Eye Tracking to Support eLearning

Leana Copeland

August 2016

A thesis submitted for the degree of
Doctor of Philosophy
of The Australian National University

Research School of Computer Science
College of Engineering and Computer Science
The Australian National University

To Mum, Dad and Michael,

My favourite people in the world!

Except where otherwise indicated, this thesis is my own original work.

Leana Copeland

August 2016

Acknowledgements

First I would like to thank my supervisor Professor Tom Gedeon. I would not have got through this without you. The support and encouragement you gave me during my honours year and then throughout my candidature will never been forgotten. You fostered within me a curiosity and drive to succeed that I never knew I had. You are a one million supervisor and I am very lucky to have had you as mine. Not every supervisor would put up with their students following them down hallways or pounce on them when they get back their office to ask non-stop questions!

Mum, Dad, and Michael, thank you does not cover my gratitude to you for not only supporting me financially, emotionally, and mentally throughout my PhD, but also for maintaining an open fridge-door policy! Thank you for not only reading all of my conference and journal papers but also for reading every single chapter of this thesis (twice)! This thesis would not be in the state that it is without your help picking up the many spelling and grammar mistakes. Thank you for everything you have done for me. It is because of you I have made it this far, and once again you have helped get me over the line. I love you all and will forever appreciate your never-ending love and support in everything I do.

Many thanks go to my panel, Dr Sumudu Mendis and Dr Kristen Palmer, for your support and help. Thank you also to the staff from CECS for your support and help, notably Dr Richard Jones, Dr Duncan Stevenson, and Lynette Johns-Boast.

Thank you to Dr Sabrina Caldwell, your chats and insightful wisdom will be missed and were much needed at many times during my candidature. Thank you also for being a fantastic co-experimenter, my thesis really would not be what it is without you.

A very big thank you goes to everyone who proofread for me and provided invaluable feedback, especially in correcting my grammar. In particular, Tom Gedeon, Wendy and Chris Copeland, Richard Jones, Jaimi Pigram, and Chris Chow.

Thank you to my wonderful friends Nandita Sharma, Maria Memmolo, Brigette Boast, Lara Connolly, Jessica Tsimeris, Jennyfer Lawrence Taylor, Oliver Thearle, Jonathan and Melanie Greenshaw and Richa Awasthy for your love and support throughout my candidature. The time would not have been as enjoyable without you all by my side.

Thank you to my amazing friends Jaimi Pigram and Katie Hotchkis who have seen me through high school, my undergraduate studies, and now my postgraduate studies. Thank you for your love and support. I know that there were many times that you had to deal with me being tired and stressed, and you were always there for me to keep me going and put a smile back on my face. I never could have got to this point without you both and I am and always will be grateful for our everlasting friendship.

Last, but by no means least, a big thank you to Chris Chow for your love and support. You have always been there to help out when I needed it; especially in expanding my vocabulary to include words such as woodpile and trunk, but also for always being available to grab a coffee and having a chat when I needed it most! Thank you for staying up until after midnight to help me print copies of this thesis, for being there while I read the examiners comments, and for making me the best cheesecake of my life! I am forever grateful for your support, advice on TV shows, modelling in my photo shoots, and for always making me smile.

Abstract

Online eLearning environments to support student learning are of growing importance. Students are increasingly turning to online resources for education; sometimes in place of face-to-face tuition. Online eLearning extends teaching and learning from the classroom to a wider audience with different needs, backgrounds, and motivations. The one-size-fits-all approach predominately used is not effective for catering to the needs of all students. An area of the increasing diversity is the linguistic background of readers. More students are reading in their non-native language. It has previously been established that first English language (L1) students read differently to second English language (L2) students. One way of analysing this difference is by tracking the eyes of readers, which is an effective way of investigating the reading process.

In this thesis we investigate the question of whether eye tracking can be used to make learning via reading more effective in eLearning environments. This question is approached from two directions; first by investigating how eye tracking can be used to adapt to individual student's understanding and perceptions of text. The second approach is analysing a cohort's reading behaviour to provide information to the author of the text and any related comprehension questions regarding their suitability and difficulty.

To investigate these questions, two user studies were carried out to collect eye gaze data from both L1 and L2 readers. The first user study focussed on how different presentation methods of text and related questions affected not only comprehension performance but also reading behaviour and student perceptions of performance. The data from this study was used to make predictions of reading comprehension that can be used to make eLearning environments adaptive, in addition to providing implicit feedback about the difficulty of text and questions.

In the second study we investigate the effects of text readability and conceptual difficulty on eye gaze, prediction of reading comprehension, and perceptions. This study showed that readability affected the eye gaze of L1 readers and conceptual difficulty affected the eye gaze of L2 readers. The prediction accuracy of comprehension was consequently increased for the L1 group by increased difficulty in readability, whereas increased difficulty in conceptual level corresponded to increased accuracy for the L2 group. Analysis of participants' perceptions of complexity revealed that readability and conceptual difficulty interact making the

two variables hard for the reader to disentangle. Further analysis of participants' eye gaze revealed that both the predefined and perceived text complexity affected eye gaze. We therefore propose using eye gaze measures to provide feedback about the implicit reading difficulty of texts read.

The results from both studies indicate that there is enormous potential in using eye tracking to make learning via reading more effective in eLearning environments. We conclude with a discussion of how these findings can be applied to improve reading within eLearning environments. We propose an adaptive eLearning architecture that dynamically presents text to students and provides information to authors to improve the quality of texts and questions.

List of Publications

The work in this thesis is the original work of the author except where specific reference or acknowledgement is made to the work or contribution of others. Some of the material in this work has appeared in publications and presentations by the author. Only the contribution made by the author has been included in this work unless specific reference to the contrary has been made.

The publications produced during the thesis are:

1. Copeland, L., & Gedeon, T. D. (2015a). Tutorials in eLearning; How Presentation Affects Outcomes. Emerging Topics in Computing, IEEE Transactions on, PP(99), 1-1.
2. Copeland, L., Gedeon, T., & Caldwell, S. (2015). Effects of Text Difficulty and Readers on Predicting Reading Comprehension from Eye Movements. Paper presented at the IEEE 6th International Conference on Cognitive Infocommunications (CogInfoCom) 2015, Győr, Hungary.
3. Copeland, L., & Gedeon, T. (2015). Visual Distractions Effects on Reading in Digital Environments: A Comparison of First and Second English Language Readers. Paper presented at the Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, 506-516.
4. Copeland, L., & Gedeon, T. (2014a). Effect of presentation on reading behaviour. In Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design (pp. 230-239). ACM.
5. Copeland, L., & Gedeon, T. (2014b). What are You Reading Most: Attention in eLearning. Procedia Computer Science, 39, 67-74.
6. Copeland, L., Gedeon, T., & Mendis, S. (2014a). Fuzzy Output Error as the Performance Function for Training Artificial Neural Networks to Predict Reading Comprehension from Eye Gaze. Paper presented at The 21st International Conference on Neural Information Processing 2014.
7. Copeland, L., Gedeon, T., & Mendis, S. (2014b). Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. Artificial Intelligence Research, 3(3), p35.

8. Copeland, L., Gedeon, T., & Mendis, B. S. (2014). An Investigation of Fuzzy Output Error as an Error Function for Optimisation of Fuzzy Signature Parameters. RCSC TR-1 2014.
9. Copeland, L., Gedeon, T., & Caldwell, S. (2014). Framework for Dynamic Text Presentation in eLearning. Procedia Computer Science, 39, 150-153.
10. Copeland, L., & Gedeon, T. (2013a). The effect of subject familiarity on comprehension and eye movements during reading. Paper presented at the Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration.
11. Copeland, L., & Gedeon, T. (2013b). Measuring reading comprehension using eye movements. Paper presented at the Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on.
12. Caldwell, S., Gedeon, T., Jones, R., & Copeland, L. (2015) Imperfect Understandings: A Grounded Theory and Eye Gaze Investigation of Human Perceptions Of Manipulated And Unmanipulated Digital Images. In Proceedings of 3rd International Conference on Multimedia and Human-Computer Interaction ***(Winner of the Best Paper award)***
13. Naqshbandi, K., Gedeon, T., Abdulla, U. A., & Copeland, L. (2015). Factors affecting identification of tasks using eye gaze. Paper presented at the Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on.
14. Taylor, J. L., Copeland, L., Chow, C., & Nitschke, K. (2015) VIRK: Virtual work environment to facilitate interaction between the unemployed. Paper presented at the Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction as part of the OzCHI 24 hour student design challenge. ***(Winner of the student design challenge)***
15. Gedeon, T., Zhu, X., He, K., & Copeland, L. (2014, January). Fuzzy Signature Neural Networks for Classification: Optimising the Structure. In Neural Information Processing (pp. 335-341). Springer International Publishing.
16. Gedeon, T., Copeland, L., & Mendis, B. S. (2012). Fuzzy Output Error. Australian Journal of Intelligent Information Processing Systems, 13(2), 37-43.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"The more that you read, the more things you will know.
The more that you learn, the more places you'll go."
— Dr. Seuss, *I Can Read With My Eyes Shut!*

Online learning could extend teaching and learning from the classroom to a wide and varied audience with different needs, backgrounds, and motivations. Particularly in tertiary education, online learning technologies are becoming ubiquitous. This is due in part to increased accessibility and availability of computer hardware but also due to an influx of eLearning software and services. Universities now frequently offer online or off-campus degrees where students may have little or no face-to-face interaction with their instructors or other students. Even for university courses that deliver traditionally using face-to-face tuition, absenteeism from lectures is more prevalent and has been shown to negatively affect learning (Romer, 1993; Woodfield et al., 2006).

However, the use of online learning can be beneficial in dealing with not only this problem but also the problems encountered by large class sizes as well as dispersed students, by providing consistency and accessibility in delivered materials (Welsh, Wanberg, Brown, & Simmering, 2003). The advent of massive open online courses (MOOCs) has not only increased the number of online learning users but also increased diversity (especially in native language) as students from around the world are able to access the content (Breslow et al., 2013; DeBoer et al., 2013). The low completion rates of MOOCs highlight the significant need to improve online learning technologies (Breslow et al., 2013). As a result of these factors there is growing importance in designing effective eLearning environments.

Most eLearning environments are one-size-fits-all, yet this does not account for differences in students' needs, backgrounds, or native language. One of the increasing diversities is the linguistic background of readers. There are an increasing number of students who are reading in their non-native language. A

study of edX's[1] first MOOC showed students came from 194 different countries and two-thirds spoke English where the other third spoke other languages (Breslow et al., 2013). It is known that first English language (L1) and second English language (L2) readers differently (Rayner, 1998). Students also vary in respect to their prior knowledge, expertise, and reading abilities. These differences can impact the processing needed to properly comprehend text. Text characteristics have been shown to affect comprehension by which, in the context of legal documents, making text simpler would benefit vulnerable populations (Scherr, Agauas, & Ashby, 2015). This can be extended to considering the differences of students in eLearning, where some students may be supported by simpler texts.

Whilst some eLearning environments provide personalisation, the learner often does this explicitly. Adaption can be based on different qualities about the learner such as the current understanding, emotional state such as stress (Calvi et al., 2008; Porta, 2008) boredom (Jaques, Conati, Harley, & Azevedo, 2014), motivation (Kareal & Klema, 2006), learner style (Mehigan et al., 2011; Spada et al., 2008; Surjono, 2011, 2014), cognitive load (Coyne et al., 2009), learner style (Bondareva et al., 2013),and skill level (Chen, 2008). Adaption achieved in real time, without disruption to the learner, is the optimal solution rather than explicitly asking the learner. Progress in technology and understanding of psychophysiological responses provide the unique opportunity of adapting eLearning environments in real time and doing so based upon implicit behaviour. These methods include the use of biometric technology (Mehigan et al., 2011; Spada et al., 2008) and psychophysiological response data (Rosch & Vogel-Walcutt, 2013), especially eye tracking (Alsobhi et al., 2015; Barrios et al., 2004; Bondareva et al., 2013; Calvi et al., 2008; Conati, Jaques, & Muir, 2013; Conati & Merten, 2007; Kardan & Conati, 2013; Merten & Conati, 2006; D'Mello et al., 2012; Gütl et al., 2005; Mehigan, 2014; Mehigan & Pitt, 2013; Mehigan, 2013; Mehigan et al., 2011; Porta, 2008).

There is a broad range of scenarios that these adaptive technologies are directed at helping students, such as plugging into traditional online learning environments (Barrios et al., 2004; De Bra et al., 2013), or providing adaption in mobile environments (Mehigan & Pitt, 2013), or accounting for dyslexia (Alsobhi et al., 2015), or foreign language reading (Hyrskykari et al., 2000). With this past research we are able to take the results from the studies presented in this thesis and add to the current knowledge base of adaptive eLearning. The contribution lies solely in the domain of text-based learning materials that have not been focused upon in the past. Eye tracking can certainly be used to make learning via reading more effective in the context of eLearning.

Using eye gaze to control adaption of eye learning environments provides the ability to go beyond the student's surface answering behaviour or preferences and adapt to the student's implicit behaviour. Eye tracking has been shown to be a powerful tool for investigating how humans interact with computer interfaces. It has also been shown to provide information about the differences between L1 and L2 readers (Dednam et al., 2014; Kang, 2014). Eye movements can reveal abundant

[1] edX is a MOOC provider - https://www.edx.org/ (Last accessed: 24th January 2016)

information about the cognitive processes behind human behaviours. Louis Emile Javal noted in the late 1800s that the eyes move in a particular way when someone is reading. Since then eye tracking technologies have vastly improved and together with new brain scanning techniques such as functional magnetic resonance imaging (fMRI) we have a greater knowledge of how humans read (Bowman et al., 2010).

While brain scans provide a good way of seeing how the brain reacts during reading, eye tracking affords the unique ability to observe the underlying cognitive processes of reading in an unobtrusive manner. There is now a plethora of research that investigates how the eye moves during the reading process which go down to the level of predicting where the eye will land on a word and for how long it will fixate (e.g. the E-Z Reader model (Reichle et al., 2006; Reichle et al., 1999, 2003, 2012; Reichle et al., 2009) and the SWIFT model (Engbert et al., 2005). Additionally, research on eye movements during reading has shown that eye movements reveal difficulties in reading (Frazier & Rayner, 1982), text difficulty and comprehension (Rayner et al., 2006), as well as differentiating between L1 and L2 readers (Dednam et al., 2014; Kang, 2014). While we now know a lot about the reading process, the application of it into the design eLearning environments is still in early days. Additionally, there is still much to be learnt about the differences between L1 and L2 readers in the context of eLearning. With eye tracking becoming increasingly more precise whilst decreasing dramatically in cost, the use of such technology in adaptive eLearning is becoming plausible.

