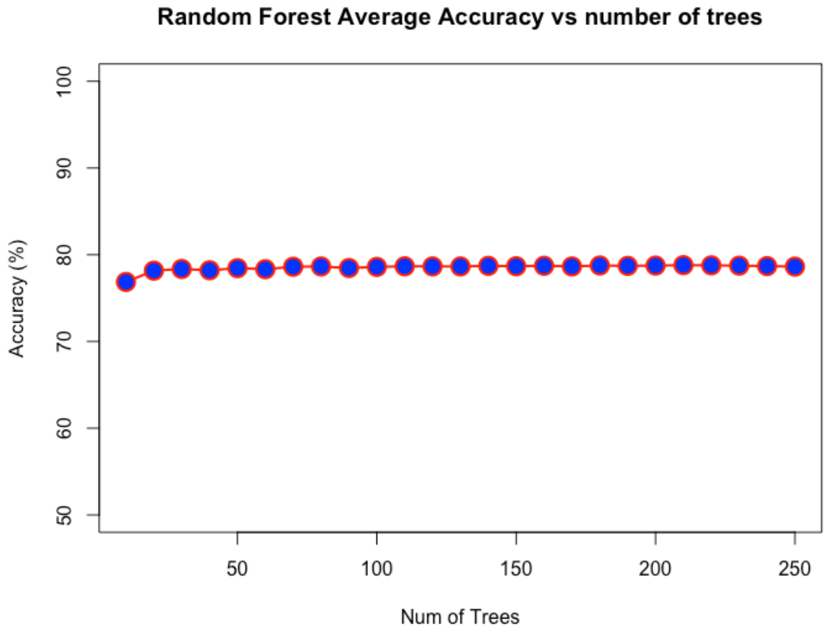


Figure 6. 6 Random forest average accuracy vs number of trees for cognitive load detection model using passage B

RF variable importance graph in the third integrated classification task (Figure 6.8), where dataset from readers of both original passages (OA, OB) and readers of both simplified passages (SA, SB) were combined and used to train the classifier. The results in figure 6.7 show that area under the curve of PSD of pupil dilation signals and average progressive saccade duration are the most discriminative features (importance values of 8.33 and 6.06) in classifying the eye movement data of participants who read original passages vs. simplified passages.

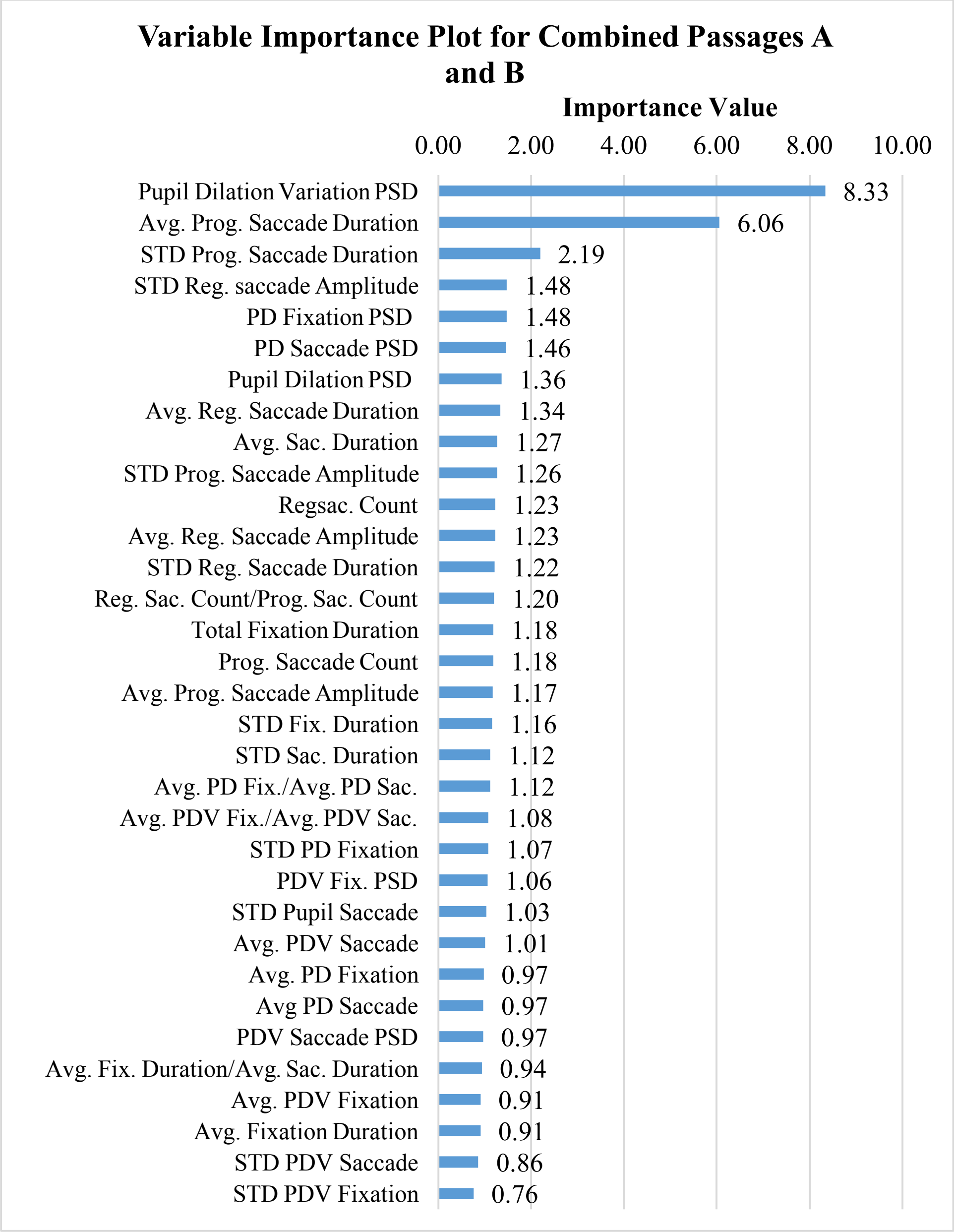


Figure 6. 7 Variable Importance Plot (Cognitive Load Classification Task 3)

Five features with the highest importance values were systematically chosen to build the integrated (OA-OB_SA-SB) machine learning model. Table 6.12 shows the five features selected for the classification task.

Table 6. 12 Selected Features for Cognitive Load Detection Model Using Passages A and B combined

1	Pupil Dilation Variation PSD
2	Avg. Progressive Saccade Duration
3	STD Progressive Saccade Duration
4	Avg. Regressive Saccade Amplitude
5	Pupil Dilation-Fixation PSD

The performance of the integrated machine learning model of cognitive load is shown in Figure 6.8. One can observe that the classifier has about 70% accuracy with more than 50 trees.

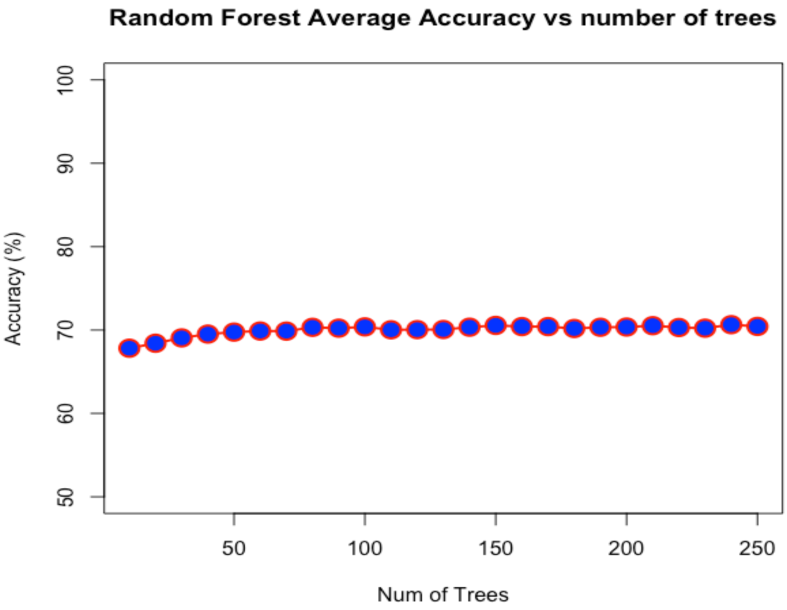


Figure 6. 8 Random forest average accuracy vs number of trees for cognitive load detection model using passages A and B combined

To investigate consistency and generalizability of feature sets in different scenarios I used the above selected features from OA-SA, OB-SB and integrated OA-SA_OB-SB models and built new models. The objective is, for example, if I use features from passage A (Table 6.10) and build a model using Passage B's data how it affects the performance of classification. Table

6.13 summarizes all the performances. It can be observed that the selected feature set in integrated machine learning model (OA-OB_SA-SB) is the most robust feature set in classification, since it results in a more consistent performance when applied to different target datasets (Table 6.13). For example, when using selected features of OA-OB_SA-SB (from Table 6.12) model to classify users who read OA and SA passages, it results in classification performance of around 72% while using the selected features of OB-SB model for the same classification task results in 54% accuracy. Similarly, when using the selected features of OA-OB_SA-SB model to classify the users who read OB and SB passages the accuracy is 78%, while using the selected features of OA-SA for the same classification task results in an accuracy of 50%. These results suggest that using eye movement features of the combined passages data results in higher and more consistent performance when used to classify different passages. Therefore, one can conclude that using more number of passages may result in enhancing the classification performance and generalized eye-movement feature set.

Table 6. 13 Performance of Cognitive Load Classifiers

		Features Selected from		
		OA_SA	**OB_SB**	**OAOB_SASB**
Target Classes	**OA vs. SA**	79.45%	54.60%	**72.49 %**
	OB vs. SB	50%	79%	**78%**
	OAOB vs. SASB	62.50%	57.16%	**70%**

Average and standard deviation of all the eye movement features used in building the feature matrix for cognitive load classification task are given in appendix A, in Tables A.1 and A.2.

6.4.2 Age Classification Results

Overall 78 young and 52 old subjects participated in the study, each read a pair of passages (OA/SB) or (OB/SA). As a result, 158 eye-movement recordings were collected from younger participants and 104 data from older subjects. Table 6.14 shows how the passages distributed among younger and older classes.

Table 6. 14 Age Classification Settings

Classes	$Class\ 1: Younger\ (156) \rightarrow \begin{cases} OA\ (41) \\ OB\ (37) \\ SA\ (37) \\ SB\ (41) \end{cases}$ $Class\ 2: Older\ (104) \rightarrow \begin{cases} OA\ (24) \\ OB\ (28) \\ SA\ (28) \\ SB\ (24) \end{cases}$
Feature Matrix Dimension	33 by 260

A) Passages Subjective Rating and Performance Results

I used two-sample t-test to verify whether the subjective SMEQ ratings of original and simplified versions of each passage were significantly different among the two age groups. Table 6.15 shows the average, STD and the results of t-test for SMEQ questions for passage A and Table 6.16 shows the same values for passage B. As the results show there is no significant differences in the rating of passages between older and younger participants for passage A. For the original version of passage B (OB), older adults' rating of difficulty of the passage was significantly lower than younger users' rating (Table 6.16 - 36.43 vs 58.57). These results suggest that except for the original passage B, there was no significant differences between perceived mental effort of older and younger adults in reading the passages.

The comparison between older and younger adults' performance results for passage A and passage B are shown in Tables 6.17 and 6.18. The results show that performance among the two age groups was not significantly different for passage A, while for passage B, older adults performed significantly better compared to younger users, for both conditions of text. The results of readability scores for original passages A and B (16.1 and 16.4 respectively) showed that passage B was harder than passage A. This was the case even after simplification (reading score 10 for simplified A vs. 11 for simplified B). The results show that older people performed better than the younger people for the harder text. This difference facilitates the possibility to test the robustness of eye movements in predicting age regardless of differences between the two groups in perceived difficulty of the task and/or differences in performance.

Table 6. 15 Subjective SMEQ Rating for Passage A (younger vs Older Adults)

	OA		SA	
	younger	older	younger	older
Mean	64.07	63.04	47.56	45.36
STD	24.94	25.86	25.51	28.51
p-value	0.87		0.74	

Table 6. 16 Subjective SMEQ Rating for Passage B (younger vs Older Adults)

	OB		SB	
	younger	older	younger	older
Mean	58.57	36.43	36.98	30.00
STD	32.16	27.28	25.16	26.27
p-value	**0.005**		0.29	

Table 6. 17 Performance – Passage A (Younger vs Older Users)

	OA		SA	
	younger	older	younger	older
Mean	1.00	1.00	0.89	1.25
STD	0.77	0.98	0.96	0.93
p-value	1		0.16	

Table 6. 18 Performance – Passage B (Younger vs Older Users)

	OB		SB	
	younger	older	younger	older
Mean	0.62	1.14	0.68	1.08
STD	0.76	0.84	0.61	0.88
p-value	**0.01**		**0.03**	

B) *Age Classification Results*

To build the age classification model, feature selection and classification processes were run on a feature matrix with original 33 features and two classes of 260 younger and older variations.

Figures 6.9 and 6.10, and Table 6.19 demonstrates the results correspond to age classification model. RF variable importance graph (Figure 6.9) shows that standard deviation of regressive saccade duration is the most discriminative feature in classifying older and young users.

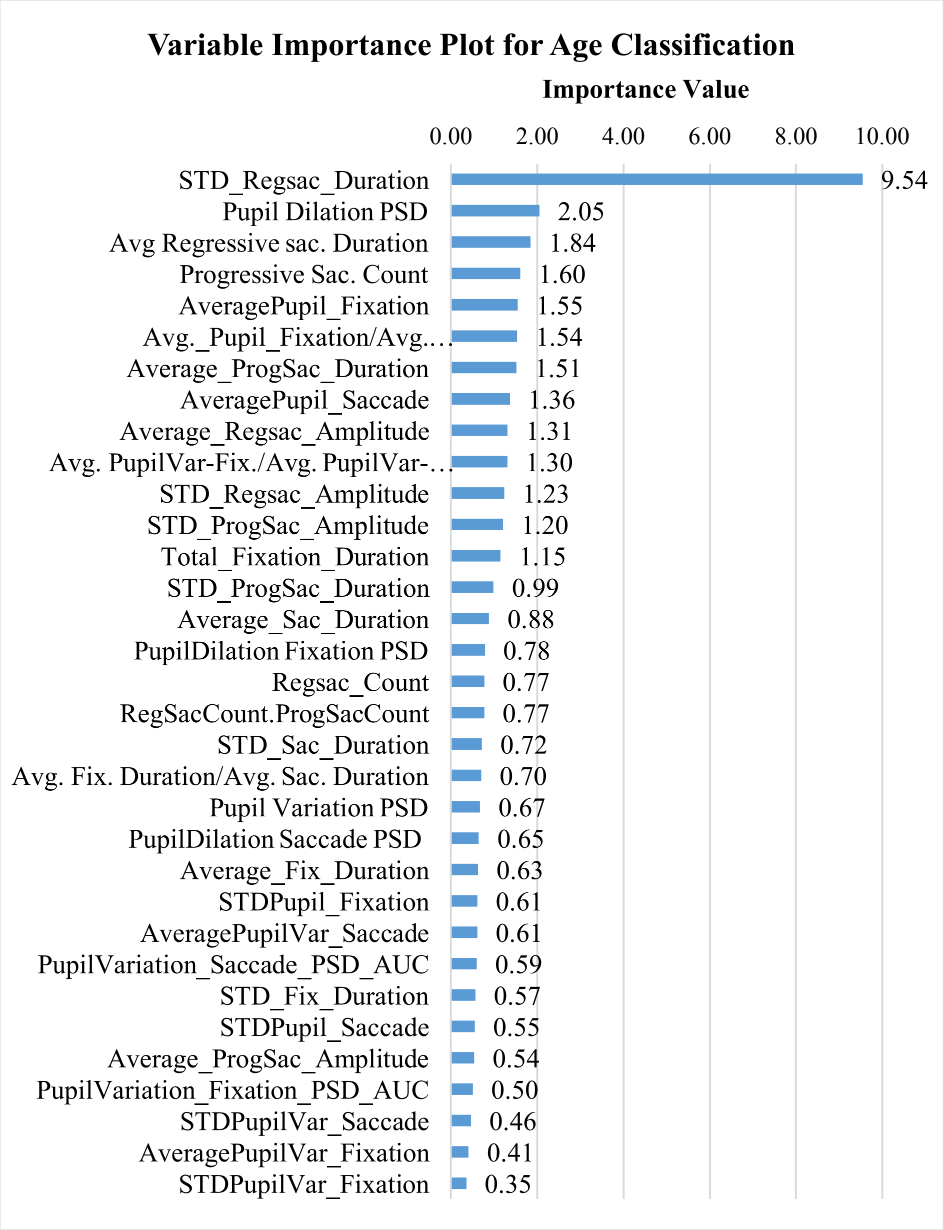


Figure 6. 9 Variable importance graph (Age Classifier)

Six features with the highest variable importance were chosen to build (Young/Old) machine learning model (Table 6.19).

Table 6. 19 Selected Features for Age Classification Model

1	STD Regressive Saccade Duration
2	Pupil Dilation PSD
3	Avg. Regressive Saccade Duration
4	Progressive Saccade Count
5	Avg. Pupil Dilation -Fixation
6	Avg. PD-Fixation/Avg. Pupil-Saccade

The performance of the age machine learning model is shown in Figure 6.11. It can be observed that the classifier has above 82% accuracy with about 60 trees.

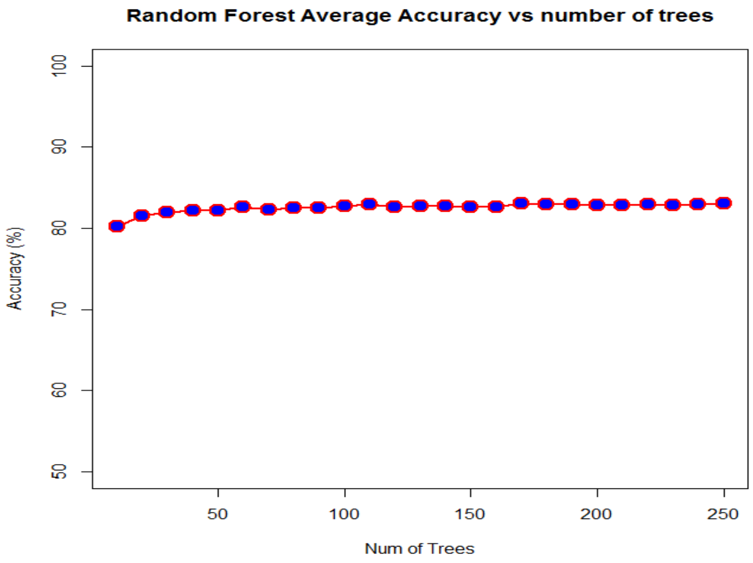


Figure 6. 10 Random Forest Average Accuracy vs. Number of Trees for Age Detection Model

6.5 Discussion

Objective user profiling has been of interest of UX and HCI researchers since recent decade. Having access to user's characteristics and mental effort during a task helps the task/interface developers to adaptively modify the content of the interface or task conditions to improve the user experience. The main goal of the present study was to obtain a better understanding of two major characteristics of user profiling (e.g. cognitive load and age) of users when reading a passage on computer screen. The results of the user profiling study suggest that eye-tracking technology when accompanied by machine learning models can be used to distinguish user's extraneous cognitive load level and age population (e.g. younger vs. older adults).