The problem of how to make eLearning environments effective to a wide and varied audience is significant; especially when learning materials come in many types and forms, and quite often depend on the subject being taught. For example a mathematics course would have exercises including many mathematical symbols as opposed to a history course, which would be more likely to have text-based materials. The focus of this thesis is on text-based materials and the use of eye tracking technology to analyse reading and learning behaviour. This thesis investigates ways of using eye tracking to make eLearning environments adaptive to the reader based upon their reading behaviours. This can mean real-time alteration of the learning environment to reflect the student's current comprehension and state. It can also mean the use of eye tracking to monitor the cohort's reading and learning behaviours and using this information as a means of improving the quality of the learning materials. Both are investigated in this thesis as a means of exploring the potential for using eye tracking to make eLearning environments better for learning.

The remainder of the introduction chapter outlines the motivations for this thesis; the primary research questions and hypotheses of the investigation; and finally the thesis structure is outlined.

1.1 Motivation

In many countries reading is part of everyday life. Such as reading signs in a building to direct you to the room you want to go to or reading the labels on food packaging. If you have an Internet connection and device capable of connecting to it

then chances are, you are reading something, like Facebook posts, tweets, online news, instant messaging, or your email, since communication is now often carried out via textual means. As Dr Seuss points out so eloquently, reading is a very good way for gaining information and the Internet and computer devices make it easier to access vast amount of information. Reading moulds what we know and what we know is used to develop opinions and base actions upon. This suggests that what a person reads can have a large bearing on their current knowledge, their beliefs, and what they are likely to be interested in. Learning itself is an ability that is shaped by what we know. Reading and learning can therefore be seen as having a somewhat reciprocal relationship.

There has been an increase is the use of eLearning systems. This can be seen both in the educational sector where tertiary institutions quite often use online learning environments in addition to the traditional face-to-face teaching, as well as in industry for employee training. This provides us with a unique opportunity to enhance both the reading and the learning processes due to the capabilities of electronic systems to provide feedback to their users. Already there are systems that record eye movements whilst reading documents to provide implicit feedback about the perceived relevance of parts of a document (Buscher et al., 2012). Furthermore, the ability to record the parts of a document that have not been comprehended properly or read thoroughly could give feedback to the user of the parts of the document that may need to be re-read for better understanding of the content. On the other hand, feedback about how a document is read can provide information to the author as to how easily it is read and understood. In turn, the author can revise the document to make it easier to read and comprehend. In education feedback often comes from assessment results. Presentation of course content may be in the form of slides, readings, tutorials, all of which are increasingly presented online. Feedback about how students comprehend and read these documents may offer insight into assessment results. This kind of feedback could provide invaluable information to instructors about how to better present course content.

1.2 Primary Research Questions

The central research question of this thesis is:

Can eye tracking be used to make eLearning environments more effective for first and second language English readers?

This is a broad question, which is broken down to look at ways in which eye tracking could be used to make eLearning environments more effective for reading English by both first and second language readers. Primarily the investigation in this thesis will be the use of eye tracking data to make predictions about reading comprehension and text properties. In this way, the broad question is divided into sub-questions that are addressed throughout this thesis:

1. Can outcomes of eye gaze analysis be used to optimise the layout of reading materials in eLearning environments for learning outcomes? How does the layout compare for L1 and L2 readers?
2. Can eye gaze be used to provide feedback about learning behaviour in eLearning environments for L1 and L2 readers?
3. Can eye tracking data be used to predict reading comprehension scores in eLearning environments for L1 and L2 readers?
    a. Does presentation of text affect predictions of comprehension?
    b. Does text difficulty affect predictions of comprehension?
4. Can participants predict text difficulty and can we predict text difficulty from their eye gaze?
5. Can eye gaze data be used to differentiate between L1 and L2 readers and to derive a measure of text difficulty?

These questions all investigate a sub-component of the overall question of whether eye tracking can be used to make eLearning more effective. In all cases we investigate this for both L1 and L2 readers, whereby we compare the outcomes for two groups. In this way, the investigation is a comparison of first and second language readers.

1.3 Hypotheses

The overall hypothesis is that using eye tracking to analyse the reading and learning behaviour of first and second English language readers can be used to improve reading and learning in eLearning environments. Within each chapter we explain the hypotheses for the investigation carried out in that chapter. However, an overview of these hypotheses is:

1. Layout of text and questions will affect eye gaze and learning outcomes as well as affect L1 and L2 readers in the same way even though there will be differences between the two groups.
2. Eye gaze can provide feedback about implicit learning behaviours, in particular, answering behaviours.
3. Different formats and different levels of text difficulty will affect prediction outcomes of reading comprehension.
4. Eye gaze data can be used to predict text difficulty.

1.4 Methodology

The research carried out for this thesis is based on data collected from user studies where the eye gaze of participants was tracked using video based eye tracking placed at the base of the display monitor. Participants were asked to sit on a chair in front of the monitor and were able to reach the keyboard and mouse. The eye tracker recorded eye gaze and pupil dilation data.

Questionnaire data was also gathered from the participants. There are two main studies that were carried out in the thesis. Each study involves in-depth analysis that is covered by more than one chapter.

1.5 Thesis Outline

The research presented in this thesis is aimed at answering the overarching question of how eye tracking can be used to make eLearning more effective. The thesis is organised in a way that follows the order of the research sub-questions.

Chapter 1: Introduction

This chapter introduces the thesis, the motivation, the research questions that will be explored, hypotheses and outline of the thesis.

Chapter 2: Literature review

The literature review presents an overview of current knowledge of eye gaze analysis and adaptive eLearning. Eye gaze has been used extensively to study the reading process. With this background on the reading process we move to the discussion of using eye gaze to make eLearning environments adaptive to students. This thesis seeks to build upon previous research and enhance the current state of adaptive online learning environments.

Chapter 3: Effect of presentation on reading behaviour

Chapter 3 addresses the first research question of whether eye gaze can be used to find appropriate layouts of reading materials in eLearning environments. This chapter describes a user study that investigated how different sequences of text and assessment questions affect performance outcomes, eye movements, and reading behaviour of L1 and L2 readers. The results from the study show that different presentation sequences induce different performance outcomes, eye movements, and reading behaviour. The presentation sequence impacts participants' ability to accurately perceive their own understanding, in addition to inducing specific reading behaviours, such as thorough reading. The outcomes from this study can be used to influence how students interact with the learning environment as well as how they learn the material.

Chapter 4: Answering questions in eLearning tutorials

A subset of the data presented in Chapter 3 is explored further by investigating the situation where participants are provided with the opportunity to read text whilst answering the questions. The eye movements that occur as a result of this presentation are characterised by transition between the questions and text to find the correct answer, or to reassure the participant that they have the correct answer. We term these eye movements as answer-seeking behaviour, and present a method for measuring and comparing this behaviour. We propose using the degree of answer-seeking

behaviour to measure how question difficulty and as an implicit measure of how difficult a participant finds a tutorial and quiz.

Chapter 5: Effects of presentation on prediction of comprehension

Using the data collected from the user study described in Chapter 3 we explore how presentation formats affect the prediction outcomes of reading comprehension from eye movements. The hypothesis being that the different eye movements caused by the formats will cause different levels of prediction accuracy. The chapter incorporates three components of analysis; the first component builds on previous work of using fuzzy output error (FOE) as an alternative performance function to mean square error (MSE) for training ANNs, as a means of improving reading comprehension predictions. The use of FOE-ANN produced better classification results compared to MSE-ANN. Additionally, the FOE trained ANN outperforms other comparison machine learning techniques. Finally, clustering of the more complex formats revealed reading behaviour properties.

Chapter 6: Effects of text difficulty on prediction of comprehension

Continuing from Chapter 5, this chapter focuses on predicting reading comprehension of text that is shown without comprehension questions. We extend the work by investigating the effect of text difficulty and machine learning techniques on prediction accuracy. Another user study was carried out to collect data from L1 and L2 participants as they read texts with differing degrees of difficulty. The grades of difficulty are based on different levels of readability and conceptual difficulty. We hypothesised that text difficulty and reader type would affect prediction accuracy. We found that neither had a significant effect on the accuracy of the k-nearest neighbour (kNN) classifier used. Whilst this is the case, we did manage to improve the classification accuracy to on average 80% for the L1 group and 73% for the L2 group, which is a substantial improvement from the 44% correct classification obtained in the previous chapter for format C. These results were achieved by using genetic algorithms (GA) for feature selection, which were significantly higher than the results produced when no feature selection is performed.

Chapter 7: Perception and prediction of text difficulty

We investigate prediction of text difficulty from eye gaze using machine learning techniques and compare these to participants' perceptions of difficulty. We show that predictions from eye tracking data are more accurate than the participants' perceptions of both readability and conceptual difficulty. We then show that prediction of participants' perceived ratings of readability and conceptual difficulty from the eye tracking data are significantly better than prediction of the predefined values. This indicates that the eye gaze measures and pupil dilation data may be more aligned with the participants' perceptions of difficulty rather than the predefined difficulty of the text. Further analysis of participants'

perceptions showed that they are poor at predicting predefined text difficulty, especially when the readability and the conceptual difficulty are not the same. Additionally, the text difficulty affected comprehension scores and confidence levels of the L1 readers.

Chapter 8: Deriving text difficulty from eye gaze

The eye tracking data from the user study in Chapter 6 was used to investigate whether L1 and L2 readers' eye gaze are distinct, and whether eye gaze measures can be used to derive text difficulty. The investigation involves clustering eye movement measures from participants using kmeans clustering. Whilst there are clusters of different reading behaviours for different levels of text difficulty, such as skimming and thorough reading, the L1 and L2 groups are not distinct. Instead, there is a tendency for L2 readers to read more thoroughly compared to skimming. The average eye gaze measures for each text were clustered using kmeans. The clusters show that there are distinct reading behaviours and that the average eye gaze measures can be used to rate the texts based on the derived reading difficulty for the L1 and L2 groups. These findings can be used to provide feedback for the purpose of adapting learning material.

Chapter 9: Discussion and Implications

This chapter discusses the results from the preceding chapters, each of which addressed a sub-question of whether eye tracking can be used to make learning more effective in eLearning environments. This overall question is essentially approached from two directions. The first approach is by investigating whether eye tracking can immediately make eLearning environments better suited to the individual learner. The demonstration of these results is through the use of adaptive eLearning whereby the system adapts to the student's understanding levels and perceptions of difficulty. The second approach is the use of historical eye tracking data to make eLearning more effective. This is through the use of eye tracking to provide information to the author of the text and comprehension questions regarding their difficulty. This information can in turn be used to improve the quality of online texts and more accurately define their complexity. To show this we have tied the results from each chapter together in the presentation of a dynamic text selection method to make eLearning environments adaptive.

Chapter 10: Conclusion

The thesis is concluded with a summary of the research findings, and a discussion of the limitations of the research and how it can be improved and extended.

Appendix A: Experiment materials for eye gaze in eLearning environments

The participant information form, consent form, texts and questionnaire used in the experiment explained in Chapter 3.

Appendix B: Experiment materials for adaptive eLearning and digital images

The participant information form, consent form, run sheet, texts and questionnaire used in the experiment explained in Chapter 6.

Appendix C: Dealing with eye gaze data that is imperfect

This appendix explains the post calibration used in the first study.

Appendix D: Reading in distracting digital environments

This appendix outlines preliminary results from a user study on reading in distracting environments.

1.6 Acronyms

The following is a list of acronyms used throughout this document:

ALE	Adaptive eLearning Environment
ANOVA	Analysis of variance
ANN	Artificial Neural Network
FOE	Fuzzy Output Error
FOE-ANN	Feed-forward ANN trained using backpropagation with FOE as the performance function
FMF	FOE Membership Function
HCI	Human Computer Interaction
KNN	k-nearest neighbour
L1	First English language reader
L2	Second English language reader
MANOVA	Multivariate analysis of variance
MCR	Misclassification rate
MOOC	Massive Open Online Course
MSE	Mean Squared Error
MSE-ANN	Feed-forward ANN trained using backpropagation with MSE as the performance function
RF	Random Forest

1.7 Glossary

This is a glossary of the terms used within this thesis.

Cloze question Assessment questions that can be a sentence or paragraph with words removed thereby requiring the reader / participant to fill them in. For example, "*This is an _____ of a cloze question*" were the missing word is example.

eLearning Learning materials presented using digital technology and usually via the Internet or Intranet.

Eye gaze pattern The combination of all the eye gaze points recorded for a participant for each screen showing a text.

Eye gaze point Eye gaze trackers take measurements of where the participants' eye is looking on the screen at regular intervals. Gaze points are used to determine fixations and saccades.

Eye tracker Equipment used to measure eye gaze location.

Fixation When the eye finishes a saccade and stays relatively still to take in visual information for processing.

Flesch-Kincaid Grade Level A readability test that returns the minimum education level (based on the USA education system) needed for the reader to understand the text.

Readability Refers to an explicit measure of text readability as calculated by readability formulae, which typically counts syllables, words, and sentences to determine readability. The readability formula used throughout this thesis is the commonly used metric Flesch-Kincaid Grade Level.

Saccade A rapid movement of the eye as it jumps from one fixation to another. Little to no visual information is taken in during a saccade.

Wattle The online eLearning environment used at the Australian National University. Accessible via: https://wattle.anu.edu.au/

Chapter 2

Literature Survey

This chapter reviews a range of research on the physiology and psychology of reading through to the practical use of human computer interaction (HCI) for eLearning. The central focus of the discussion is on the use of eye tracking to record and analyse eye gaze. Since its invention, eye tracking has proven to be an effective way of analysing human behaviours. This is particularly true for reading, as the eyes have been shown to move in a unique way during reading. These movements consequently reveal much about the underlying cognitive functions involved in reading (review by Rayner (1998)).