Results of cognitive load classification models indicate that frequency domain features of pupil dilation and pupil dilation variation along with duration of regressive and progressive saccades are the most important eye-movement features in predicting the task condition or user's level of cognitive load when reading passages with different levels of difficulty. From the age classification task, it is concluded that frequency domain feature of pupil dilation as well as regressive saccadic duration are the most important features to distinguish the age population of a user during a reading task. Time domain features, such as the ratio of pupil dilation during fixation to pupil dilation during saccade, were only found in the age classification task as one of the important features. Prior eye tracking/reading literature have used length, duration, and number of regressive and progressive saccades as well as number and duration of fixations to compare the reading behavior of older and younger adults (e.g., Coyne et al., 2002, Rayner et al., 2006) and cognitive load during reading (e.g., Rayner 2009, Rayner et al., 2012, Campbell et al., 2014). To the best of my knowledge, however, pupillary data have not been used in previous research. While the results of this study showed that power spectral density of pupillary data are among the most important features in classification of younger and older adult's eye movements as well as the level of cognitive load. To best of my knowledge, the frequency domain features of pupil dilation have not been used by researchers for cognitive load assessment. Therefore, this study contributes to the related literature by using frequency domain features. Using frequency domain features of eye movements facilitates a new direction for eye tracking and reading research.

7 Conclusion & Future Work

Advances in technology make it possible to embed eye-tracking technology in computing devices at affordable prices. The data produced holds a wealth of information to improve the understanding of user behavior and decision making. The advent of robust machine learning approaches provides an attractive opportunity to capitalize on this information. Hence, designing machine learning predictive models using eye tracking is very likely to continue as a productive line of research and development. Using eye movements to detect reactions to task demands is an important first step in designing interfaces that can more effectively respond to user needs based on their age and the level of cognitive effort they experience.

Cognitive effort affects how people make decisions, including the potential adoption of a system, and its effective use (Gregor & Benbasat, 1999; Payne et al., 1993). Detecting the cognitive effort that a task demands can improve the design of adaptive user interfaces that can respond to user needs in real time. A first step in designing such responsive computerized tool is to build an advanced system that can detect cognitive load unobtrusively and automatically.

In the first study of this dissertation (Chapter 3), grounded in the Adaptive Decision Making and eye tracking literature, I argued that task demand can be detected unobtrusively and automatically via eye movement data. I developed an eye tracking machine learning model to test this assertion. The results showed that eye movements indeed carry information about cognitive load and that pupil data, in particular the ratio of pupil dilation during saccades and fixations, was the most important predictor factor for task demand in a math problem solving task under time limit. Another novel finding of this study was that pupillary data when discriminated during fixation and saccade can provide invaluable insight about the relation between cognitive load and eye movements. Additionally, results of the study showed that the random forest machine learning system can predict task demand with approximately 70% accuracy. These results show that building such an advanced system is possible and is computationally practical.

Being able to predict cognitive load of a user via eye movements and machine learning model in a math problem solving task, the next step was to assess cognitive load in a different task, such as reading. The research question was whether differences in task demand could be detected via pupillary data during a reading task. To address this research question, I designed study II (Chapter 4), and investigated whether reducing cognitive load of readers by simplifying text passages can affect their pupillary data during reading and whether this effect remained steady over different time intervals of reading. Time series analyses of pupillary data were performed and then were compared among the two groups of participants, one with higher task load (reading original text), and another one with lower task load (reading a simplified version of text). The results showed that pupil dilation and pupil dilation variation were significantly different among two different task condition. Further, the results provided evidence that examining pupillary data in various time intervals can provide additional information for understanding cognitive load. Time series analysis of eye-tracking data is important because it provides a continuous measure of eye-movement data, which allows to examine moment by moment analysis of eye-movement data.

The objective of the last study of the dissertation was to develop a user profiling framework for reading tasks based on two machine learning (ML) models. The proposed ML based profiling process consists of user's age characterization and user's cognitive load detection. To this end, detection of cognitive load through eye-movement features was investigated during reading with different task conditions. Furthermore, relationship between user's eye-movements and their age population (e.g. younger and older adults) were carried out during the reading task. Tobii X300 eye tracking device was used in all the above mentioned studies to record the eye movement data from participants. Eye-movements were acquired from Tobii eye tracking software, and then were preprocessed and analyzed in R. Random forest classifier with bootstrapping was used to build machine learning models. The aggregated results of the studies indicate that machine learning once accompanied with a NeuroIS tool, like eye-tracking can be used to model user characteristics like age, and user mental states like cognitive load, automatically and implicitly with accuracy above chance (range of 70-92%).

The results of this dissertation can be used in a more general framework to adaptively modifying textual information to better serve the users mental and age needs.

Future Work

This study can be enhanced to provide more robust and generalized machine learning models by doing the following enhancements:

- Increased number of passages from a wide variety of topics (e.g., politics, history, sport, health, etc.).
- Increased sample size with different demographics (e.g., people with different educational level or reading habits could be controlled for). This will likely improve the robustness of machine learning models.
- More investigation in frequency domain as it has been done for other physiological signals (e.g., heart rate) would provide a better insight into how different frequency bands of eye-tracking signals might respond to different level of cognitive load or user's characteristics.
- Age characteristic can be expanded into more categories other than what was investigated in this study (18-30 vs. 50-69). This would increase the personalization level and make the model more specific to user's age group needs.
- Other user characteristics such as proficiency level (e.g., native vs. non-native English readers) can be also modeled via a separate classification task. This would be beneficial since the ultimate goal of the user personalization task is to characterize users objectively via their biometric and cognitive behaviors.
- Using other NeuroIS technologies (e.g., EEG brain activity) besides eye-tracking would provide more reliable and richer bio-feedback that might increase the accuracy of the machine learning models.

The above scenarios are suggested for the future studies that focus on the enhancement of user profiling models based on user's needs and characteristics.

References

Aarts, H., Bijleveld, E., Custers, R., Dogge, M., Deelder, M., Schutter, D., & Haren, N. E. M. (2012). Positive priming and intentional binding: Eye-blink rate predicts reward information effects on the sense of agency. *Social Neuroscience*, *7*(1), 105–112. https://doi.org/10.1080/17470919.2011.590602

Abel, L. A., & Douglas, J. (2007). Effects of age on latency and error generation in internally mediated saccades. *Neurobiology of Aging*, *28*(4), 627–637. https://doi.org/https://doi.org/10.1016/j.neurobiolaging.2006.02.003

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623–636. https://doi.org/https://doi.org/10.1016/j.ergon.2006.04.002

Albers, M. J. (2011). Tapping as a measure of cognitive load and website usability. *Proceedings of the 29th ACM International Conference on Design of Communication*. Pisa, Italy: ACM. https://doi.org/10.1145/2038476.2038481

Andrzejewska, M. ., & Stolińska, A. (2016). Comparing the Difficulty of Tasks Using Eye Tracking Combined with Subjective and Behavioural Criteri. *Journal of Eye Movement Research*, *9*(3), 1–16.

Arch, A., Abou-Zahra, S., & Henry, S. L. (2009). older-users-online. Retrieved from https://www.w3.org/WAI/posts/2009/older-users-online

Arthur D. Fisk, Sara J. Czaja, Wendy A. Rogers, Neil Charness, J. S. (2012). Books @ Books.Google.Com. Retrieved from

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *58*(6), 1065–1086. https://doi.org/10.1080/02724980443000476

Attar, N. (2016). *Enhancing Cognitive Load Measurement and User Performance in Human-Computer Interaction*. *Computer Science*. UMass Boston.

Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: technological antecedents and implications. *MIS Q.*, *35*(4), 831–858.

Baby Boomers and Credit generational. (n.d.).

Bailey, B., & Iqbal, S. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Trans. Comput.-Hum. Interact.*, *14*(4), 1–28. https://doi.org/citeulike-article-id:4679115 doi: 10.1145/1314683.1314689

Barkhi, R., Rolland, E., Butler, J., & Fan, W. (2005). Decision Support System induced guidance for model formulation and solution. *Decision Support Systems*, *40*(2), 269–281. https://doi.org/https://doi.org/10.1016/j.dss.2003.12.006

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bednarik, R., Kinnunen, T., Mihaila, A., & Fränti, P. (2005). Eye-Movements as a Biometric BT - Image Analysis. In H. Kalviainen, J. Parkkinen, & A. Kaarna (Eds.) (pp. 780–789). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bednarik, R., Vrzakova, H., & Hradis, M. (2012). What do you want to do next: a novel approach for intent prediction in gaze-based interaction. *Proceedings of the Symposium on Eye Tracking Research and Applications*. Santa Barbara, California: ACM. https://doi.org/10.1145/2168556.2168569

Beran, R. (1992). Introduction to Efron (1979) Bootstrap Methods: Another Look at the Jackknife. In S. Kotz & N. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 565–568). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_40

Bergstrom, J. R., & Schall, A. (2014). *Eye Tracking in User Experience Design*. Morgan Kaufmann Publishers Inc.

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction* (Vol. 140). John Wiley & Sons, Inc. https://doi.org/10.2307/2344882

Boechler, P., Foth, D., & Watchorn, R. (2006). The Influence of Reading and Memory Skills on Older Adults' Information Search and Learning in Educational Hypermedia. (T. Reeves & S. Yamashita, Eds.), *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006*. Honolulu, Hawaii, USA: Association for the Advancement of Computing in Education (AACE). Retrieved from https://www.learntechlib.org/p/23984

Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, *14*(3), 29. Retrieved from http://dx.doi.org/10.1167/14.3.29

Brandtzæg, P. B., Lüders, M., & Skjetne, J. H. (2010). Too Many Facebook "Friends"? Content Sharing and Sociability Versus the Need for Privacy in Social Network Sites. *International Journal of Human–Computer Interaction*, *26*(11–12), 1006–1030. https://doi.org/10.1080/10447318.2010.516719

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brian, A., & Ripley, M. B. (2018). Package " tree ."

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist*, *38*(1), 53–61. https://doi.org/10.1207/S15326985EP3801_7

Brünken J. L.; Moreno, R., R. . P. (2010). Current issues and open questions in cognitive load research. *Cognitive Load Theory*, 253–272.

Buettner, R. (2014). Analyzing Mental Workload States on the Basis of the Pupillary Hippus. In *NeuroIS* (p. 52).