Eye tracking is a relatively recent technology (Huey, 1968) but advances in hardware and software for eye tracking have seen an increased popularity of eye tracking for many uses. Initially eye tracking was primarily used for reading analysis, but this technology has proven to be useful in usability testing and HCI (Jacob & Karn, 2003; Poole & Ball, 2005). Reading in a digital environment is now ubiquitous. Concurrently, eLearning technologies have become popular. Given that a primary form of educational material is text and that eye tracking provides an invaluable method of analysing reading behaviour, this raises the question of how eye tracking can be used to make the learning process more effective in eLearning.

The review begins with the discussion of the reading process. This leads to the discussion of how the eye moves during the reading process to reveal the cognitive process of reading. With this background on the reading process we move to the discussion of using these research findings to make eLearning environments adaptive to eye movements.

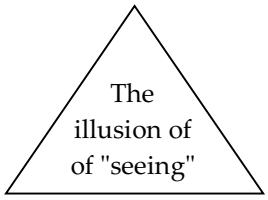
2.1 Attention and Effort

Attention and effort are two cognitive experiences that impact what will be discussed in this thesis. In later sections discussion of the human eye, its functions, and how they apply to reading behaviours will be centred somewhat on attention and effort. This brief discussion about attention and effort is focussed on its applicability to reading.

Attention is focused awareness and can be thought of as the allocation of cognitive resources to deal with some stimuli over others (Buscher et al., 2012; Kahneman, 1973). Selective attention is conscious; an example being that one can move one's head and/or eyes to either look or not look at something. Reading is an example of selective attention where the reader has to consciously choose to allocate attention to the task of looking at the page and reading. On the other hand, involuntary attention occurs when one's attention is allocated to a sudden change in the environment with no conscious control over this allocation. Hearing a loud or surprising noise and turning to see what made it, where it came from, and if it is a threat, is an example of involuntary attention.

There are limited cognitive resources in the human brain and thus limits to human attention. When attention is subjectively allocated it reflects, at least in some way, the person's preference. A person will focus on what they consider most relevant, interesting, or useful in a given situation (Buscher et al., 2012). For humans a reliable measure of attention is eye movement (Henderson, 2003).

2.1.1 Attention and Visual Processing



Due to the anatomy of the human eye, humans do not view scenes in full; they only view parts of it, and only the essential parts in detail. This leads to the intriguing fact that just because the eyes are directed upon a stimulus does not guarantee that all parts of the stimulus are seen; only that which is needed. This observation has been demonstrated many times; perhaps most famously in *The Invisible Gorilla* experiment (Simons & Chabris, 1999). In the experiment, participants watched a short film where two teams, wearing black and white shirts respectively, were passing basketballs between members of their own teams. The players are moving around rapidly, weaving in between one another. Participants are asked to count the number of passes made only by the white team. Halfway through the video a person wearing a gorilla suit crosses the court, thumps their chest and moves on. What they found was that half of the participants did not see the gorilla. This is a demonstration of selective attention where participants are forced to focus on a task and become effectively blind to everything else, termed *inattentional blindness*.

This experiment has been replicated many times, in different settings and confirms that about half of the observers never see an unexpected stimulus. The obvious question is whether the observers actually looked at the gorilla at all. In a study by Memmert (2006) eye tracking was used to record the eye gaze of the

observers in the gorilla experiment. Participants who did not perceive the gorilla had their gaze fixed upon the gorilla for about a second, which is the same time as those who did perceive it. Furthermore, factors such as age and expertise in dealing with certain stimuli are correlated with perceiving unexpected stimuli. So looking at something does not equate to perceiving it; something that has to be kept in mind when analysing eye gaze data.

Inattentional blindness is related to *change blindness*, a phenomenon where humans are seemingly blind to visual change in a stimulus, not always caused by focusing on an absorbing task. Simons and Levin (1998) showed this in a remarkable experiment where an experimenter initiated a conversation with a pedestrian and half way through the conversation the experimenter was replaced by another person. Only half the participants realised that the experimenter had been changed. Even if the eyes are directed upon a stimulus there is no guarantee that all parts will be seen. We have included an example that is designed to show this point. In the triangle[2] figure shown on the previous page most people are not aware that there are is a duplication of the word "of" until it is pointed out to them. The choice of how eye movements are used in real world situations has to take this non-direct relationship into account.

The last point of this subsection is that not all features of a visual stimulus can be reportable. In short, shown the string "aaaaaaa" one could quite easily report that it is a group of a's but most likely not that is it a group of 7 a's without taking a longer to look in order to count them. These phenomena illustrate that the brain does not need to know everything and in fact would not be as efficient if it did. Instead it calculates what it needs only when it is needed. This is why attention and effort are important concepts. Eye gaze provides the remarkable ability to actually identify what is seen in fine detail and to some degree where attention lies. The goal of this research is not to investigate attention; however these concepts must be kept in mind when undertaking reading analysis. Just putting text in front of someone will not guarantee that it is read or even seen.

2.2 The Human Eye

Eyes are the small but complex, organs that enable vision in humans. The human eye is capable of responding to a portion of the electromagnetic spectrum, referred to as visible light. In most cases, reading is possible because our eyes give us the ability to see. This section contains an overview of the physiology of the human eye as an introduction to how humans can take in visual information to be processed by the brain. Finally, how the eye moves to take in information is discussed.

2.2.1 Visual Processing in Humans

The sensory organs collect information about the environment and physical state. The brain processes all of the complicated information that streams in from the sensory organs and then decides what to do with it. Evolutionary processes favour

[2] Taken from (Eagleman, 2011) page 26.

brains that can process the complex information in the most beneficial and effective ways in order to promote survival or even more significantly, reproductive capacity. A large portion of the human brain is used in visual processing, which itself is a complex array of neural processes. Consequently, vision actually occurs in the brain and not in the eyes, which are just there to take in the information (Gehring, 2005).

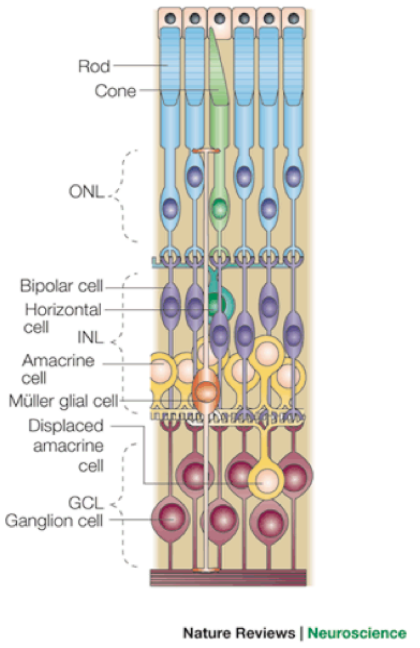


Figure 2.1. The layers of the retina. Image taken from (Dyer & Cepko, 2001).

To illustrate this fact, when humans who have been blind for most of their lives are given surgery to give them eyesight, such as corneal implants, they do not miraculously start "seeing" the world in the way that someone with normal vision from birth does. Instead they have to learn to see; the neural networks in the brain have to be reorganised to provide this ability. If vision was lost early in life or a person is blind from birth, it is believed parts of the visual processing system never completely develop to the extent of an individual with unimpaired vision (Cohen et al., 1997). This highlights the fascinating point that some brain function is dependent upon input from the sensory organs. From an evolutionary point of view it is very likely that the eyes came before the brain (Gehring, 2005). Intuitively this is because there is no point having such an intricate information-processing unit if it has no information to process.

When humans read, the eyes are the starting point of the process, (excluding reading Braille). Given that the eye is a critical part of reading we discuss further how letters on a piece of paper or on an electronic display make their way into the human brain for interpretation. Firstly, light enters the eye and is passed through the cornea and projected onto the retina, which is a light sensitive layer of tissue at the back of the eye. The cornea is a transparent covering of the iris and pupil at the front of the eye. The iris dilates and constricts the pupil to regulate the amount of light that enters the eye and the lens focuses this light onto the retina (Burton et al., 2009).

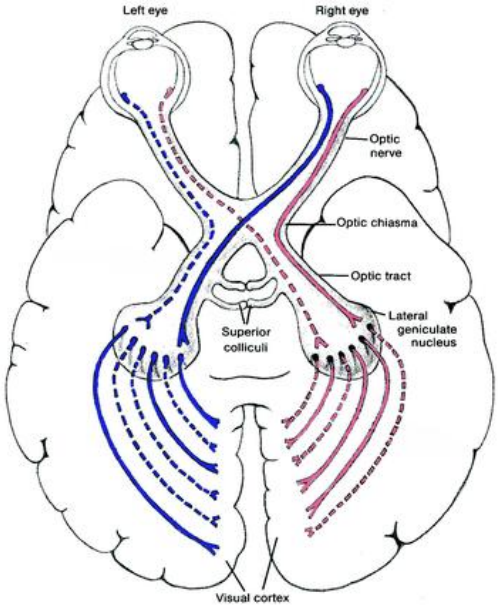


Figure 2.2. The optic tract in the human brain. Image taken from https://senseofvision.wikispaces.com/ (Last accessed: 8th November 2015).

The retina transforms light that enters the eye into an electrical signal using photoreceptors. The retina is a complex multilayer structure, as shown in Figure 2.1 and the following is an overview of how the light that enters the eye then makes its way to the brain. The photoreceptor layer in the human eye contains two types of light receptors: rods and cones, shown at the top of Figure 2.1. Rods are responsible for vision in low level light and are used in peripheral vision, cones are responsible for vision in higher levels of lights and for the ability to see colour. When a rod or cone absorbs light energy, an electrical signal is generated. These signals are passed through a layer of bipolar cells onto ganglion cells that integrate the electrical signals from many photoreceptors. The resulting signals are transmitted through the long axons of the ganglion cells that bundle together to form the optic nerve. The optic nerve transmits the signal to the brain via the optic chiasma where information from the left half of each visual field goes to the right hemisphere and similarly for the right (Schwarz & Schmückle, 2002).

The optic tract projects to three major subcortical structures (Schwarz & Schmückle, 2002) that make use of the visual information for different purposes. These structures are: the pretectum which controls pupillary reflexes; the superior colliculus which controls saccadic eye movements; and, the lateral geniculate nucleus (LGN), which is a thalamic nuclei and is the major relay for input to the visual cortex (Schwarz & Schmückle, 2002).

The sensory information from the eyes first goes through the LGN and then proceeds to the primary visual cortex in the occipital lobe at the back of the brain. Most sensory and motor information only reaches the cortex via the thalamus; there are exceptions to this, such as smell. Visual information then flows through the hierarchy of the visual cortex, where V1 is the point of entry of the visual sensory

information that is then passed through to V2, and so on to V5[3]. Complexity of the neural representation increases as the information flows through the cortical hierarchy (Dehaene, 2009).

2.2.2 Types of Vision

The eye is capable of two types of vision, peripheral and detailed. Peripheral vision is hazy and occurs outside of the centre of gaze. Whilst peripheral vision is not very good at distinguishing colours and shapes, it is sensitive at detecting movements, and mostly used to gather information about the present surroundings. The human brain prioritises the information to give attention only to what it somehow deems important. For example, Itti and Baldi (2009) found that humans orient their attention and gaze toward surprising stimuli in the context of watching television. The reason for this orientation towards a stimulus is so that detailed vision can be used to examine the stimuli further and to manage the limited resources at hand in order that the most important stimuli are tended to first. The peripheral region of the visual field encompasses the whole retina apart from the foveal and parafoveal regions.

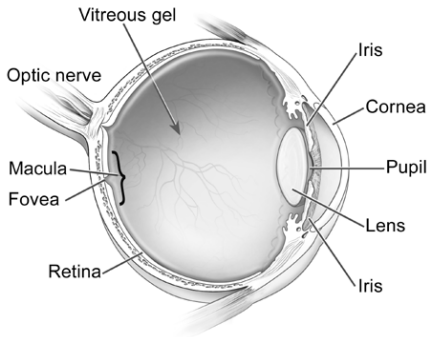


Figure 2.3. Diagram of the anatomy of the eye. Image taken from: https://nei.nih.gov/sites/default/files/nehep-images/eyediagram.gif (Last accessed: 29th January 2016)

Detailed vision is handled by the fovea and to some degree, the parafovea. The fovea is the small central region of the retina that is sensitive to fine detail. The fovea only sees the central 2° of the visual field (Rayner & Bertera, 1979), and comprises of a region of only cone photoreceptors. The parafovea extends 10° of the visual field around the centre of gaze and provides less detailed visual information than the fovea but more than the periphery (Rayner & Bertera, 1979). Whilst the fovea takes up less than 1% of the retina, the processing of this information accounts for over 50% of the activity of the visual cortex in the brain (Mason & Kandel, 1991). The fovea is necessary in humans for reading (Rayner & Bertera, 1979). Since only a very small part of the eye is capable of seeing in detail, the eye is constantly on the move to assimilate information about the visual environment. How and why the eye moves in the way that it does will be discussed in the following subsection.

[3] also known as the middle temporal area (MT)

2.2.3 Types of Eye Movements

Eye movement is somewhat sporadic and complex, with the eyes moving at high velocity before stopping for a period to take in information before moving on again. Louis Emile Javal first described this process in 1879 by direct observation. It was not until almost a century later that Edmund Huey developed the first eye tracker (Huey, 1968). To explain this phenomenon, we must consider detailed vision again. The foveal region is where 2° of visual acuity extends across the fixation point (Rayner, 1998; Rayner & Bertera, 1979; Underwood & Batt, 1996). The parafoveal region is just outside the foveal region and it comprises 5° on either side of the fixation point (Rayner & Bertera, 1979). The peripheral region is the rest of the visual field.

Due to this limited area of detailed vision, the eyes are constantly on the move so that the fovea can be oriented upon different parts of the environment. The visual information is taken in when the eye has been reoriented and is relatively still; this is termed a fixation. The rapid jumps between fixations are termed saccades, and little to no visual information is taken in then (Rayner, 1998). Humans, therefore, do not view an image of the environment or scene as a whole, instead it is viewed in parts and in differing detail depending on where the centre of gaze is oriented (Henderson, 2003). The attention given to certain stimuli can be quite dependent upon the reasons for looking at them. This was shown by Alfred Yarbus in his early work on eye movements in scene perception (Yarbus, 1967). In his work, he showed that an individual's eye gaze was dependent upon the question they were asked.