Buettner, R. ., Sauer, S. ., Maier, C. ., & Eckhardt, A. (2015). Towards Ex Ante Prediction of User Performance: A Novel NeuroIS Methodology Based on Real-Time Measurement of Mental Effort. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 533–542). https://doi.org/10.1109/HICSS.2015.70

Buscher, G., Dengel, a., & van Elst, L. (2008). Eye movements as implicit relevance feedback. *CHI'08 Extended Abstracts on Human Factors in Computing Systems, ACM*, 2991–2996. https://doi.org/10.1145/1358628.1358796

Campbell, R. J., & Bovee, J. C, G. E. (2014). Using Eye Movements to Evaluate the Cognitive Processes Involved in Text Comprehension. *Journal of Visualized Experiments*, *83*, e50780.

Capozzo, D., Groezinger, R. L. ., Ng, K.-F. F. ., & Siegel, M. J. . (2008). Appeal of Web Page Layout and Characteristics Based on Age: Usability Research through Eye Tracking at Fidelity Investments Inc.

Chadwick-Dias, A., McNulty, M., & Tullis, T. (2003). Web usability and age: how design changes can improve performance. *Proceedings of the 2003 Conference on Universal Usability*. Vancouver, British Columbia, Canada: ACM. https://doi.org/10.1145/957205.957212

Chadwick-Dias, A., Tedesco, D., & Tullis, T. (2004). Older adults and web usability. *Extended Abstracts of the 2004 Conference on Human Factors and Computing Systems - CHI '04*, 1391. https://doi.org/10.1145/985921.986072

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye Activity As a Measure of Human Mental Effort in HCI. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (pp. 315–318). New York, NY, USA: ACM.

Choi, H.-H., van Merriënboer, J. J. G., & Paas, F. (2014). Effects of the Physical Environment on Cognitive Load and Learning: Towards a New Model of Cognitive Load. *Educational Psychology Review*, *26*(2 LB-Choi2014), 225–244. https://doi.org/10.1007/s10648-014-9262-6

Clark;, R. E., & Clark, V. P. (2010). *From Neo-behaviorism to Neuroscience: Perspectives on the Origins and Future Contributions of Cognitive Load Research*. New York.

Cooley, J. W., & Tukey, J. W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, *19*(90), 297–301. https://doi.org/10.2307/2003354

Coyne, K. P. and J. N. (2002). *Web Usability for Older People*. Retrieved from https://www.nngroup.com/articles/usability-for-senior-citizens/

Craik, F. I. M., & Salthouse, T. A. (Eds.). (2000). The handbook of aging and cognition, 2nd ed. *The Handbook of Aging and Cognition, 2nd Ed.*, ix, 755-ix, 755.

Cutler, A. (2014). *Random Forests*. (N. B. T. C. B. E. W. P. F. R. J. L. Teugels, Ed.). Wiley StatsRef:Statistics Reference Online. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06520

Cyr, D., Head, M., Larios, H., & Pan, B. (2009). Exploring Human Images in Website Design: A Multi-Method Approach. *MIS Quarterly*, *33*(3), 539–566. https://doi.org/10.2307/20650308

Czaja, S., & Lee, C. C. (2007). The Human-Computer Interaction Handbook Fundamentals, Evolving Technologies and Emerging Applications, Second Edition Information Technology and Older Adults. *The Human-Computer Interaction Handbook*, 777–792. https://doi.org/10.1201/9781410615862.ch39

Dähne, S., Wilbert, N., & Wiskott, L. (2014). Slow Feature Analysis on Retinal Waves Leads to V1 Complex Cells. *PLoS Computational Biology*, *10*(5), e1003564. https://doi.org/10.1371/journal.pcbi.1003564

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

De Luca, M. ., Borrelli, M. ., Judica, A. ., Spinelli, D. ., & Zoccolotti, P. (2002). Reading Words and Pseudowords: An Eye Movement Study of Developmental Dyslexia. *Brain and Language*, *80*(3), 617–626. https://doi.org/https://doi.org/10.1006/brln.2001.2637

Dimoka, A. (2012). How to conduct a Functional Magnetic Resonance (fMRI) study in Social Science research. *MIS Quarterly*, *36*(3), 811–840.

Djamasbi, S. (2014). Eye Tracking and Web Experience . *AIS Transactions on Human-Computer Interaction*, *6*, 37–54.

Djamasbi, S., Rochford, J., DaBoll-Lavoie, A., Greff, T., Lally, J., & McAvoy, K. (2016). Text Simplification and User Experience BT - Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience. In D. D. Schmorrow & C. M. Fidopiastis (Eds.) (pp. 285–295). Cham: Springer International Publishing.

Djamasbi, S., Shojaeizadeh, M., Chen, P., Rochford, J., Chen, P., & Rochford, J. (2016). Text Simplification and Generation Y : An Eye Tracking Study Text Simplification and Generation Y : An Eye Tracking.

Djamasbi, S., Siegel, M., Skorinko, J., & Tullis, T. (2011). *Online Viewing and Aesthetic Preferences of Generation Y and the Baby Boom Generation: Testing User Web Site Experience Through Eye Tracking*. *International Journal of Electronic Commerce* (Vol. 15). https://doi.org/10.2753/JEC1086-4415150404

Djamasbi, S., Siegel, M., & Tullis, T. (2010). Generation Y, web design, and eye tracking. *International Journal of Human Computer Studies*, *68*(5), 307–323. https://doi.org/10.1016/j.ijhcs.2009.12.006

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Doubé, W., & Beh, J. (2012). Typing over autocomplete: cognitive load in website use by older adults. *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI 2012)*, 97–106. https://doi.org/10.1145/2414536.2414553

Duric, Z., Gray, W. D., Heishman, R., Li, F., Rosenfeld, A., Schoelles, M. J., … Wechsler, H. (2002). Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE*, *90*(7), 1272–1289. https://doi.org/10.1109/JPROC.2002.801449

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife, 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Taylor & Francis. Retrieved from https://books.google.com/books?id=gLlpIUxRntoC

Einhäuser, W., Stout, J., Koch, C., & Carter, O. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(5), 1704–1709. https://doi.org/10.1073/pnas.0707727105

Eivazi, S., & Bednarik, R. (2011). Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data. In the proceedings of 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI 2011.

Eye tracker accuracy and precision. (n.d.). Retrieved from https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-affects-the-accuracy-and-precision-of-an-eye-tracker/

Faraday, P. (2000). Visually Critiquing Web Pages. In T. Chambel & G. Davenport (Eds.) (pp. 155–166). Vienna LB - 10.1007/978-3-7091-6771-7_17: Springer Vienna.

Fehrenbacher, D. D. ., & Djamasbi, S. (2017). Information systems and task demand: An exploratory pupillometry study of computerized decision making. *Decision Support Systems*, *97*, 1–11. https://doi.org/10.1016/j.dss.2017.02.007

Felthousen, M. (2008). The trail less traveled. *Proceedings of the 36th Annual ACM SIGUCCS Conference on User Services Conference - SIGUCCS '08*, 83. https://doi.org/10.1145/1449956.1449984

Ferrari, J. R. (2001). Procrastination as self-regulation failure of performance: effects of cognitive load, self-awareness, and time limits on "working best under pressure." *European Journal of Personality*, *15*(5), 391–406. https://doi.org/10.1002/per.413

Fisk, A. ., Rogers, W. A. ., Charness, N. ., Czaja, S. J. ., & Sharit, J. (2012). *Designing for Older Adults: Principles and Creative Human Factors Approaches*.

Fukuda, R., & Bubb, H. (2003). Eye tracking study on web-use: Comparison between younger and elderly users in case of search task with electronic timetable service. *PsychNology Journal*, *1*(3), 202–228.

Goldberg, J. H. ., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645.

Grahame, M., Laberge, J., & Scialfa, C. T. (2004). Age Differences in Search of Web Pages: The Effects of Link Size, Link Number, and Clutter. *Human Factors*, *46*(3), 385–398. https://doi.org/10.1518/hfes.46.3.385.50404

Gregor, S., & Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, *23*(4), 497–530. https://doi.org/10.2307/249487

Gupta, A., Li, H., & Sharda, R. (2013). Should I send this message? Understanding the impact of interruptions, social hierarchy and perceived task complexity on user performance and perceived workload. *Decision Support Systems*, *55*(1), 135–145. https://doi.org/https://doi.org/10.1016/j.dss.2012.12.035

Gustavsson, C. J. (2010). Real Time Classification of Reading in Gaze Data, 1–42.

Hans, A., Borchers, W., & Borchers, M. H. W. (2018). Package pracma.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/https://doi.org/10.1016/S0166-4115(08)62386-9

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer. https://doi.org/citeulike-article-id:7765923

He, J., & McCarley, J. S. (2010). Executive working memory load does not compromise perceptual processing during visual search: Evidence from additive factors analysis. *Attention, Perception, & Psychophysics*, *72*(2), 308–316. https://doi.org/10.3758/APP.72.2.308

Henderson, J. M., Shinkareva, S. V, Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting Cognitive State from Eye Movements. *PLOS ONE*, *8*(5), e64937. Retrieved from https://doi.org/10.1371/journal.pone.0064937

Hertzum, M., & Hornbæk, K. (2010). How Age Affects Pointing With Mouse and Touchpad: A Comparison of Young, Adult, and Elderly Users. *International Journal of Human–Computer Interaction*, *26*(7), 703–734. https://doi.org/10.1080/10447318.2010.487198

Hess, E. H. ., & Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science*, *143*(3611), 1190–1192. Retrieved from http://www.jstor.org/stable/1712692

Hess, T. J., Rees, L. P. ;, & Rakes, T. R. (2005). Using Autonomous Planning Agents to Provide Model-based Decision-making Support. *Journal of Decision Systems*, *14*(3), 261–278. https://doi.org/10.3166/jds.14.261-278

Hill, R. L. ., Dickinson, A., Arnott, J. L. ., Gregor, P., & McIver, L. (2011). Older web users' eye movements: experience counts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1151–1160). ACM. https://doi.org/10.1145/1978942.1979115

Hogan, M., Torres, A., & Barry, C. (2015). An eye tracking pilot study of optional decision constructs in B2C transactional processes. *14th International Conference on WWW/INTERNET (ICWI)*.

Holmqvist, K. ., Nystrom, M. ., Anderson, R. ., Dewhurst, R. ., Jarodzka, H. ., & & Van de Weijer, J. (2011). *Eye tracking: A Comprehensive Guide to Methods and Measures*. OUP Oxford.

Ikehara, C. S., Crosby, M. E., & Silva, P. A. (2013). Combining Augmented Cognition and Gamification. In C. M. Fidopiastis (Ed.) (pp. 676–684). Berlin, Heidelberg LB - 10.1007/978-3-642-39454-6_72: Springer Berlin Heidelberg.

Iqbal, S., Adamczyk, P., Zheng, X., & Bailey, B. (2005). Towards an index of opportunity: understanding changes in mental workload during task execution. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 311–320). New York, NY, USA: ACM. https://doi.org/citeulike-article-id:311608 doi: 10.1145/1054972.1055016

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. *CHI '04 Extended Abstracts on Human Factors in Computing Systems*. Vienna, Austria: ACM. https://doi.org/10.1145/985921.986094

Jacob, R. J. K. ., & Karn, K. S. (2003). Commentary on Section 4 - Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises A2 - Hyönä, J. In R. Radach & H. Deubel (Eds.), *The Mind's Eye* (pp. 573–605). Amsterdam: North-Holland. https://doi.org/https://doi.org/10.1016/B978-044451020-4/50031-1

James, A. G., Witten, D., Hastie, T., Tibshirani, R., & Hastie, M. T. (2018). Package " ISLR ."

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer New York. Retrieved from https://books.google.com/books?id=at1bmAEACAAJ

Josephson, S., & Holmes, M. (2004). Age differences in visual search for information on web pages. In *Proceedings of the 2004 symposium on Eye …* (p. 62). https://doi.org/10.1145/968363.968379

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. https://doi.org/10.1016/0010-0285(76)90015-3

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2003). How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 447–457. https://doi.org/10.1037/0278-7393.29.3.447

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583–1585.

Kalyuga, S. (2011). Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, *23*(1 LB-Kalyuga2011), 1–19. https://doi.org/10.1007/s10648-010-9150-7

Kardan, S., & Conati, C. (2012). Exploring Gaze Data for Determining User Learning with an Interactive Simulation BT - User Modeling, Adaptation, and Personalization. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.) (pp. 126–138). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kemper, S., Crow, A., & Kemtes, K. (2004). Eye-Fixation Patterns of High- and Low-Span Young and Older Adults: Down the Garden Path and Back Again. *Psychology and Aging*. Kemper, Susan: Gerontology Center, University of Kansas, 3090 Dole Building, 1000 Sunnyside, Lawrence, KS, US, 66045, skemper@ku.edu: American Psychological Association. https://doi.org/10.1037/0882-7974.19.1.157

Kemtes, K. A., & Kemper, S. (1997). Younger and older adults' on-line processing of syntactically ambiguous sentences. *Psychology and Aging*, *12*(2), 362–371. https://doi.org/10.1037/0882-7974.12.2.362

Kinnunen, T., Sedlak, F., & Bednarik, R. (2010). Towards Task-independent Person Authentication Using Eye Movement Signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research &#38; Applications* (pp. 187–190). New York, NY, USA: ACM. https://doi.org/10.1145/1743666.1743712

Klami, A., Saunders, C., de Campos, T. E., & Kaski, S. (2008). Can relevance of images be inferred from eye movements? *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval - MIR '08*, 134. https://doi.org/10.1145/1460096.1460120

Klami, A., Saunders, C., Te, #243, Campos, filo E. de, & Kaski, S. (2008). Can relevance of images be inferred from eye movements? *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*. Vancouver, British Columbia, Canada: ACM. https://doi.org/10.1145/1460096.1460120

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1–2), 262–284. https://doi.org/10.1080/09541440340000213

Klingner, J. (2010). Fixation-aligned pupillary response averaging. *Proceedings of the 2010*

Symposium on Eye-Tracking Research \&\#38; Applications. Austin, Texas: ACM. https://doi.org/10.1145/1743666.1743732

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker, *1*(212), 69–72. Retrieved from http://portal.acm.org/citation.cfm?doid=1344471.1344489

Korbach, A., Brünken, R., & Park, B. (2017). Measurement of cognitive load in multimedia learning: a comparison of different objective measures. *Instructional Science*, *45*(4 LB-Korbach2017), 515–536. https://doi.org/10.1007/s11251-017-9413-5

Król, M., & Król, M. E. (2017). A novel approach to studying strategic decisions with eye-tracking and machine learning. *Judgment and Decision Making*, *12*(6), 596–609.

Kruger, J.-L., Hefer, E., & Matthew, G. (2013a). Measuring the impact of subtitles on cognitive load. In *Proceedings of the 2013 Conference on Eye Tracking South Africa - ETSA '13* (Vol. 1, pp. 62–66). https://doi.org/10.1145/2509315.2509331

Kruger, J.-L., Hefer, E., & Matthew, G. (2013b). Measuring the Impact of Subtitles on Cognitive Load: Eye Tracking and Dynamic Audiovisual Texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa* (pp. 62–66). New York, NY, USA: ACM. https://doi.org/10.1145/2509315.2509331

Laeng, B. ., Sirois, S. ., & Gredebäck, G. (2012). Pupillometry:A Window to the Preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Ledger, H. (2013). The effect cognitive load has on eye blinking. *The Plymouth Student Scientist*, *6*(1), 206–223.

Levine, R., Locke, C., Searls, D., & Weinberger, D. (1999). *The Cluetrain Manifesot*. Basic Books. Retrieved from http://cluetrain.com/Cluetrain_10/index.html

Lin, T., & Imamiya, A. (2006). Evaluating usability based on multimodal information: an empirical study. *Proceedings of the 8th International Conference on Multimodal Interfaces*. Banff, Alberta, Canada: ACM. https://doi.org/10.1145/1180995.1181063

Liu, Y., Hsueh, P. Y., Lai, J., Sangin, M., Nussli, M. A., & Dillenbourg, P. (2009). Who is the expert? Analyzing gaze data to predict expertise level in collaborative applications. In *2009 IEEE International Conference on Multimedia and Expo* (pp. 898–901). https://doi.org/10.1109/ICME.2009.5202640

Locher, P. (2009). A Framework for Aesthetic Experience. *CHI 2009 Conference*, 9–12. https://doi.org/10.1162/DESI_a_00017</p>

Loos, E. F. ., & Romano Bergstrom, J. (2011). *Design and Development a Social Networks Platform for Older People* (Vol. pt.II). https://doi.org/10.1007/978-3-642-21663-3_20

Maintainer, M. S., & Seilmayer, M. (2016). Package "spectral" Common Methods of Spectral Data Analysis. Retrieved from https://cran.r-project.org/web/packages/spectral/spectral.pdf

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviat Space Environ Med*, *78*(5 Suppl), B165-75.

Max, A., Contributions, K., Weston, S., Keefer, C., Engelhardt, A., Cooper, T., … Hunt, T. (2018). Package " caret ."

McKinley, S., & Levine, M. (2002). Cubic Spline Interpolation. *Acta Mathematica Hungarica*, *107*(May), 493–507. https://doi.org/10.1007/s10474-005-0180-4

Meghanathan, R. N., van Leeuwen, C., & Nikolaev, A. R. . (2014). Fixation duration surpasses pupil size as a measure of memory load in free viewing. *Frontiers in Human Neuroscience*, *8*, 1063. https://doi.org/10.3389/fnhum.2014.01063

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2017). R Package e1071 Version 1.6-8. *Gpl-2*. Retrieved from https://cran.r-project.org/web/packages/e1071/e1071.pdf

Money, A. G., Fernando, S., Lines, L., & Elliman, A. D. (2010). Assessing online –form complexity for the development of assistive technologies for older adults. *Journal of Computing and Information Technology*, *18*(3), 257–274. Retrieved from http://eprints.kingston.ac.uk/19846/

Moreno, R. (2006). Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. *Journal of Computer Assisted Learning*, *22*(3), 149–158. https://doi.org/10.1111/j.1365-2729.2006.00170.x

Najar, A. S., Mitrovic, A., & Neshatian, K. (2014). Utilizing Eye Tracking to Improve Learning from Examples BT - Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge. In C. Stephanidis & M. Antona (Eds.) (pp. 410–418). Cham: Springer International Publishing.

Oviatt, S. (2006). Human-centered Design Meets Cognitive Load Theory: Designing Interfaces That Help People Think. In *Proceedings of the 14th ACM International Conference on Multimedia* (pp. 871–880). New York, NY, USA: ACM. https://doi.org/10.1145/1180639.1180831

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The use of multiple strategies in judgment and choice.* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a Measure of Cognitive Effort in Younger and Older Adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Poole, A., & Ball, L. (2005). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects. In C. Ghaoui (Ed.), *Encyclopedia of*

Human Computer Interaction. IGI Global. https://doi.org/citeulike-article-id:3431568

Porter, G., Troscianko, T., & D Gilchrist, I. (2007). *Effort during visual search and counting: Insights from pupillometry*. *Quarterly journal of experimental psychology (2006)* (Vol. 60). https://doi.org/10.1080/17470210600673818

Priest, L., Nayak, L., & Stuart-Hamilton, I. (2007). Website task performance by older adults. *Behaviour & Information Technology*, *26*(3), 189–195. https://doi.org/10.1080/01449290500402668

Ragu-Nathan, T. S., Tarafdar, M., Ragu-Nathan, B. S., & Tu, Q. (2008). The Consequences of Technostress for End Users in Organizations: Conceptual Development and Empirical Validation. *Information Systems Research*, *19*(4), 417–433. https://doi.org/10.1287/isre.1070.0165

Rao, R. B., Fung, G., & Rosales, R. (2008). On the Dangers of Cross-Validation. An Experimental Evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 588–596). https://doi.org/doi:10.1137/1.9781611972788.54

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465. https://doi.org/10.1037/0882-7974.21.3.448

Rayner K., Pollatsek A., et al. (2012). Psychology of Reading: 2nd Edition - Keith Rayner, Alexander Pollatsek, Jane Ashby, Charles Clifton Jr. - Google Libri.