Gaze control is influenced by many factors, including information about the environment or stimulus and several cognitive systems (Henderson, 2003). This includes past memories of the scene, whether the individual is searching or memorising the scene, and its spatial and semantic properties. This type of gaze control is said to be knowledge-driven (Henderson, 2003). More precisely, the spatial and semantic properties of the scene refer to the fact that you can anticipate where a particular object will be found. For example, you would expect to see a stapler on a desk and not on the floor. Further, there is a difference in the distribution of fixations and their durations based on whether the individual is trying to memorise or scan the scene (Henderson, 2003). Short, sparse and highly distributed fixations are observed for scanning and frequent, long and clustered fixations are observed for memorisation.

The control of when and where a fixation will occur involves coordination of information from several areas of the brain. In a general sense, we can say that it is the oculomotor system that oversees the process of directing the fovea to particular regions of interest. To accomplish this task, six different control systems are involved, which are grouped into two classes of gaze control mechanisms; intentional gaze shifting mechanisms and reflex gaze stabilizing mechanisms (Schwarz & Schmückle, 2002).

The gaze shifting mechanisms include saccadic movements, smooth pursuit movement, and vergence movement. Saccadic eye movements are the rapid ballistic

movements that move the fovea to another point of fixation (Purves et al., 2001). Pursuit eye movements keep the fovea on a moving target and are slower than saccades (Purves et al., 2001). Finally, vergence eye movements change the orientation of the eyes in accordance with the distance from which a target is being viewed. That is, the eyes rotate toward the nose when looking a close target.

The gaze stabilising mechanisms include vestibular eye movements and optokinetic eye movements (Schwarz & Schmückle, 2002). Vestibular eye movements are rotations of the eye produced in order to maintain vision in the same direction when there are head and body movements (Purves et al., 2001). Optokinetic eye movements are the combination of saccade and pursuit eye movements. These eye movements are seen when the observed target is moving fast across the visual field.

The fixations are characterised by the relative stillness of the eye to take in visual information. Although fixations are characterised by suppression of gaze shifting eye movements, the eye actually never stays completely still. This is due to three types of small eye movements: tremors, drifts, and microsaccades (Martinez-Conde, 2006). The eye constantly tremors; these are the smallest of any eye movements and are hard to record (Martinez-Conde, 2006). Drifts and microsaccades are larger movements, but are still quite small. Drifts appear to be random and caused by instability of the oculomotor system (Martinez-Conde, 2006). Microsaccades, similar to saccades, are jerking motions. They are differentiated from saccades as being the movements that happen whilst you are fixating. These small movements are usually regarded as noise as it is the larger eye movements that are of importance, especially in reading.

The main types of eye movements to consider when investigating reading are saccades, fixations and regressions. Saccades[4] are high velocity ballistic movements of the eyes. At the end of a saccade the eye stays relatively still for a period of time; (a fixation) and is the only point during reading that visual information is encoded. Since visual information is taken in during fixations, there is often a focus on analysis of fixations, in particular the duration and location. No visual information is taken in during saccades under normal reading conditions (Underwood & Batt, 1996). However, they cannot be discounted, as lexical processing occurs during saccades (Yatabe et al., 2009) and that during long saccades, readers perform more lexical processing than during short saccades.

Saccades are motor movements and therefore require time to plan and execute. Saccade latency is the period associated with making a saccade (Rayner, 1998). Saccade latency still exists even if uncertainty about where and when to move the eyes is eliminated so saccade programming is believed to be done in parallel with comprehension processes during reading (Rayner, 1998). Engbert and Kleigl (2001) found that initiations of saccades are not completely driven by lexical processing and that in fact saccades can be autonomous (with foveal inhibition).

[4] Saccade is French for jump

2.2.4 Pupillometry

Movements are not the only source of information about cognitive processes that can be gathered from eye tracking. Pupil dilation provides abundant information about the cognitive state of the person. Experiments have shown that pupil size correlates to cognitive load or effort, where the pupil dilates further as effort increases and constricts as it decreases (Kahneman & Beatty, 1966). Pupil dilation is a good indicator of effort (Kahneman, 1973; Pomplun & Sunkara, 2003) and as a result pupil dilation has become widely used as an involuntary indicator of mental effort and cognitive load.

Whilst pupil dilation as an indicator for mental effort was first described by Hess and Polt (1964), it has been greatly studied after popularisation by (Kahneman & Beatty, 1966). The correlation between cognitive load and pupil dilation has been confirmed in many contexts since including assessing task difficulty in response preparation (Moresi et al., 2008), software development (Fritz et al., 2014), listening comprehension (Engelhardt et al., 2010; Zekveld et al., 2014), as well as to detect decision to change task (Katidioti et al., 2014) and difficulties in making decision (Satterthwaite et al., 2007).

The rest of this section is concerned with outlining how pupil response can be used as an indicator of learning and in terms of HCI. Work on pupillary response as an index of learning (Sibley et al., 2011) show that pupil diameter drops as participants learn tasks reflecting decreased effort required in performing that task and increases at the beginning of another level of difficulty. The implication is that pupillary response could be used to assess whether an individual has learned a task sufficiently or if they need more training. Further, the pupillary response could be used to speed up and slow down training procedures by judging the rate at which the individual is learning. This is significant due to the implications of the use of pupil response in adaptive learning and training environments - a major aspect of this research.

Pupil response could therefore be used as an accurate indicator of task difficulty (Iqbal & Bailey, 2004; Pomplun & Sunkara, 2003; Zekveld et al., 2014). However, the averaged value of pupil response ignores the effects of the fluctuations of pupil dilation due to lower and higher loads of mental effort within tasks (Iqbal & Bailey, 2004). Pupil dilation can therefore be used to measure the changes in workload during a task (Iqbal et al., 2005). The pupil is seen to increase is dilation during a subtask but decrease when the subtask is finished. These results have interesting implications on our current research, as averaging of pupil response cannot be considered a viable measure in assessments of reading comprehension. This differs from the eye movements where commonly averaged numbers of fixations, saccade lengths, etc., are used as measures.

2.3 Reading

Language is one of the important characteristics that have set humans apart from any other organism on the planet. This extensive communication device has enabled

intricate social behaviour that has seen the human race flourish. Along with tool making and teaching, humans have become masters of invention. Amongst the greatest of these inventions is that of writing and reading language, and is indeed a very important part of human behaviour.

When a human reads, the eyes quickly and almost unconsciously move to acquire the text on display so that the brain can piece them all together and make logical sense out of it. Reading requires numerous cognitive processes to work together including visual information processing, word recognition, attention, language processing, and oculomotor control.

Up to 30% of Australian children have difficulty learning to read even with normal schooling (Burton et al., 2009). The process of learning to read is less natural than learning to speak, as written language is a much later addition to spoken language. Reading requires complex interpretation of symbols in order to derive meaning from them. This is termed comprehension and is the main objective of reading. Proficient readers quickly and unconsciously recognise words; if a word is not familiar it requires more cognitive processing in order to discern the meaning of the word. Reading, therefore, requires continuous education to ensure this processing time is minimised.

2.3.1 Human Language

Language is a complex communication system. Humans created written language as a way to communicate through time and space. Before reviewing written language, which is the foundation of reading, there is a short discussion about human language to present the foundations of language. A language system is made up of a set of symbols, sounds, meanings (semantics), rules (syntax) and interpretation (pragmatics). Language can be conceptualised in a hierarchical structure, consisting of basic elements at the lowest level called phonemes. Phonemes are the smallest elements of sound that form coherent speech such as how vowels and consonants are pronounced in English. Phonemes make up morphemes, which are the smallest units of meanings, e.g. words. Morphemes make up phrases that in turn combine with more words to make up sentences. The rules that govern these combinations are called syntax. Syntax is a part of grammar, which is the system of generating correctly structured expressions. Semantics is used alongside syntax to understand meaning of expressions (Burton et al., 2009). Semantics are the rules behind the meanings of the morphemes, words, phrases and sentences. Language is generative and diverse, allowing humans to express themselves in a potentially infinite number of ways. From a finite set of elements that make up language (phonemes) a very large number of words, phrases and sentences can be generated. Human languages are forever growing, changing and evolving to humans needs.

2.3.2 Written Language

Human language developed at least 45,000 years ago whereas written language only occurred about 3500 BC (Barton, 2007; Burton et al., 2009). Children very easily

learn to speak a language but must be taught how to read and write. Writing systems were built on to spoken language and have evolved according to the neural network capabilities of the human brain (Dehaene, 2009). More precisely the human brain allows neuronal recycling so that new activities can be learnt by humans to deal with new or changing situations. This means that new activities are constrained by the limits of the brain structures. Reading and written language are as mentioned very recent in human history. Neither our eyes nor brains have evolved to read or write, instead it is our reading and writing system that have developed according to the constraints imposed by our visual system and brains' capabilities (Dehaene, 2009).

Writing was invented independently in three different areas; the Fertile Crescent of Mesopotamia and Egypt, China, and pre-Columbian America (Barton, 2007). Cuneiform is the earliest known writing system, which can be dated back to about 3500 BC in Mesopotamia. There are now numerous writing systems, however, they share much in common because they are limited by the same brain structures (Dehaene, 2009). Characters from all writing systems have visual features that rely on basic shapes that provide optimal contrast of contours on the retina.

2.3.3 Eye Movement during Reading

We have already discussed the types of eye movements in the preceding sections. This section discusses the movements in the context of reading and is broken down into three subsections; types of eye movements observed during reading (saccades, fixations and regressions) and the reasons for why these movements are observed, perceptual span and parafoveal preview, and finally, where and when fixations and saccades occur.

2.3.3.1 Reading and the Fovea: Why and What

The human brain and eyes have not evolved to read. Instead written language has been constrained by the anatomy of our eyes and brains (Dehaene, 2009). The fovea, which is essential in the reading process (Rayner & Bertera, 1979), developed an extremely long time before language in general was even conceived. It is the fovea and paraforveal regions that are critical for reading (Rayner & Bertera, 1979; Rayner & McConkie, 1976). Masking foveal and parafoveal vision have showed this experimentally (Rayner & Bertera, 1979). The results showed that although participants could see words, they would get the words wrong in the sentence. Longer fixation times were recorded when vision was masked and reading time increased. The larger the mask, the greater the percentage of words incorrectly identified. Even though the participants were aware that words were in the parafovea and peripheral view they could not report what the words were. Masking of the fovea resulted in severe reading difficulties compared to masking of only the parafovea. Rayner & McConkie (1976) concluded that information necessary for meaningful identification of a word is obtained from the fovea and near parafovea. Additionally, information such as that used to guide the eye to the next location in the text is collected by the parafovea.

The further a word is presented from the fovea, the greater the decrease in ability to identify that word (Rayner & McConkie, 1976). The eye must move frequently in order to orient the fovea for highly detailed vision in different places to assimilate information. This is why the eyes do not move in a smooth pattern taking in a constant amount of information. Contrary to what may be assumed, the eyes do not necessarily move from left to right, line by line. Therefore the eye movements generated during the reading process can be quantified into measures that can be used to infer cognitive processes that are required for reading. The variability of these measures then reflects real time processing (Rayner, 1998).

Generally when reading English, fixation duration is around 200-300 milliseconds, with a range of 100-500 milliseconds and saccadic movement is between 1 and 15 characters with an average of 7-9 characters (Liversedge & Findlay, 2000). The majority of saccades are to transport the eye forward in the text when reading English, however, a proficient reader exhibits backward saccades to previously read words or lines about 10-15% of the time (Rayner, 1998). These backward saccades are termed regressions. Short regressions can occur within words or a few words back and may be due to problems in processing the currently fixated word, overshoots in saccades, or oculomotor errors (Rayner, 1998). However, longer regressions occur because of comprehension difficulties, so the reader tends to send their eyes back to the part of the text that caused the difficulty (Rayner, 1998). Frazier and Rayner (1982) demonstrated this elegantly by showing that when readers were subjected to garden-path sentences regressions could be systematically induced. A garden path sentence[5] is designed in such a way to mislead the reader into incorrectly interpret the sentence though the sentence is grammatically correct. They are used in psycholinguistics to illustrate the fact that during reading, humans process language one word at a time. Readers make regressions back to the point of difficulty and then re-interpret the sentence (Frazier & Rayner, 1982).

2.3.3.2 Reading and the Parafovea: Perceptual Span and Parafoveal Preview

As mentioned earlier, the parafoveal region includes 10° of the visual field around the point of fixation. Although the parafovea is not responsible for fine detail vision it plays an important role in reading. In this subsection we will discuss two of these roles in regards to reading: perceptual span and the effect of parafoveal preview.

The perceptual span in reading is the region of the visual field where visual information is encoded. In previous sections we have distinguished between the areas of the visual field: foveal, parafoveal and peripheral. It is established that the foveal region is where fine detail vision is encoded, but the parafoveal region also encodes visual information that is useful, though in less detail then the foveal region. Information from the parafoveal region is gathered on most fixations (Rayner, 1998). An approach used to assess the perceptual span in reading is using a

[5] An example of a garden path sentence is "The old man the boat."

gaze-contingency paradigm[6] technique called the *moving window paradigm* (McConkie & Rayner, 1975). In this paradigm, the experimenter controls the visual information available at each fixation as the text within a defined window is distorted in some way (see Figure 2.4). Although the reader is free to look anywhere, the letters outside of a window spanning a given number of character spaces are distorted. It is possible to determine the perceptual span of the reader by changing the size of the window and making its location dependent on where the reader is looking. A notable observation from experiments using this technique is that given the correct window size and properly functioning equipment, participants are not aware of the changed text outside of the window (Pollatsek & Rayner, 2009).