Reinsch, C. H. (1971). Smoothing by spline functions, II. *Numer. Math.*, *16*(3), 451–454.

https://doi.org/10.1007/BF02169154

Richstone;, L., Schwartz;, M. J., Seideman;, C., Cadeddu;, J., Marshall;, S., & Kavoussi, R., L. (2010). Eye metrics as an objective assessment of surgical skill. *Annals of Surgery*, *252*(1), 177–182. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20562602

Riedl Rajiv D.; Benbasat, Izak; Davis, Fred D.; Dennis, Alan R.; Dimoka, Angelika; Gefen, David; Gupta, Alok; Ischebeck, Anja; Kenning, Peter; Müller-Putz, Gernot; Pavlou, Paul A.; Straub, Detmar W.; vom Brocke, Jan; and Weber, Bernd, R. B. (2010). On the foundations of NeuroIS: reflections on the Gmunden Retreat 2009. *Communications of the Association for Information Systems*, *27*, 243–264.

Romano Bergstrom, J. C., Olmsted-Hawala, E. L., & Jans, M. E. (2013). Age-Related Differences in Eye Tracking and Usability Performance: Website Usability for Older Adults. *International Journal of Human–Computer Interaction*, *29*(8), 541–548. https://doi.org/10.1080/10447318.2012.728493

Rosch, J. L., & Vogel-Walcutt, J. J. (2013). A review of eye-tracking applications as tools for training. *Cogn. Technol. Work*, *15*(3), 313–327. https://doi.org/10.1007/s10111-012-0234-7

Salojärvi, J., Kojo, I., Simola, J., & Kaski, S. (2003a). Can relevance be inferred from eye movements in information retrieval. *Proceedings of the 4th Workshop on Self-Organizing Maps (WSOM 2003)*, (September), 261–266. https://doi.org/10.1145/1460096.1460120

Salojärvi, J., Kojo, I., Simola, J., & Kaski, S. (2003b). Can Relevance be Inferred from Eye Movements in Information Retrieval? In *WSOM Workshop on Self-Organizing Maps, Hibikino, Kitakyushu, Japan, September 2003* (pp. 261–266).

Salojärvi, J., Puolamäki, K., Simola, J., Kovanen, L., Kojo, I., & Kaski, S. (2005). Inferring Relevance from Eye Movements: Feature Extraction. *Tech Rep. A82 Helsinki Univ. of Technology. Publication in Computer and Information Science*, 1–23. https://doi.org/10.1.1.96.6356

Salthouse, T. A. (1996). The Processing-Speed Theory of Adult Age Diff erences in Cognition. *Psychological Review*, *103*(3), 403–428. https://doi.org/10.1037/0033-295X.103.3.403

Sauro;, J., & Lewis, J. R. (2012). *Quantifying The User Experience: Practical Statistics For User Research* (2nd ed.). Morgan Kaufmann. Retrieved from https://measuringu.com/book/quantifying-the-user-experiencepractical-statistics-for-user-research/

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, *8*(4), 375–397. https://doi.org/10.1007/s11412-013-9181-4

Schultheis, H., & Jameson, A. (2004). Assessing Cognitive Load in Adaptive Hypermedia Systems: Physiological and Behavioral Methods. In W. Nejdl (Ed.) (pp. 225–234). Berlin, Heidelberg LB - 10.1007/978-3-540-27780-4_26: Springer Berlin Heidelberg.

Shah, J., Wiken, J., Williams, B., & Breazeal, C. (2011). Improved human-robot team performance using chaski, a human-inspired plan execution system. *Proceedings of the 6th International Conference on Human-Robot Interaction*. Lausanne, Switzerland: ACM. https://doi.org/10.1145/1957656.1957668

Shojaeizadeh, M., Djamasbi, S., Chen, P., & Rochford, J. (2017). Task Condition and Pupillometry . In *Twenty-third Americas Conference on Information Systems*. Boston, MA.

Shojaeizadeh, M., Djamasbi, S., Chen, P., & Rochford, J. (2017a). Task Condition and Pupillometry Full Papers, (2011), 1–8.

Shojaeizadeh, M., Djamasbi, S., Chen, P., & Rochford, J. (2017b). Text Simplification and Pupillometry: An Exploratory Study BT - Augmented Cognition. Enhancing Cognition and Behavior in Complex Human Environments. In D. D. Schmorrow & C. M. Fidopiastis (Eds.) (pp. 65–77). Cham: Springer International Publishing.

Shojaeizadeh M.; Trapp A.; Djamasbi S. (2015). Does Pupillary Data Differ During Fixations and Saccades? Does it Carry Information About Task Demand? In *Thirteenth Annual*

Workshop on HCI Research in MIS. Fort Worth, Texas, USA.

Simola, J., Salojärvi, J., & Kojo, I. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, *9*(4), 237–251. https://doi.org/http://dx.doi.org/10.1016/j.cogsys.2008.01.002

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Soussan, Djamasbi, Marisa, Siegel, & Tullis, T. (2011). Visual Hierarchy and Viewing Behavior: An Eye Tracking Study. In J. Jacko (Ed.), *Human-Computer Interaction. Design and Development Approaches* (Vol. 6761, pp. 331–340). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21602-2_36

Specification of Gaze Precision and Gaze Accuracy. (2016). Retrieved from http://www.tobiipro.com/siteassets/tobii-pro/brochures/tobii-pro-t60xl-brochure-ux-market-research.pdf

Stassen, H. G., Johannsen, G., & Moray, N. (1990). Internal representation, internal model, human performance model and mental workload. *Automatica*, *26*(4), 811–820. https://doi.org/https://doi.org/10.1016/0005-1098(90)90057-O

Steichen, B., Conati, C., & Carenini, G. (2014). Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Trans. Interact. Intell. Syst.*, *4*(2), 11:1--11:29. https://doi.org/10.1145/2633043

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (2011). CHAPTER TWO - Cognitive Load Theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of Learning and Motivation* (Vol. 55, pp. 37–76). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Tarafdar, M., Tu, Q., & Ragu-Nathan, T. (2010). Impact of Technostress on End-User Satisfaction and Performance. *J. Manage. Inf. Syst.*, *27*(3), 303–334. https://doi.org/10.2753/mis0742-1222270311

Theofanos, M. F., & Redish, J. (Ginny). (2003). Guidelines for Accessible and Usable Web Sites: Observing Users Who Work With Screen Readers. *Interactions*, *X*(6), 38–51. https://doi.org/10.1145/947226.947227

Todd, P., & Benbasat, I. (1991). An Experimental Investigation of the Impact of Computer Based Decision Aids on Decision Making Strategies. *Information Systems Research*, *2*(2), 87–115. Retrieved from http://www.jstor.org/stable/23010636

Todd, P., & Benbasat, I. (1992). The Use of Information in Decision Making: An Experimental Investigation of the Impact of Computer-Based Decision Aids. *MIS Quarterly*, *16*(3), 373–393. https://doi.org/10.2307/249534

Todd, P., & Benbasat, I. (1994). The Influence of Decision Aids on Choice Strategies: An Experimental Analysis of the Role of Cognitive Effort. *Organizational Behavior and Human Decision Processes*, *60*(1), 36–74. https://doi.org/https://doi.org/10.1006/obhd.1994.1074

Todd, P., & Benbasat, I. (1999). Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection. *Information Systems Research*, *10*(4), 356–374. https://doi.org/10.1287/isre.10.4.356

Tullis, T. S. (2007). Older Adults and the Web: Lessons Learned from Eye-Tracking BT - Universal Acess in Human Computer Interaction. Coping with Diversity. In C. Stephanidis (Ed.) (pp. 1030–1039). Berlin, Heidelberg: Springer Berlin Heidelberg.

Turns, J., & Wagner, T. S. (2004). Characterizing Audience for Informational Web Site Design. *Technical Communication*, *51*(1), 68–85.

van den Brink, R. L., Murphy, P. R., & Nieuwenhuis, S. (2016). Pupil Diameter Tracks Lapses of Attention. *PLOS ONE*, *11*(10), e0165274. Retrieved from https://doi.org/10.1371/journal.pone.0165274

Van Orden, K., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(1), 111–121. https://doi.org/citeulike-article-id:10126630 doi: 10.1518/001872001775992570

Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems*, *62*, 1–10. https://doi.org/10.1016/j.dss.2014.02.007

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

Wong, K. K. W. ., Wan, W. Y. ., & Kaye, S. B. (2002). Blinking and operating: cognition versus vision. *The British Journal of Ophthalmology*, *86*(4), 479. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1771097/

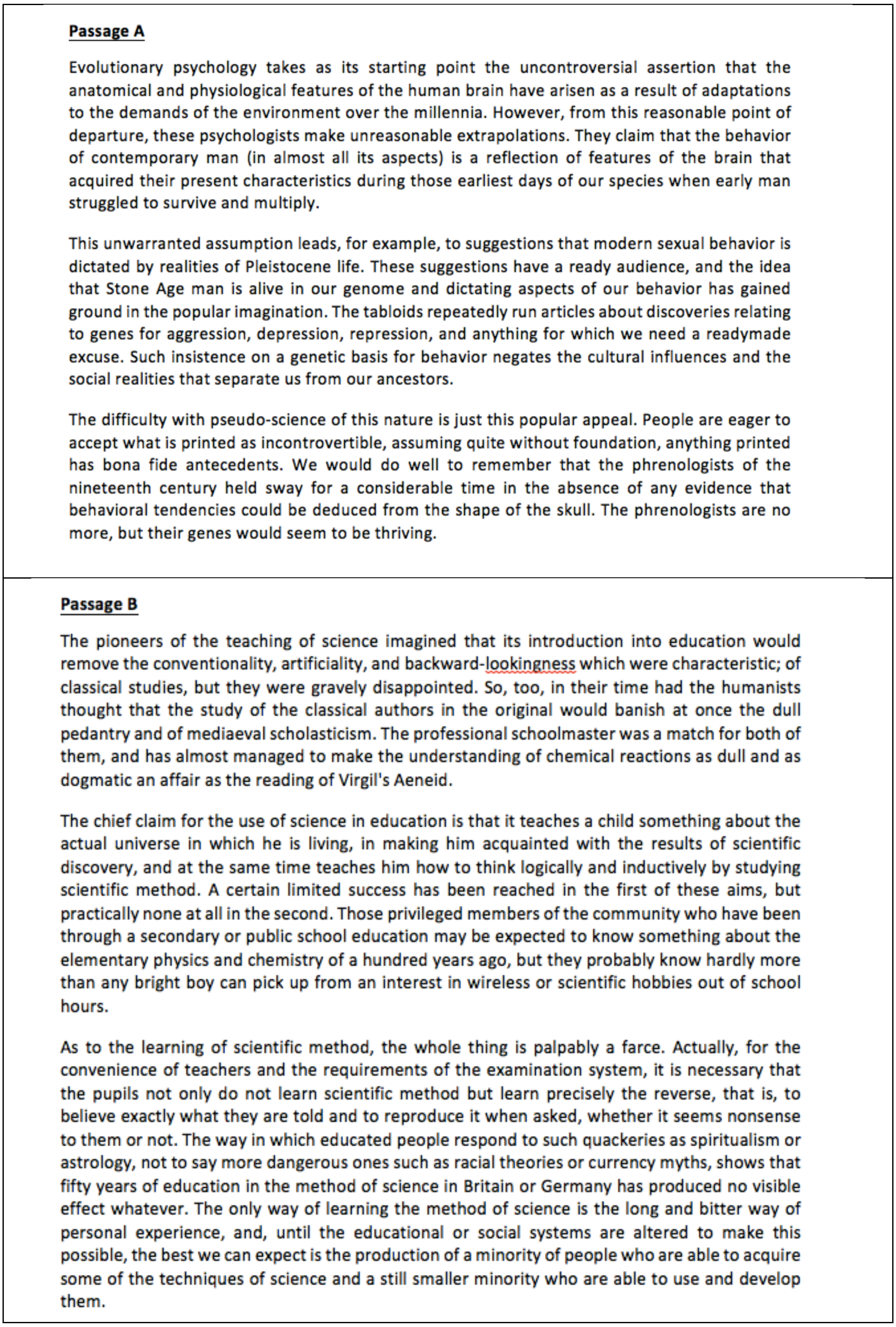
Yıldırım, P. (2016). Pattern Classification with Imbalanced and Multiclass Data for the Prediction of Albendazole Adverse Event Outcomes. *Procedia Computer Science*, *83*, 1013–1018. https://doi.org/https://doi.org/10.1016/j.procs.2016.04.216

Zaphiris, P., & Savtich, N. (2008). *Age-related Differences in Browsing the Web. SPARC Strategic Promotion of Ageing Research Capacity*. Lon.

Zijlstra, F. ;, & van Doorn, L. (1985). *The construction of a scale to measure subjective effort*. Retrieved from https://adasgeek.wordpress.com/2013/12/19/rsme/

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577. https://doi.org/ROC; Receiver-Operating Characteristic; SDT; Signal Detection Theory

Appendix A



Passage A

Evolutionary psychology takes as its starting point the uncontroversial assertion that the anatomical and physiological features of the human brain have arisen as a result of adaptations to the demands of the environment over the millennia. However, from this reasonable point of departure, these psychologists make unreasonable extrapolations. They claim that the behavior of contemporary man (in almost all its aspects) is a reflection of features of the brain that acquired their present characteristics during those earliest days of our species when early man struggled to survive and multiply.

This unwarranted assumption leads, for example, to suggestions that modern sexual behavior is dictated by realities of Pleistocene life. These suggestions have a ready audience, and the idea that Stone Age man is alive in our genome and dictating aspects of our behavior has gained ground in the popular imagination. The tabloids repeatedly run articles about discoveries relating to genes for aggression, depression, repression, and anything for which we need a readymade excuse. Such insistence on a genetic basis for behavior negates the cultural influences and the social realities that separate us from our ancestors.

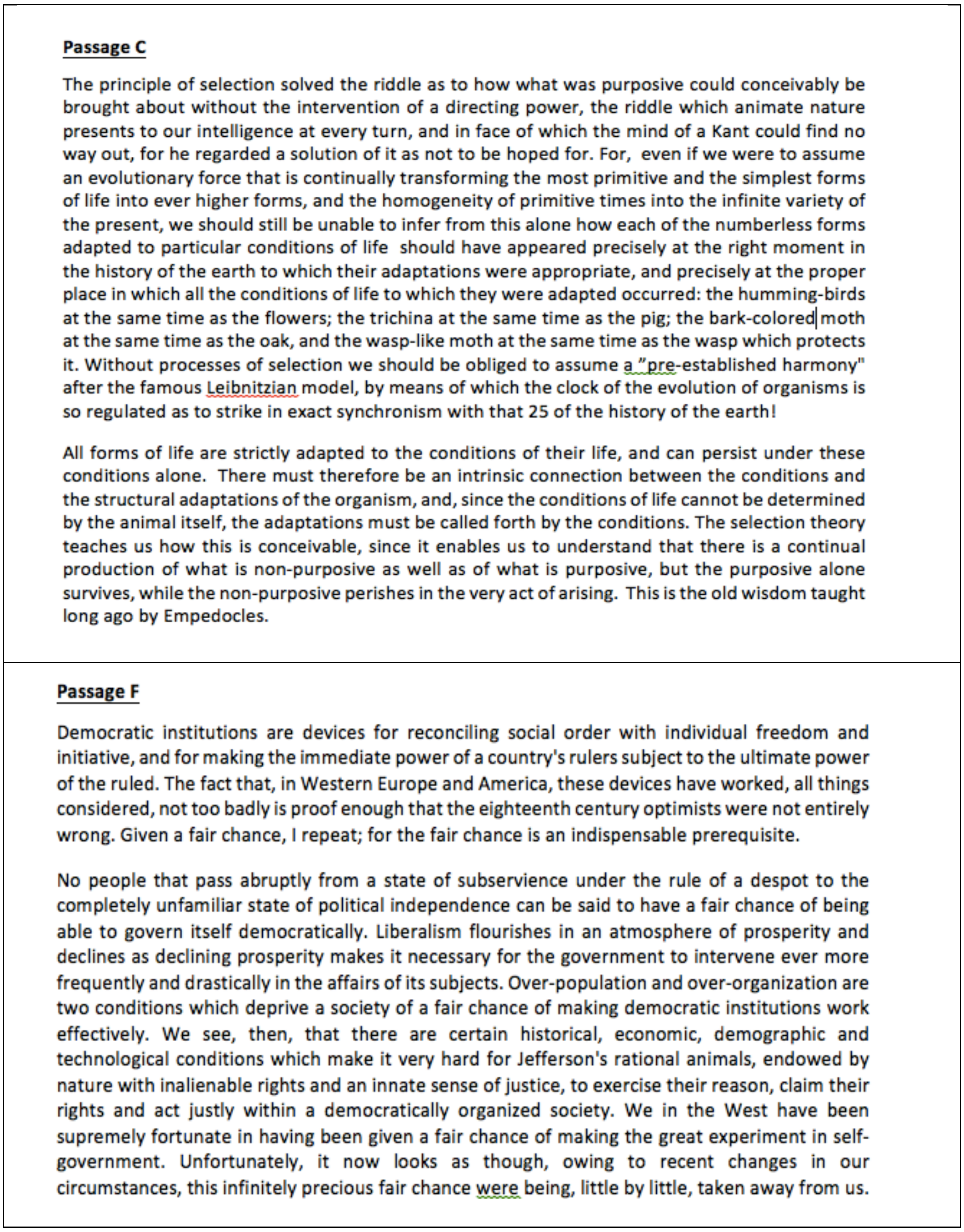
The difficulty with pseudo-science of this nature is just this popular appeal. People are eager to accept what is printed as incontrovertible, assuming quite without foundation, anything printed has bona fide antecedents. We would do well to remember that the phrenologists of the nineteenth century held sway for a considerable time in the absence of any evidence that behavioral tendencies could be deduced from the shape of the skull. The phrenologists are no more, but their genes would seem to be thriving.

Passage B

The pioneers of the teaching of science imagined that its introduction into education would remove the conventionality, artificiality, and backward-lookingness which were characteristic; of classical studies, but they were gravely disappointed. So, too, in their time had the humanists thought that the study of the classical authors in the original would banish at once the dull pedantry and of mediaeval scholasticism. The professional schoolmaster was a match for both of them, and has almost managed to make the understanding of chemical reactions as dull and as dogmatic an affair as the reading of Virgil's Aeneid.

The chief claim for the use of science in education is that it teaches a child something about the actual universe in which he is living, in making him acquainted with the results of scientific discovery, and at the same time teaches him how to think logically and inductively by studying scientific method. A certain limited success has been reached in the first of these aims, but practically none at all in the second. Those privileged members of the community who have been through a secondary or public school education may be expected to know something about the elementary physics and chemistry of a hundred years ago, but they probably know hardly more than any bright boy can pick up from an interest in wireless or scientific hobbies out of school hours.

As to the learning of scientific method, the whole thing is palpably a farce. Actually, for the convenience of teachers and the requirements of the examination system, it is necessary that the pupils not only do not learn scientific method but learn precisely the reverse, that is, to believe exactly what they are told and to reproduce it when asked, whether it seems nonsense to them or not. The way in which educated people respond to such quackeries as spiritualism or astrology, not to say more dangerous ones such as racial theories or currency myths, shows that fifty years of education in the method of science in Britain or Germany has produced no visible effect whatever. The only way of learning the method of science is the long and bitter way of personal experience, and, until the educational or social systems are altered to make this possible, the best we can expect is the production of a minority of people who are able to acquire some of the techniques of science and a still smaller minority who are able to use and develop them.