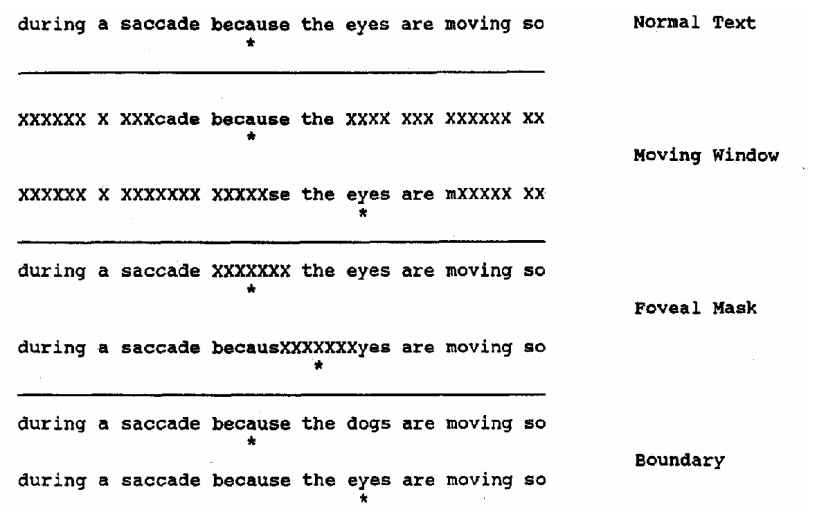


Figure 2.4. Figure 2 from (Rayner, 1998) examples of the moving window paradigm

Studies have shown that when the window extends 14-15 character spaces to the right of the fixation point readers can read alphabetic text, such as English, without disturbance (McConkie & Rayner, 1975; Rayner & Bertera, 1979). The same has been shown for the window extending just 3-4 characters spaces to the left of the fixation point (Rayner & McConkie, 1976). This observation shows that there is an asymmetry to the perceptual span, which is most likely language specific as in English readers read from left to right and in other languages where the direction of reading is different, so too is the perceptual span (Rayner, 1998; Reichle et al., 2003). However it is important to note that word encoding does not occur outside of 7-8 characters to the right of the fixation, only information about letter shape and word length is taken in (Rayner, 1998; Reichle et al., 2003).

This leads to the next point of discussion: parafoveal preview. Studies have shown that even before a word is fixated upon, orthographic and phonological processing already has begun (Rayner, 1998; Reichle et al., 2003). The parafoveal preview of a word can therefore decrease subsequent fixation duration of a word. This is seen most prominently by the observation that even when a word has not been fixated upon it is still processed (Rayner, 1998). This was demonstrated in a

[6] Gaze-contingency techniques are where the display on a computer screen is changed as function of where the viewer is looking.

study where an initial group of participants had their eye gaze monitored as they read text. Another group read the same text with the words that the first group skipped removed. The study showed that the second group had difficulty understanding the modified text (Rayner, 1998). Word skipping can therefore be induced by parafoveal information that allows the word to be identified when in the parafoveal visual region.

The effects of parafoveal preview are seen most prominently in the observation that predictable words are skipped more than unpredictable words and that short function words are skipped more than content words. The effect of the parafoveal preview is diminished when processing of the fixated word is difficult. There is some support of this in that durations of the fixations before and after a skip are longer (Rayner, 1998; Reichle et al., 2003).

2.3.3.3 Eye movements: Where and When

Where the eyes move is based largely on low-level visual information such as word length and spacing between words. When the eyes move, the movement is believed to be largely based upon lexical processing. That is, until lexical processing has concluded the eye will not move to the next word. Discussion in the previous subsection has already introduced the fact that fixations are not evenly distributed over the words in the text read. Firstly, not all words are fixated upon during reading, with many being skipped. Secondly, some words are fixated upon more than once. Interestingly, content words are fixated upon 85% of the time and function words are only fixated upon about 35% of the time (Rayner, 1998). Rayner and McConkie (1976) found that there is a relationship between the probability of fixating upon a word and its length, so as word length increases so too does the probability of fixation. Since function words are usually short words this is one explanation for why they are fixated upon less than content words.

Word length is also useful in determining where in the word a fixation will occur. The first fixation on a word has been shown to, in general, be between the beginning and the middle of the word (McConkie et al., 1988, 1989; O'Regan, 1981). This was termed the *preferred viewing location*. Later, the *optimal viewing position* of a word was defined as the location in a word at which recognition time is minimised. The optimal viewing position is closer to the centre of the word. The likelihood of re-fixation increases as the fixations become further away from the optimal viewing position, which is termed the re-fixation effect (O'Regan, 1984). However, as word length increases, first fixations tend to occur near to the beginning of the word and a second fixation will occur toward the end of the word.

The length of the fixated word and the word to the right of the fixation influence saccade length. The fixation position on a word is dependent on the fixation position on the currently fixated word. When readers have no information about where the spaces are between upcoming words, saccade length decreases and reading is slowed considerably (McConkie & Rayner, 1975).

Linguistic processing also has a great bearing on how long a fixation is on a particular word. Evidence for this is that low frequency words are fixated upon for

longer that high frequency words. In addition to this, numerous studies have shown that predictable words are more likely to be skipped than unpredictable words and fixations are more likely to occur on low frequency words (Rayner, 1998). Furthermore, longer fixations have been observed for misspelled words (Rayner, 1998), which in essence is a product of word frequency and word predictability.

The effect of the text and the presentation of the text can be observed in eye movements. Factors such as the quality of the print, line length, and letter spacing influence eye movements (Rayner, 1998). The format in which text is presented in terms of length can have effect on eye movements observed during reading. Sharmin et al. (2012) showed that by altering text presentation length from paragraphs, to individual sentences and to line-by-line presentation of text that fit a computer screen, made a considerable effect on fixation duration, number of fixations per minute and number of regressions. Furthermore, as text becomes more difficult to understand, an increase in fixation duration is seen along with decreases in saccade length and increased frequency of regressions (Rayner et al., 2006).

2.3.4 Eye Movement Measures

Eye tracking produces a considerable amount of data. As established in the previous subsections, the basic units of analysis when considering eye gaze data are fixations and saccades. Converting eye gaze data into fixation points first reduces the eye gaze data. However, the fixation data is still quite large and fixations and saccades, alone, tell us very little about the nature of reading behaviour. Eye movement measures are a way to reduce the amount of data to investigate the nature of reading. There are many commonly used measures, which are described in Table 2.1. Typically word-based measures are used, especially when investigating lexical access and syntactic parsing. Eye movement measures are used to identify different patterns in the data to tell us different facts about the data. An eye movement measure is a form of description about the fixations and saccades that are observed for given section of text, which may be a word, sentence, paragraph, etc.

When considering global text processing, recognition of individual words that have been read is not appropriate for assessing comprehension. This is because global text comprehension not only involves the assimilation of words in individual sentences to form a conceptual meaning and build relationships between sentences in the text. Listed in Table 2.1 are eye movement measures that can possibly be used in assessing global text processing. These measures include eye movement matrices and regional gaze duration.

The majority of the research using eye gaze to analyse reading behaviour is on small units of text such as words, phrases, or sentences for the purpose of studying lexical access and syntactic parsing (Hyona et al., 2003). Models such as the E-Z reader model (Reichle et al., 1998; Reichle et al., 2006; Reichle et al., 1999, 2003, 2012; Reichle et al., 2009) are used for local processing analysis. However, in real life situations, in particular HCI situations, often the text being read is quite a lot longer, being paragraphs, articles, books, etc. When analysing the comprehension of such long pieces of text, global text processing must be assessed. Global text processing is

where relationships are identified and symbolised in constructing a mental model of the meaning of the text (Hyona et al., 2003). The relationship not only spans sentences, but also paragraphs. This type of text processing requires recall of not just working memory but short-term memory, and often long term memory.

Table 2.1. Eye movement measures

Measure	Definition
Number of fixations	The number of fixations can be affected by the reading behaviour, text difficulty, and reading skill (Rayner, 1998).
Average fixation duration	The sum of the durations of all fixations on a paragraph divided by the number of fixations on that paragraph. This measure has been used to predict reading comprehension (Underwood et al., 1990).
Average forward saccade length	The average length of the left to right saccades. Saccade length is affected by characteristics of the text (Rayner, 1998).
Regression ratio	The number of regressions divided by the total number of saccades on a paragraph. There is evidence that when reading harder text more regressions are observed (Rayner et al., 2006).
Average Regression Length	The average length of regressions. Regressions are affected by text complexity and inconsistencies in the text (Rayner et al., 2006).
Coherently read text length	The length of text in characters that has been read without skipping any text in between according to the reading detection algorithm (Buscher et al., 2012). Used in assessing whether users find a piece of text relevant or irrelevant. However, the assumption for this thesis is that the longer the length of text read the more likely that the text has been understood.
Thorough reading ratio	The length of text that has been detected as read by a reading detection algorithm divided by the length of read or skimmed text (Buscher et al., 2012).
Total fixation time	Sum of all fixations on complete text. This measure is useful in global text processing analysis because this measures immediate as well as delayed effects of comprehension (Hyona et al., 2003).

2.3.5 Reading Comprehension

Reading comprehension is the capacity to make sense from written language. In alphabetic languages such as English, this requires assimilating symbols to make them into words and then sentences, and deducing meaning of the bigger picture. Simply looking at the alphabetic symbols on a page, electronic display, packaging, etc. involves little cognitive ability as our eyes have evolved to take in fine detail

information (see section 2.2). Interpreting these symbols requires somewhat more skill. This involves knowledge of the alphabet system and language so that individual letters can be recognised as words, that is, identification of the orthographic form of a word and lexical processing of that word to identify phonological and/or semantic forms. However, the individual words and letters often have little importance on their own; it is their combination that carries value including meaning (Snow, 2002; Underwood & Batt, 1996). Reading for the most part requires making inferences both locally and globally in the text and conceptualising the ideas expressed in the text. The reader then incorporates their knowledge and experience to build a model of the ideas being expressed within the text (Kintsch & Rawson, 2005). This is why reading comprehension is dependent on many different variables, not just knowledge and experience but also motivation and context.

Reading comprehension is a skill that must be taught and requires constant education. It is a skill that requires making relationships between not only the words in a sentence but also in the multiple sentences and paragraphs by concurrently finding and forming meaning from what has been read (Snow, 2002; Underwood & Batt, 1996). When understanding language, we integrate ideas in the text and form a mental model that is an abstraction of the conglomeration of ideas (Kintsch & Rawson, 2005; Underwood & Batt, 1996). The actual text read is not remembered verbatim, it is the ideas and constructed representation that are remembered (Bransford & Franks, 1971; Kintsch & Rawson, 2005; Underwood & Batt, 1996).

One of the main parts of reading comprehension is inference. This is often referred to in terms of inferring co-reference but can also be inferring meaning from context. Words can often have multiple meanings (lexical ambiguity e.g. bank, right) and multiple words that are spelt differently and mean different things have the same sound (phonological ambiguity, e.g. homophones such as to, too and two). Furthermore, phrases and sentences can have different meanings (syntactic ambiguity). An example of this is the sentence "Visiting relatives can be boring". This sentence is ambiguous because it can both be interpreted as relatives that have come to visit are boring or the act of going to visit relatives is boring. .

In general, semantically and phonologically ambiguous words are resolved by context. This is more prominent with semantic ambiguity where the meaning of the word is derived from the context from the sentence, or prior sentences. Of course this may not always be the case, where both meanings may be equally probable, and it has been found that the more prominent meaning of the word is usually inferred. If the inference is wrong then the reader often directs his eyes back to the word to re-interpret (Frazier & Rayner, 1982).

Reading comprehension involves several levels of processing (Kintsch & Rawson, 2005). The first and most basic level has just been described; this is the linguistic level where word recognition and parsing occurs. The next level is to derive meaning from the text, the semantic analysis of the text, which requires inference. A classic example of inference is anaphoric co-reference, where basic

inference is required to resolve quite simple meaning. An example of this is the sentence: “The carnation won a prize. It was the best flower in the show.” (Underwood & Batt, 1996). These are two simple sentences that are easy to understand. However, the two sentences only make sense in combination when the conclusion is made about the pronoun “it” refers to the carnation in the first sentence. Resolution of anaphoric references is fundamental in sentence comprehension.

The example of an anaphor given above is simple, however, there are many factors that can affect how anaphors are resolved, such as linguistic, semantic, and pragmatic information. Linguistic factors include context such as gender information, for example, from (Kintsch & Rawson, 2005): "Leonard handed Michael a sandwich. Then he passed Carla an apple. Then Carla passed him an apple.” In this sentence is it more likely that the pronoun “he” in the second sentence is resolved to refer to Leonard and the pronoun “him” in the last sentence is resolved to refer to Michael?

Semantic factors that affect anaphor resolution include implicit causality of verbs. Take for example the two sentences: “John questioned Chris because he wanted the correct answers. John praised Chris because he knew the correct answers.” Here the verb question in the first sentence implies that “he” refers to John, in contrast with the second sentence where the verb praise implies that “he” refers to Chris. Another factor is pragmatic plausibility, where contextual information has implications on the interpretation of anaphors. For example: “Scott stood watching while Henry fell down some stairs. He ran for a doctor.” Here, the referent “he” is most likely to be Scott because it is less plausible that Henry, who has just fallen down the stairs, is able to run to find a doctor.

These are basic relationships and are local in effect. Sections of text are often related as well, which requires recognition of global interrelationships. It is the combination of the local and global relationships within the text that represent the meaning of the text. This requires identifying important themes or topics in a text. The ideas expressed in the text are tied together, which results in a meaningful interpretation of several sentences. This involves abstraction of ideas and their integration into an overall mental model of the text. So ideas are not kept in isolation, but integrated together to form general meaning. In fact, the original piece of text is most likely not remembered verbatim and instead the abstract ideas are remembered in the mental model.

This has elegantly been shown experimentally by Bransford and Franks (1971) who presented several sentences to participants. Following the first presentation, another set of sentences is shown to the participants and they are asked if the sentences in the second presentation had been presented in the first presentation as well. The interesting part of this study is that the first set of sentences only contained singular ideas that where semantically related so could be linked together to form an overall coherent idea. In the presentation of the second set of sentences, the participants were more likely to say that they had been presented with sentences that incorporated all of the ideas, even though they never had been.

2.3.5.1 Testing Comprehension

Reading comprehension is not a straightforward quantity or characteristic that can be measured because it is not an explicit process that the brain performs. It is the product of a number of cognitive processes, including visual processing, lexical processing, linguistic and semantic processing, high-level integration of concepts, memory, and reasoning. It is the products of these processes that observed and from which assumptions about reading comprehension must be made.

Unfortunately, this is not the only reason why reading comprehension is hard to measure. A person's understanding of text is reliant on many factors including their overall cognitive capabilities (intelligence), motivation, knowledge, experiences, and even the purpose of reading (Snow, 2002). Variations in any of these factors can attribute to different measurements of comprehension. Both the text and the type of assessment should be considered when making conclusions about reading comprehension (Fletcher, 2006). This section begins with an example to highlight a critical problem that must be addressed when testing for comprehension.

In a study that examined speed-reading, normal readers and speed-readers were required to read a text and then answer a reading comprehension test. It was found that normal readers had a higher percentage of questions answered correctly compared to the speed-readers, 72% to 68% respectively (Crowder & Wagner, 1992). However, the interesting part of this example is that the same comprehension test was given to individuals who had not even read the text. These individuals managed to on average correctly answer 57% of questions on the test. Guessing and common sense was enough for the participants to pass the comprehension test. Although the point of this study was to show that speed-readers underperform in reading comprehension compared to normal readers, this example serves to highlight the necessity of appropriately testing comprehension. The remainder of this section will discuss current methods for measuring reading comprehension as well as the advantages and disadvantages of each method.

Typical informal[7] methods for assessing reading comprehension include: question-answer tests, recall procedures, oral passage reading measures, and cloze[8] techniques (Fuchs et al., 1988). These are all relatively simple to construct and can be tailored to the purpose of the teacher or experimenter. Usually one method of assessment (e.g. multiple choice, cloze, etc.) is used in assessing reading comprehension making the assessments one-dimensional. However, this is often not sufficient to accurately assess comprehension in reading. Cutting and Scarborough (2006) showed that results from different assessment methods produced differing levels of comprehension for the same material. Their results suggest that commonly used tests of reading comprehension may not require the same cognitive processes to complete. Keenan et al. (2008) demonstrated similar results in testing different

[7] As opposed to standardised tests, informal methods can be constructed by a teacher or experimenter within their own bounds and are more flexible in what is tested.

[8] Cloze techniques refer to methods where words are deleted from text and replaced with blanks. Students then insert words into the blank spaces to complete and construct meaning from the text. This procedure can be used as a diagnostic reading assessment technique. Also, note that this is used as a method of calculating predictability of words in context.

standard reading comprehension tests and found that even though the tests were supposedly measuring the same outcome, the results from each test were only modestly correlated to the others. Furthermore, Francis et al. (2006) showed that any single, one-dimensional attempt to assess reading comprehension is inherently imperfect. This is because only parts of the comprehension process can be observed from which conclusions are made, so inherently these conclusions may not be representative of the true quality of the comprehension. In brief, this illustrates that the method of assessment used to evaluate reading comprehension can be a determinant of the conclusions drawn.

Further to this point, the difficulty of text as well as its characteristics (e.g. semantic, syntactic) can play a large role in whether an individual will understand it or not. Therefore, the text plays a key role in determining level of comprehension (Fletcher, 2006). This is demonstrated in the eye movement study by (Rayner et al., 2006) and the attempt to minimise the role of certain text characteristics in Francis et al. (2006). It is important to understand that text variability is a determinant of the inferences made about reading comprehension.

2.3.5.1.1 Standardised Reading Comprehension Tests

A common standardised test for reading comprehension is the *Critical Reading* section of the Stanford Achievement Test Series (SAT). The SAT is a widely used achievement test set in the American schooling system. It is used to measure academic knowledge of elementary and secondary school students. The whole set of test covers subjects including mathematics, science, social science, spelling, listening comprehension, and importantly for this analysis, reading comprehension.

Another standardised achievement test is the Wechsler Individual Achievement Test Second Edition (WIAT-II). This test is different from the SAT in that is can be used to assess academic achievement of children right through to adults (ages 4 to 85), where SAT is designed for school students. WIAT-II is used to assess the four general areas of: Reading, Math, Writing, and Oral Language. The reading comprehension subtests include: matching a written word with its representative picture, reading passages and answering content questions, and reading short sentences aloud, and responding to comprehensive questions.

Other standardised tests of reading comprehension include: Passage Comprehension from Woodcock-Johnson-III, Diagnostic Assessment of Reading Comprehension (DARC) (which are both analysed in (Francis et al., 2006)), the Gates-MacGinitie reading test, and the Gray Oral Reading test (which are both analysed in (Cutting & Scarborough, 2006)).

The materials for these standardised achievement tests are not openly available and must be purchased. Furthermore, they cannot be altered to fit a specific situation, such as those found in experiments. The main reason for listing some standardised tests is to show that there are many ways to assess reading comprehension and no particular way may be better than another. In fact, Cutting and Scarborough (2006) showed that individual tests vary in assessment of different measures of comprehension.

2.3.5.2 Eye Movements and Comprehension

Eye movements can be used to understand the on-going cognitive processes that occur during reading (Rayner, 1998). Models for reading based on the premise that lexical processing is driving eye movements have been built on these findings, such as the E-Z Reader (Reichle et al., 1998; Reichle et al., 2006; Reichle et al., 1999, 2003; Reichle et al., 2009) and SWIFT model (Engbert et al., 2002; Engbert et al., 2005). These models serve as default models for the reading process where lexical processing is assumed to drive eye movements. However, lexical processing is not the only factor that affects eye movements; comprehension of the text can have significant effects on the eye movements observed.

As a number of studies have shown, there are numerous variables that are largely based around comprehension functions that can have influence on eye movements during reading. The variables include: semantic relationships between words, anaphora and co-reference, lexical ambiguity, phonological ambiguity, discourse factors and stylistic conventions, and syntactic disambiguation (Rayner, 1998). These variables have different effects on eye movement that cause them to deviate from the default reading process. For example, garden-path sentences are syntactically ambiguous and induce regressions to resolve the comprehension problems (Frazier & Rayner, 1982). Amongst these findings it is believed that as text becomes increasingly hard to understand, fixation duration and number of regressions is observed to increase along with shorter saccades observed. This is due to the fact that higher order comprehension processes supersede the default reading process.

Indeed, experimental results show that eye movements reflect text difficulty (Rayner et al., 2006). As the difficulty in comprehending text increases so too does average fixation duration, the number of fixations and the total time taken to read the text (Rayner et al., 2006). Furthermore, this study showed that there is a higher probability of regressions when text was difficult. It is important to note that the text difficulty in this experiment was assessed independently by a group of students who had to rate the passages between 1 and 10. The authors note that although there was some correlation between poor comprehension and text difficulty, it was not statistically significant. The method of testing comprehension was not specified. Although there was no statistically significant correlation between text difficulty and comprehension, there was no mention of whether there were correlations between eye movement measures and comprehension. Nevertheless, the results from the Rayner et al. (2006) study confirm that eye movements are affected by overall text difficulty and that regressive eye movements can indicate comprehension failures.

Eye movements can reveal much about readers such as when readers encounter difficulties in reading (Frazier & Rayner, 1982), and text incomprehension (Okoso et al., 2015). Furthermore, fixation duration has been shown to be a predictor of reading comprehension (Underwood et al., 1990). Yet, the task of predicting quantified measures of reading comprehension has been attempted with poor results (Copeland et al., 2014b; Martínez-Gómez & Aizawa, 2014).

2.4 Eye Tracking

The discussion so far has been on eye movements, but we have not discussed how such movements are recorded. This section provides a brief overview of how some eye-tracking systems work followed by a discussion of some of the challenges faced dealing with this type of data. Put simply, an eye tracker is a device that captures eye position at regular time intervals to give estimates of where a person's gaze is (Morimoto & Mimica, 2005; Poole & Ball, 2005). Eye tracking is a relatively recent technology; the first being invented by Edmund Huey, for the purpose of reading analysis, and whose results were published in the late 1960's (Huey, 1968). This tracker, along with many early eye trackers, was intrusive and designed specifically for scientific research (Morimoto & Mimica, 2005). It involved the use of contact lenses that are embedded with a device such as a magnetic field sensor to estimate the person's eye gaze. These systems are very accurate but expensive and invasive (Morimoto & Mimica, 2005).

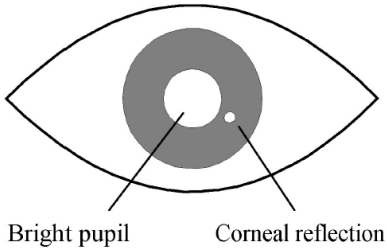


Figure 2.5. Pupil and corneal reflection are tracked with camera-based eye tracking to estimate eye gaze. Image take from (Poole & Ball, 2005).

At present, camera based eye trackers are most often used commercially. These types of eye trackers are usually non-intrusive to the user where a camera and light source are situated in front of the person whose gaze is being tracked (Poole & Ball, 2005). The person is usually not restricted in any way. However, sometimes to increase accuracy of tracking head mounted cameras are used (Morimoto & Mimica, 2005). Camera based eye trackers track at least one feature of the eye, such as the corneal reflection method. Many camera based eye trackers at present use light, usually infrared, to track the pupil and corneal reflection, as shown in Figure 2.5 (Goldberg & Wichansky, 2003; Morimoto & Mimica, 2005). An infrared light is targeted at the eye that generates a reflection off the surface of the eye and causes the pupil to appear as a bright disk (Poole & Ball, 2005). This is because the pupil reflects almost all of the infrared light. A camera can then be used to capture images of the eye and then the information is used to determine the eye rotation. However, most current camera based eye trackers are based on pupil-corneal reflection (Morimoto & Mimica, 2005). Corneal reflection is a glint on the cornea surface, also referred to as the first Purkinje image. The corneal reflection and the centre of the pupil are used to track the eye and determine where gaze is directed (Morimoto & Mimica, 2005). Specialised image processing software is needed to generate such results (Poole & Ball, 2005). Camera based tracking systems have to be calibrated to the participants' eyes for every session and potentially multiple times per session (Goldberg & Wichansky, 2003). This requires the participant to watch a dot appear

in several different locations of the screen. An example of this is shown in Figure 2.6.

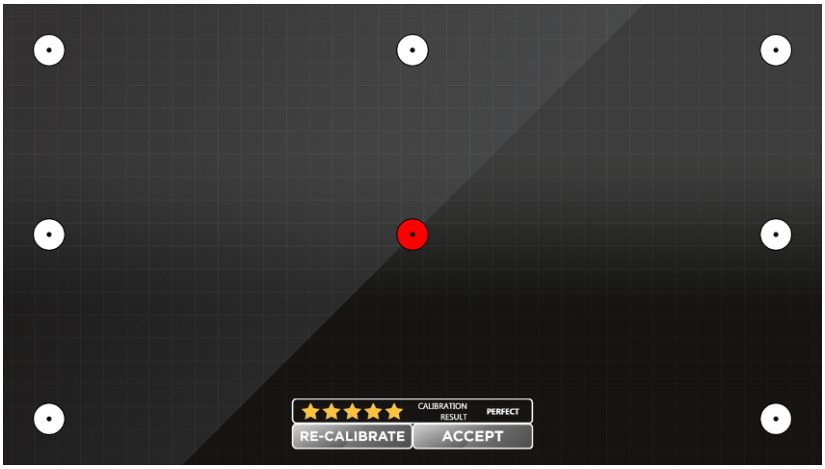


Figure 2.6. Example of calibration screen for "The Eye Tribe" eye tracker. Image taken from The Eye Tribe website: http://dev.theeyetribe.com/start/.

Head mounted eye tracking is useful for situations where the participant needs to move around the environment (Goldberg & Wichansky, 2003). However remote eye tracking is often used to study onscreen eye motion. This method has the disadvantage of requiring the participant to stay relatively still and often can be quite susceptible to equipment error or noise (Hornof & Halverson, 2002).

Recently remote camera based eye trackers have become inexpensive. Examples are the Tobii EyeX[9] and The Eye Tribe[10] which at the time of this writing were $USD 139 and $USD 99, respectively. The accuracy of the Eye Tribe is 0.5° – 1° which means it is able of determining the on-screen gaze position with only a 10mm error. This device is small, as shown in Figure 2.7, and attachable to tablet devices. These affordable devices expand the use of eye tracking from research or commercial use to the wider community.

Figure 2.7. The Eye Tribe eye tracker. Image taken from https://theeyetribe.com/order/ Last accessed: 27th January 2016

Remote eye trackers can also be used to measure pupil diameter (Klingner et al., 2008). In the fields of psychology and cognitive science many studies performed on pupillary response use specialist pupillometry systems. However, eye trackers are

[9] http://www.tobii.com/en/eye-experience/eyex/ Last accessed: 22nd August 2015
[10] http://theeyetribe.com/ Last accessed: 22nd August 2015

more readily available and inherently perform the task of tracking where the eye is positioned so it makes sense to use the eye trackers for recording pupil diameter as well. Klinger et al. (2008) demonstrated that classic pupillometry studies (Kahneman & Beatty, 1966) performed with specialist pupillometry equipment could be replicated with a video based remote eye tracker.

The output of eye trackers is time series data that is in the form of coordinates where the eye gaze position was captured at regular intervals (Goldberg & Wichansky, 2003). This is usually in x-y coordinates and a time stamp along with any other measurements that are requested by the experimenter such as pupil diameter. The following subsections will outline analysis of such data.

2.4.1 Dealing with Error

Modelling eye movement patterns is challenging. To add to this problem, analysis of fixation locations is often complex for several reasons: equipment noise, user variability and the size of the data set. The gaze data is often cleaned up and inference must be made about where fixations actually occur subsequent to data collection. First, we will consider the issues of eye tracking inaccuracies, which can be due to: inaccuracy of the equipment; the participants moving in front of the tracker causing drift from calibration; or simply that the participants eyes are hard to track (Hyrskykari, 2006). There are methods for adjusting and recalibrating the eye tracker during use such as the use of implicit required fixation locations (RFLs) (Hornof & Halverson, 2002). Implicit RFLs are locations on a screen that a participant must look at as part of a task and therefore provide a location from which the eye gaze data can be recalibrated if deviation has been encountered. Other algorithms such as those presented by Hyrskykari (2006) are highly related to reading tasks and involve using lines of text as the locations where fixations are reference points for mapping of the gaze data. This algorithm is used in real time as part of a reading aid called iDict and allows for manual corrections to be made if the fixations are not mapped to the right words (Hyrskykari, 2006). This algorithm focuses more on the vertical disposition of gaze points rather than the horizontal disposition. For post-collection recalibration of data, inference about where the fixations should occur can use the same logic as the above examples of recalibration of eye gaze trackers during experimentation.