Passage C

The principle of selection solved the riddle as to how what was purposive could conceivably be brought about without the intervention of a directing power, the riddle which animate nature presents to our intelligence at every turn, and in face of which the mind of a Kant could find no way out, for he regarded a solution of it as not to be hoped for. For, even if we were to assume an evolutionary force that is continually transforming the most primitive and the simplest forms of life into ever higher forms, and the homogeneity of primitive times into the infinite variety of the present, we should still be unable to infer from this alone how each of the numberless forms adapted to particular conditions of life should have appeared precisely at the right moment in the history of the earth to which their adaptations were appropriate, and precisely at the proper place in which all the conditions of life to which they were adapted occurred: the humming-birds at the same time as the flowers; the trichina at the same time as the pig; the bark-colored moth at the same time as the oak, and the wasp-like moth at the same time as the wasp which protects it. Without processes of selection we should be obliged to assume a ”pre-established harmony" after the famous Leibnitzian model, by means of which the clock of the evolution of organisms is so regulated as to strike in exact synchronism with that 25 of the history of the earth!

All forms of life are strictly adapted to the conditions of their life, and can persist under these conditions alone. There must therefore be an intrinsic connection between the conditions and the structural adaptations of the organism, and, since the conditions of life cannot be determined by the animal itself, the adaptations must be called forth by the conditions. The selection theory teaches us how this is conceivable, since it enables us to understand that there is a continual production of what is non-purposive as well as of what is purposive, but the purposive alone survives, while the non-purposive perishes in the very act of arising. This is the old wisdom taught long ago by Empedocles.

Passage F

Democratic institutions are devices for reconciling social order with individual freedom and initiative, and for making the immediate power of a country's rulers subject to the ultimate power of the ruled. The fact that, in Western Europe and America, these devices have worked, all things considered, not too badly is proof enough that the eighteenth century optimists were not entirely wrong. Given a fair chance, I repeat; for the fair chance is an indispensable prerequisite.

No people that pass abruptly from a state of subservience under the rule of a despot to the completely unfamiliar state of political independence can be said to have a fair chance of being able to govern itself democratically. Liberalism flourishes in an atmosphere of prosperity and declines as declining prosperity makes it necessary for the government to intervene ever more frequently and drastically in the affairs of its subjects. Over-population and over-organization are two conditions which deprive a society of a fair chance of making democratic institutions work effectively. We see, then, that there are certain historical, economic, demographic and technological conditions which make it very hard for Jefferson's rational animals, endowed by nature with inalienable rights and an innate sense of justice, to exercise their reason, claim their rights and act justly within a democratically organized society. We in the West have been supremely fortunate in having been given a fair chance of making the great experiment in self-government. Unfortunately, it now looks as though, owing to recent changes in our circumstances, this infinitely precious fair chance were being, little by little, taken away from us.

Figure A. 1 Passages Selected for the Passage Selection Process

Table A. 1 Mean and STD of Eye Movement Features Used for Cognitive Load Classification for Passage A

Eye Features	**OA (Mean ± STD)**	**SA (Mean ± STD)**
Avg. Saccade Duration	27.36 ± 3.12	27.74 ± 3.08
STD Saccade Duration	17.35 ± 2.50	17.30 ± 2.64
Avg. Progressive Saccade Amplitude	2.69 ± 0.56	2.71 ± 0.59
STD Progressive Saccade Amplitude	2.53 ± 0.75	2.54 ± 0.76
Avg. Regressive Saccade Amplitude	8.17 ± 2.15	7.40 ± 2.15
STD Regressive Saccade Amplitude	9.17 ± 1.42	8.49 ±1.69
Regressive Saccade Count	0.22 ± 0.10	0.22 ± 0.08
Avg. Regressive Saccade Duration	38.71 ± 2.91	36.63 ± 3.27
STD Regressive Saccade Duration	23.75 ± 2.44	22.75± 3.02
Avg. Progressive Saccade Duration	26.28 ± 2.16	26.45 ± 1.77
STD Progressive Saccade Duration	13.97 ± 5.94	12.63 ± 2.67
Progressive Saccade Count	0.94 ± 0.33	0.89 ± 0.27
Regressive Saccade Count/Progressive Saccade Count	24.90 ± 8.50	25.53 ± 7.54
Total Fixation Duration	276.32 ± 112.80	279.97 ± 91.65
Avg. Fixation Duration	227.85 ± 24.04	232.37 ± 21.60
STD Fixation Duration	103.03 ± 20.88	101.55 ± 19.60
Fixation Duration/ Saccade Duration	8.45 ± 1.43	8.50 ± 1.36
Pupil Dilation PSD	0.98 ± 0.40	1.85 ± 0.31
Pupil Variation PSD	1.56 ± 0.25	1.55 ± 0.21
Pupil Dilation –Fixation –PSD	5.90 ± 3.09	4.02 ± 3.13
Pupil Dilation-Saccade-PSD	5.99 ± 3.1	5.39 ±3.08
Pupil Variation-Fixation-PSD	0.60 ± 0.29	0.56 ± 0.28

Pupil Variation-Saccade-PSD	0.80 ± 0.38	0.77 ± 0.38
Avg. Pupil Dilation-Fixation	2.80 ± 0.38	2.83 ± 0.38
STD Pupil Dilation-Fixation	0.108 ± 0.034	0.112 ± 0.033
Avg. Pupil Dilation-Saccade	2.80 ± 0.37	2.84 ±0.36
STD Pupil Dilation-Saccade	0.114 ± 0.036	0.119 ± 0.37
Avg. Pupil Variation-Fixation	0.039 ± 0.016	0.040 ± 0.016
STD Pupil Variation-Fixation	0.054 ± 0.021	0.055 ± 0.021
Avg. Pupil Variation-Saccade	0.043 ± 0.016	0.044 ± 0.016
STD Pupil Variation-Saccade	0.062 ± 0.022	0.064 ± 0.022
Avg. Pupil Dilation-Fixation/Avg. Pupil Dilation –Saccade	0.99 ± 0.0025	0.99 ± 0.0026
Avg. Pupil Variation-Fixation/Avg. Pupil Variation –Saccade	0.9157 ± 0.103	0.9156 ± 0.102

Table A. 2 Mean and STD of Eye Movement Features Used for Cognitive Load Classification for Passage B

Eye Features	**OB (Mean ± STD)**	**SB (Mean ± STD)**
Avg. Saccade Duration	28.85 ± 3.47	28.24 ± 3.08
STD Saccade Duration	18.25 ± 3.53	17.41 ± 2.58
Avg. Progressive Saccade Amplitude	3.07 ± 0.77	2.95 ± 0.63
STD Progressive Saccade Amplitude	2.95 ± 0.93	2.65 ± 0.86
Avg. Regressive Saccade Amplitude	8.01 ± 2.21	7.97 ± 1.98
STD Regressive Saccade Amplitude	8.73 ± 1.50	8.82 ±1.55
Regressive Saccade Count	0.26 ± 0.09	0.24 ± 0.09
Avg. Regressive Saccade Duration	38.76 ± 4.11	40.35 ± 3.47
STD Regressive Saccade Duration	25.23 ± 6.32	24.80 ± 3.11
Avg. Progressive Saccade Duration	28.16 ± 1.84	25.22 ± 2.09

STD Progressive Saccade Duration	14.32 ± 5.04	12.50 ± 3.04
Progressive Saccade Count	0.93 ± 0.3	0.93 ± 0.36
Regressive Saccade Count/Progressive Saccade Count	28.69 ± 9.37	26.78 ± 8.45
Total Fixation Duration	291.59 ± 105.88	282.91 ± 79.89
Avg. Fixation Duration	228.28 ± 20.06	223.78 ± 20.82
STD Fixation Duration	101.75 ± 19.84	98.96 ± 19.32
Fixation Duration/ Saccade Duration	8.03 ± 1.24	8.04 ± 1.31
Pupil Dilation PSD	1.87± 0.31	1.80 ± 0.37
Pupil Variation PSD	1.58 ± 0.22	1.54 ± 0.24
Pupil Dilation –Fixation –PSD	0.82 ± 0.29	0.73 ± 0.28
Pupil Dilation-Saccade-PSD	1.042 ± 0.45	0.89 ± 0.35
Pupil Variation-Fixation-PSD	0.59 ± 0.23	0.52 ± 0.23
Pupil Variation-Saccade-PSD	0.80 ± 0.38	0.69 ± 0.29
Avg. Pupil Dilation-Fixation	2.86 ± 0.33	2.79 ± 0.36
STD Pupil Dilation-Fixation	0.115 ± 0.03	0.105 ± 0.03
Avg. Pupil Dilation-Saccade	2.86 ± 0.33	2.79 ± 0.36
STD Pupil Dilation-Saccade	0.125 ± 0.39	0.112 ± 0.038
Avg. Pupil Variation-Fixation	0.042 ± 0.013	0.039 ± 0.015
STD Pupil Variation-Fixation	0.059 ± 0.021	0.053 ± 0.021
Avg. Pupil Variation-Saccade	0.05 ± 0.01	0.42 ± 0..02
STD Pupil Variation-Saccade	0.07 ± 0.021	0.06 ± 0.021
Avg. Pupil Dilation-Fixation/Avg. Pupil Dilation –Saccade	1 ± 0.004	0.99 ± 0.002
Avg. Pupil Variation-Fixation/Avg. Pupil Variation –Saccade	0.93 ± 0.11	0.92 ± 0.09

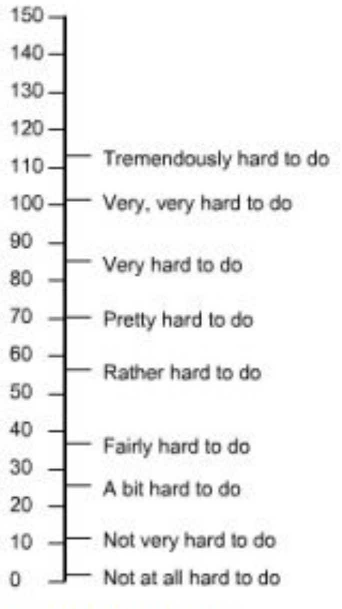


Figure A. 2 Subjective Mental Effort Questionnaire