2.4.2 Fixation Identification

Eye tracking can result in large data sets from monitoring even quite short tasks. Trackers typically sample many times per second, such as 50 to 60Hz. This means that even for a 10-minute task sampling at 60Hz there will be 36,000 data points generated. Fixation and saccade identification is the first essential step to take when analysing eye gaze data and can reduce the data considerably (Salvucci & Goldberg, 2000). However, fixation identification can have a large impact on results (Jacob & Karn, 2003). There is no standard fixation identification algorithm in current use, (Salvucci & Goldberg, 2000) so it is hard to compare results regarding fixations across experiments that do not use the same algorithms or even the same parameters (Jacob & Karn, 2003).

To complicate things further, during fixations the eye does not stay completely still. The eye can make very small rapid movements, or occasional drifts and sometimes microsaccades to bring the eye back to the original position (Salvucci & Goldberg, 2000). These mean very little to high level analysis of eye movements such as in reading analysis. Nevertheless they can make it harder to establish when fixations begin and end. Poor fixation identification may result in too few or too many fixations being extracted from the sequence of gaze points, which could in turn have dramatic effects on observations and further analysis.

Salvucci & Goldberg (2000) performed a comparison study on fixation identification algorithms. They divided them into two characteristic groups, spatial and temporal. Spatial algorithms are based on the velocity of saccades or based on dispersion of gaze points. Alternatively, temporal algorithms are time sensitive. In this thesis we use the dispersion-threshold based identification (I-DT) algorithm that is described in generic terms by Salvucci & Goldberg (2000). This algorithm is straightforward to understand and implement, as it relies on the underlying nature of fixations and saccades. That is, when the eye fixates on an object in the visual field it remains relatively still. This means that eye gaze points that are in close proximity, for a specific time frame, are likely to make up a fixation. Gaze points that are sparse are therefore more likely to be part of a saccade. There is average fixation duration of about 200–250ms and a range from 100ms to over 500ms, so on average about 12 to 15 gaze points make up a fixation with a range of about 5 to 30 fixations in each trial. Since there is a minimum of the range a fixation duration, the I-DT algorithm uses a minimum duration threshold to ensure a fixation meets the duration criterion. To check for fixation, a moving window approach is used. The initial window is set to encompass the minimum duration threshold and then the dispersion of points within the window is checked. If the distance between points is less than the dispersion threshold, the window is expanded to enclose more gaze points until the dispersion threshold is reached. At this point the fixation is closed off and a new window is created. For each fixation that is identified, a centre point and encompassing diameter must be calculated.

2.4.3 Interpretation of Eye Movements

There are two difficulties faced when interpreting eye movements that are due to human visual physiology. These difficulties are incidental fixations and off centre fixations (Salvucci, 1999). Although this has less impact on analysis of reading eye gaze patterns, it is important to keep in mind whilst modelling the data.

Incidental fixations are fixations that are accidental; these types of fixations are not of much interest when looking at reading eye gaze patterns (Salvucci, 1999). Gaze points recorded by eye trackers can be off centre over visual targets. This creates off centre fixations. Also humans can fixate within 1° visual angle of the target and still encode information in the fovea. To add to this, eye trackers have a typical accuracy of approximately 1°. This adds to the problem of mapping user actions to user intentions based on eye movement (Salvucci, 1999). This is a problem in terms of analysis of reading eye gaze patterns because calibration needs to be done to bring the fixation points in line with what the participant is actually fixating

on. If the points are not brought in line with actual fixation points, there could be misinterpretation of the gaze patterns.

Interpretation of eye gaze data can take many different approaches and is often based upon the purpose of the original research and the area that the research falls under. For instance, in usability studies in HCI eye tracking is often used to measure relative visual attention and how users look at a specific stimulus. In this field, there are approaches taken that are considered top-down and bottom-up interpretation of the data (Jacob & Karn, 2003). The top-down approach can be either based on cognitive theory or a hypothesis about the design. Analysis of eye gaze is therefore based upon either a cognitive theory such as longer fixations imply difficulty interpreting the interface or observations that change of a design causes longer fixations and therefore is harder to understand and use (Jacob & Karn, 2003). The bottom-up approach is used when there are no hypotheses before recording the data and instead patterns in the data are found and extrapolations can be made from there.

Eye movement patterns can be quite different depending on what task is being performed. The task a person is performing can be predicted based on their eye movement (Iqbal & Bailey, 2004; Simola et al., 2008). Even in the general task of reading there are differences in eye movement patterns (Fahey, 2009; Gustavsson, 2010; Vo et al., 2010). In most cases, there appears to be a difference in eye movement patterns when you visually compare gaze paths for reading the paragraphs to reading the questions. Copeland (2011) showed the types of movements (forward, backward and no movement) were statistically significantly different when comparing paragraphs to questions.

Figure 2.8 shows there is a difference between eye movements of participants recorded reading paragraphs compared to reading questions. This difference is expected, as when individuals answer questions they may study the questions and the answers more closely than they study the paragraph material the questions are based on.

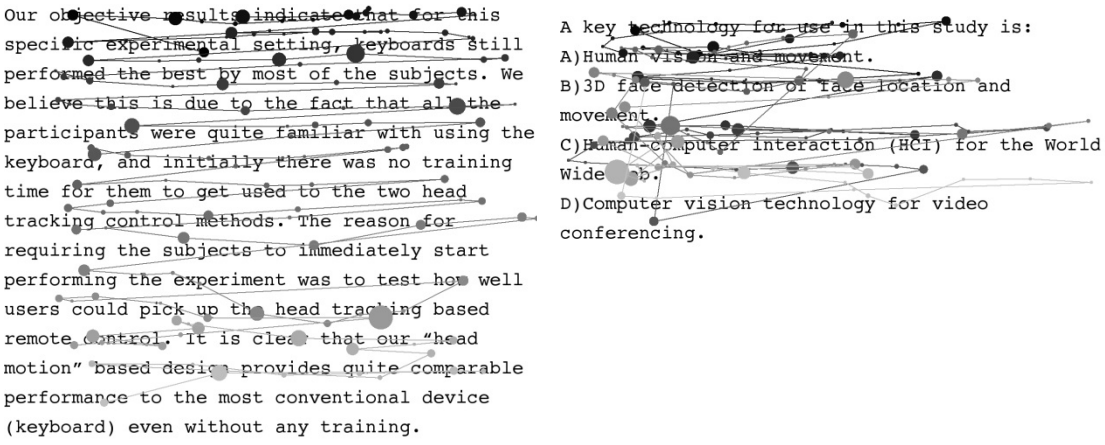


Figure 2.8. Eye movement trajectories of one participant; to the left is the eye movement whilst reading a paragraph and the right is the eye movement pattern whilst reading a question. Images taken from (Fahey, 2009).

Salvucci & Anderson (2001) put forward the idea of eye movement as protocols, which they describe as "tracing", which is plotting eye movements to predictions of a cognitive model. The three tracing methods are target, fixation and point tracing. These methods can be used differently in applications such as equation solving, reading and eye typing. They indicated that for reading, only fixation and point tracing are relevant. These three scenarios are an example of the different patterns that can be generated from eye movement. All three are essentially related to normal reading; when solving an equation you must first read the equation and when typing you must read the letters before you type. They generate quite different patterns where fixation trends tend to be focussed on the elements of the equation or the keyboard. There are many more applications of analysis of eye movement in tasks such as viewing faces, driving, watching television where complex patterns can be seen in eye movement. Salvucci & Anderson (1998) showed that "tracing" eye movement data is effective at interpreting the intent of eye movements using hidden Markov models. Tracing has been shown to generate accurate interpretations of these actions in areas such as eye typing and has been proposed to improve flexibility and design of eye based user interfaces (Salvucci, 1999).

2.5 The Use of Eye Tracking in HCI

Eye tracking has been used extensively in human computer interaction (HCI). Eye tracking can be utilised for quite different purposes and outcomes. This includes using eye movements to perform useability evaluation or using the eye movements as inputs to drive interaction with the system. The topics that will be discussed in this section include analysis of eye gaze data, usability analysis, and eye movement interactions with interfaces.

2.5.1 Eye Gaze Analysis in HCI

Eye gaze data can provide a wealth of knowledge about different tasks, not just about the cognitive functions that occur during reading. This subsection is highly related to the current research as it focuses on analysis of eye gaze data for the purpose of drawing conclusions about the nature of the eye movements and not the cognitive processes that make them occur. This will lead into the final subsection in this section that centres on providing feedback based on eye gaze. Different data analysis techniques have been used to achieve different outcomes. These will be discussed in the following subsections.

2.5.1.1 Task Identification

Tracking a person's eye gaze while they perform a task produces a pattern in which they view a visual scene. The patterns in which the eyes move can vary greatly between tasks, making eye tracking a useful tool for discriminating between tasks. Eye gaze patterns have been used to detect the following: what kind of task the participant is performing (Iqbal & Bailey, 2004; Salojarvi et al., 2005; Simola et al., 2008); when a person is viewing particular expressions on an individual (Kozek, 1997); and when a person is reading or not reading (Campbell & Maglio, 2001).

Previous studies have shown that even within the activity of reading, eye gaze patterns can be used to differentiate when individuals are reading different types of content (Vo et al., 2010) and that there is a correlation between the eye gaze patterns observed from reading (subjectively) hard or easy content (Rayner et al., 2006).

Distance and direction in letter spaces	Feature	Reading detector score s_r	Skimming detector score s_s
0 < x <= 11	Read forward	10	5
11 < x <= 21	Skim forward	5	10
21 < x <= 30	Long skim jump	-5	8
-6 <= x < 0	Short regression	-8	-8
-16 <= x < -6	Long regression	-5	-3
x < -16 and y according to line spacing	Reset jump	5 and line delimiter	5 and line delimiter
All other movements	Unrelated move	Line delimiter	

Figure 2.9. The scoring system for fixation transitions for the reading algorithm outlined in (Buscher et al., 2008). Taken from (Buscher et al., 2008).

Reading detection algorithms use eye gaze to detect reading, skimming and scanning behaviour (Buscher et al., 2008; Campbell & Maglio, 2001). The reading detection algorithm put forward by Campbell and Maglio (2001) uses averaged gaze points to differentiate between reading and scanning. The algorithm uses cumulative evidence for reading that is assigned for both the horizontal and vertical movements. The system starts in scanning mode. Points are associated to these movements and when enough evidence has been accumulated the system goes into reading mode. The system can be reset back into scanning mode by encountering a scan jump.

The reading detection algorithm put forward by Buscher et al. (2008) is an extension of algorithm just described. Buscher et al.'s (2008) reading detection algorithm uses sequences that are separated by reset jump features or unrelated moves, as defined in Figure 2.9, to separate behaviour. A reading and a skimming score are kept for each sequence, denoted s_r and s_s in Figure 2.9. The sum of each score is found for each sequence and if that sum is above a given threshold[11], $t_r = 30$ and $t_s = 20$, that is, $\sum_{f \in DF} s_d(f) > t_d$. If only one detector is above the threshold, then behaviour is defined as detected. If both reading and skimming behaviours are detected, the algorithm moves to the next line to discern between the two behaviours. The scores for the detection algorithm are shown in Figure 2.9. Further differences in the algorithms are that the parameters do not need to be set, as they are defined by the authors; gaze points are not used, instead fixation points are; the

[11] Authors specify that threshold is based on the literature

movement in the y axis is confined to movement between lines of text; and finally skimming behaviour is also detected as well as reading.

2.5.2 Usability Testing

Usability testing involves a wide range of methods and techniques. Methods include heuristic evaluation, cognitive walk through, pluralistic walk through and task analysis (Ramakrisnan et al., 2012). Techniques used include interviews, questionnaires, direct observation, video recorded observation and eye gaze tracking (Ramakrisnan et al., 2012). Often the methods and techniques are somewhat interrelated and more than one technique is often used. The technique that will be discussed in this section is eye gaze tracking.

Eye gaze tracking provides the ability to observe implicit behaviour during a task; that is, the user may not be aware they are exhibiting a certain behaviour or method to completing a task. Eye movements are the product of complicated cognitive and oculomotor processes, which have provided researchers with a bridge to the underlying workings of the human brain. This is what makes eye movement analysis such a diverse tool in HCI. Useability studies based on the implicit feedback of eye movements give researchers a unique method of assessing where people look, what catches their visual attention, and more importantly what they do not look at or miss in a visual stimulus An example of an application for analysing people's eye gaze as they view web pages is WebGazeAnalyzer (Beymer & Russell, 2005).

Many usability studies are designed to discover differences in expert and novice users, or investigate how users search for something in the interface. Results from both types of investigations can have great impact in the design and layout of an interface as well as give insight into training for use of an interface. An example of discovering design issues of an interface is analysis of the design of learning management systems (LMS) (Ramakrisnan et al., 2012). LMSs now play an important role in education as Web based delivery of content has become ubiquitous. The LMS interface must therefore be usable and beneficial to the learning process, as hindrance would detract from the learning process. In the study by Ramakrisnan et al. (2012), several design issues with the experimental LMS were discovered using eye tracking. From this information the authors outlined suggestions for improvements and potential guidelines for designing LMS interfaces.

In terms of marketing, this gives companies an important tool in assessing whether their marketing campaign is designed in such a way to attract people's attention to the right parts of the advertisement. As a tool in interface design, eye movement analysis can result in redesign of interfaces or displays to reduce error, increase efficiency or appeal. Somewhat related to this is the investigation of eye gaze whilst searching through Web search engine results. Whilst Web search engines return organic results they also display ads and pay-per-click placements of search results. Eye tracking allows researchers to see if there are factors that affect how those results are viewed. Interestingly, the visual attention paid to an ad is

dependent on not only its quality but also the quality of the ads in prior searches (Buscher et al., 2010). In the same study, Buscher et al. (2010) found that the visual attention paid to the organic search results, was dependent upon the task type and ad quality. Insights like these allow search engines to optimise their search results and advise how to produce effective ads.

2.5.3 Eye Movement Interaction

Eye gaze can also be used as an input device for human-computer communication, as opposed to a keyboard and mouse. This type of interaction has been investigated for the purpose of providing interaction with computers for disabled users (Jacob & Karn, 2003). However, as gaze tracking becomes more readily available and cheaper so too does the possibility of using eye movements to control hands free devices and computers in new and proactive ways.

The advantages of using eye gaze as an input medium for interaction with a computer are that eye movement is fast, natural, and the user will most likely have to look at a visual stimuli, such as a button, menu item, and so on, in order to click it anyway. The problem, however, with using eye movements as a control medium is in finding ways to respond to the eye movement inputs appropriately, that is, how do you differentiate eye movements for viewing the scene to eye movements for control purposes (Jacob & Karn, 2003). A common problem with using eye movements as a control device is how to design the interface so that it does not over-respond (Jacob & Karn, 2003). This is called the "Midas Touch" whereby everything the user looks at turns into a command. Of course the opposite of this is that the interface under-responds and the user has to look at something so long for the command to be issued that the interface becomes unnatural and too slow to use. The use of eye gaze to control on-screen keyboard input has been investigated (Lankford, 2000; Salvucci & Anderson, 2001), as well as to control a clicking device[12] with the eyes (Lankford, 2000; Murata, 2006).

Eye gaze in gaming has become more popular as eye-tracking technology becomes better and less expensive. The use of eye gaze in entertainment mediums such as in gaming not only could benefit people with disabilities but there is the possibility of using eye gaze as a sort of secondary input that rather than controlling the game, provides a more dynamic experience. A review of the use of eye tracking in gaming is given by Isokoski et al. (2009) and will not be discussed further in this thesis as it has little relevance to reading analysis and instead serves as an illustration the diverse uses of eye gaze.

2.5.3.1 Attention Aware Systems

Attention aware systems are where the eye gaze is integrated into the use of an interface in an implicit way, so that the user may not even be aware it. An example of this is in gaze-based rendering where high-resolution display is only rendered at the point of the users fixation (Jacob & Karn, 2003). This type of display exploits the fact that the fine detail vision occurs only in the fovea and so only high resolution of

[12] Analogous to a mouse controlled by the hand

image display is necessary at the point of fixation. Another example is the use of eye gaze in interactive environments. Gedeon et al. (2008) demonstrated that fuzzy signatures could be used to infer actions based upon eye gaze in an interactive task.

To an extent these systems are more of an amalgamation of usability and control based on eye gaze, whereby, they can provide feedback or assistance. The main goal of these systems is to act almost as a transparent interface whose input is eye gaze but where feedback is based upon use. An example of an implicit feedback system is an "attentive" document (Buscher et al., 2012). Eye gaze is recorded to give implicit information about the users' perceived relevance of pieces of text in a document. Two experiments are presented in this paper (Buscher et al., 2012); the first looks at providing implicit feedback to users about the ways in which they read documents in relation to how relevant or important they deem the documents. The second demonstrates the effect of implicit feedback for personalising web search. The aim is to work toward "attentive" documents that keep track of how they are read.

The Text 2.0 framework enables applications to use eye gaze to be used in real time to provide help with comprehension difficulties (Biedert et al., 2010; Biedert et al., 2010). The framework allows applications that analyse eye gaze to plug-in to gaze handlers. Several applications have been created to aid in reading comprehension for different purposes. An example is the use of eye gaze in creating footnotes that contain information about words to assist when reading in foreign languages. For further information see Biedert et al. (2010).

There are several applications that are used in reading assistance. iDict is a reading aid designed to help readers of a foreign language (Hyrskykari et al., 2000). iDict uses eye gaze to predict when a reader is having comprehension difficulties. If the user hesitates whilst reading a word then a translation of the word is provided along with a dictionary meaning. This is somewhat similar to The Reading Assistant (Sibert et al., 2000), which uses eye gaze to predict failure to recognise a word. The Reading Assistant then provides auditory pronunciation of the word to aid in reading.

2.6 Digital Text and eLearning

Digital environments are dynamic and immersive. The rise of the Internet, and ever growing expansion of the World Wide Web, has seen an increase in reading in many countries (Bohn & Short, 2009). This increase is growing with the proliferation of mobile technology such as smart phones and tablets. The Internet is now available almost anywhere at any time given you have a smart device. The debate on the effects of digitisation and rapid access to vast quantities of information ranges from ergonomics (Dillon, 1992, 2004), effect on memory (Sparrow et al., 2011), reading comprehension and effects on learning (DeStefano & LeFevre, 2007; Dillon & Gabbard, 1998; Mangen et al., 2013; Rockinson-Szapkiw et al., 2013). This section will begin with a discussion of these debates, and then lead into making learning environments adaptive.

2.6.1 Electronic text (eText)

Electronic text (eText) is the general term for digital presentation and storage of text. eText is read via digital devices, such as a computer, laptop, tablet, smart phone, or eReader. The advent of these devices has meant that eText is becoming more prevalent. The digitalization of text has spawned a great deal of research into what effects this has on the reading process. Initially, much research went into comparing reading digital to paper based texts (Dillon, 1992; Rho & Gedeon, 2000). We now give a brief overview of differences that have been found in the context of educational materials.

Hypertext is a prominent form of eText, in that it is the primary delivery of information on the web. Broadly, a hypertext document enables the reader to navigate via links to other resources or pieces of text. The resulting structure of hypertext documents can be complex and requires the reader to make decisions about where to go next. The consensus now is that hypertext structure negatively impacts the reading processes due to increased cognitive demand needed for decision-making and visual processing (DeStefano & LeFevre, 2007).

Hypertext is of course not the only form of eText. Quite often documents are read that are linear, such as PDFs (portable document format) or eBooks (electronic books). Such eTexts are therefore much closer to traditional print media. When print and PDF text comprehension was tested on students it was shown that students who read the print version of the text achieved significantly higher comprehension results than those who read a PDF version (Mangen et al., 2013). However, looking at the issue more abstractly, it has been shown that students who purchase electronic textbooks perform no differently in a university course (Rockinson-Szapkiw et al., 2013).

Paper offers advantages over digital presentation that has been studied to provide design suggestions for better reading technologies (O'Hara & Sellen, 1997). These include supporting annotation, quick and easy navigation, as well as control of spatial layout. Meanwhile, eText does itself have advantages over paper that include increased accessibility, easy storage and retrieval, ubiquity, and flexibility. Flexibility refers to the ability to dynamically change the way text is read. Changes can be simple, for example changing of font size, colour, or typeface. Changes can also be complex, such as verbalizations of the text, embedded definitions, and links to background information (Anderson-Inman & Horney, 2007). The ability to dynamically change eText presents the opportunity to make transformations to promote learning and comprehension. (Anderson-Inman, 1999) produced a typology of resources for supported eText that consists of presentational, navigational, translational, explanatory, illustrative, summarizing, enrichment, instructional, notational, collaborative, and evaluation resources. The typology is a list of ways in which eText can be supported; they vary vastly in method and purpose. Perhaps for this reason there is no consensus which supports should be provided (Anderson-Inman & Horney, 2007).

Many studies have examined navigation through eTexts, as it is often non-trivial (Dillon, 2004). Studies have investigated navigation in eBooks (McKay, 2011) and periodicals (Marshall & Bly, 2005) as well as the impact of screen size on document triage (Marshall & Bly, 2005). Navigation can be affected by the medium and familiarity with the book, whereby there is no difference in search efficiency between paper books and PC however the same task is performed significantly slower on a tablet PC (Shibata et al., 2015).

Additionally, the effects of highlighting, hyperlinks, fonts, distractions such as alerts, as well as embedded videos and sounds have long been investigated. The insight gained from these studies is beneficial in designing online reading materials. Inappropriate highlighting of words negatively affects reading comprehension whereas appropriate highlighting enhances comprehension (Beymer & Russell, 2005). The effects of font and font size used in eText have been investigated, where the focus has been on comparing serif and san-serif fonts (Bernard & Mills, 2000; Beymer et al., 2008; Mansfield et al., 1996). Smaller font sizes tend to induce slower reading speeds (Bernard & Mills, 2000; Beymer et al., 2008). This was found to result from increased fixation duration (Beymer et al., 2008).

The increased ease at which we can now locate information has changed the way in which we remember information (Sparrow et al., 2011). Knowing the information can be gathered from the Internet almost anywhere and anytime means that we often do not remember what we read on the Web and instead remember where to find information. People learn that the Internet "knows" certain information, which results in the tendency to not remember that information, instead remembering only the information that cannot found on the Internet (Sparrow et al., 2011).

2.6.2 eLearning

Associated with the rapid increase in digitised media is the rapid rise of eLearning. There are clear benefits to providing learning materials in digital format on the Web, namely making these materials accessible virtually anywhere at any time, to a wide and varied audience. In tertiary education, eLearning materials are becoming increasingly ubiquitous. This is due in part to increased accessibility and availability of computer technologies, but also because of the problems that large class sizes cause, such as limiting class discussions, assessment and time for teacher-student interaction (Longmore et al., 1996). Universities now frequently offer online and off-campus degrees where students may have little or no face-to-face interaction with their instructors or other students. Students are increasingly skipping lectures in favour of accessing digital copies or recordings of the lectures. The availability of lecture webcasts and PowerPoint slides negatively impacts student attendance (Traphagan et al., 2010). Whilst missing face-to-face tuition can have a negative effect on learning (Romer, 1993; Woodfield et al., 2006), webcasts of lectures actually nullify the negative effects on student performance and are instead associated with higher learning experience satisfaction (Traphagan et al., 2010). This means that eLearning has potential for making the learning process more enjoyable whilst increasing the amount learnt.

The advent of massive open online courses (MOOCs) has also increased the importance of designing effective eLearning materials. MOOCs have become popular in the past couple of years. The goal of MOOCs is to provide free or low cost but quality education that is available to anyone who wishes to take part. There are now many examples of reputable websites that offer MOOCs, such as Udacity, Coursera, edX, and Khan Academy. Whilst on one hand MOOCs do achieve the goal of making educational resources available to people who would not have access to them otherwise, they suffer from extremely low completion rates. An analysis of edX's first MOOC, Circuits and Electronics 6.002x, completion rate was below 5% (Breslow et al., 2013). One of the problems identified with MOOCs is that they are indeed massive, making them easy to get lost in and likely to end up unhappy, frustrated or overwhelmed. Students that are likely to succeed in completing MOOCs tend to be self-motivated, self-directed, and independent; they tend to be students who would succeed in a classroom setting and who are probably doing the MOOC out of interest rather than necessity (Howland & Moore, 2002).

The problem of how to make eLearning effective to a wide and varied audience is significant especially when learning materials come in many types and forms, quite often dependent on the subject being taught. For example, a mathematics course would have mathematical exercises as opposed to a history course, which would be more likely to have text-based materials. One solution is to use technology to provide personalised learning to students. The focus of this study is on text-based materials with assessment questions, and the use of eye gaze.

2.6.3 Providing Adaptivity in eLearning

Adaptive learning is the modification of educational material to suit a student's needs. Traditionally, a skilled instructor, who would observe a student's performance, would change the learning material to reflect the student's needs. However, the advent of computers allows for the automation of such a process and takes away the responsibility of a human instructor to make such judgments. Broadly these types of software packages are termed adaptive learning environments (ALEs) although they are also referred to as adaptive learning management systems (ALMS) and intelligent tutoring systems (ITS).

Adaptive eLearning has already started to show great promise in improving education. Adaptive tutorials have been harnessed to decrease failure rates in early year engineering subjects and drastically increase student enrolment and satisfaction (Prusty & Russell, 2011). In the area of learning from information visualisation, adaptivity to the user using innervations has been shown to improve performance (Carenini et al., 2014). For MOOCs adaptive support has been shown to improve user's acceptance as well as to isolate areas for improvement (Kardan & Conati, 2015). Finally, analysing attention of users in educational games increases performance when providing hints to help learning and completion (Conati et al., 2013).

Adaption can be performed by the student or by the system. Adaptability on the other hand is the term used when the student performs the adaption (Surjono,

2014). In these environments, the student is in control of how the adaption occurs by changing certain parameters. Adaptivity is the term used when the system performs the adaption (Surjono, 2014). In these environments student characteristics are detected and used to determine the adaption. These characteristics are determined using non-trivial means such as intelligent algorithms and machine learning. This thesis will focus on environments that provide adaptivity.

ALEs are typically composed of different components referred to as models (Paramythis & Loidl-Reisinger, 2003). These models include the expert model, which contains the learning material. The student model tracks and acquires information about the students' behaviour. The instructional model is where the learning material is delivered, and finally the instructional environment is the user interface for the ALE (Kareal & Klema, 2006). The student model is the main driver for how the system will be adapted. It is important that the model be as accurate as possible because the adaption performed can only be as good as the model.

There has been work to create adaptive learning environments in many different respects. Some examples of adaptive learning systems include InterBook (Eklund & Brusilovsky, 1999), which is a web-based adaptive tutoring system that allows textbooks to be navigated in multiple ways. The navigation of the textbook is personalized to assist the learner. Another example is Web-based Intelligent Design and Tutoring System (WINDS), which uses a student's past and current behaviour to predict their knowledge and goals, as well as record progress (Specht, Kravcik, Klemke, Pesin, & Hüttenhain, 2006). This information is used to provide adaptive learning material by annotating material and guiding students to suitable learning material. Generic Responsive Adaptive Personalized Learning Environment (GRAPPLE) project (De Bra et al., 2013) is another adaptive learning environment through adaptive guidance and personalized learning content. The authors of GRAPPLE show how they can integrate their system with currently used LMSs such as Claroline, Moodle, Sakai, Clix and learneXact. Other frameworks take into account students with learning problems such as dyslexia (Alsobhi et al., 2015). The Dyslexia Adaptive eLearning (DAEL) framework is designed to tailor learning materials according the dyslexia type (Alsobhi et al., 2015).

There are also companies actively involved in producing adaptive learning technology. There are several examples of commercially available adaptive learning technology; two such examples include DreamBox[13] and Smart Sparrow[14]. DreamBox is an adaptive learning platform that provides individualized learning paths based on the users measured skill level and use of gamification. Smart Sparrow is a web based adaptive learning environment that allows instructors to create, deploy and report on adaptive learning material. It is an intelligent tutoring system in which adaption comes from the answers that the student provides to questions.

[13] http://www.dreambox.com/ Last accessed: 7th January 2016
[14] https://www.smartsparrow.com/ Last accessed: 7th January 2016

2.6.4 Making eLearning adaptive using psychophysiological data

The basis of what drives changes in an eLearning environment can be based on different factors such as the learners current understanding, emotional state, such as stress (Calvi et al., 2008; Porta, 2008), emotions (Jaques et al., 2014), learner style (Mehigan et al., 2011; Spada et al., 2008; Surjono, 2014), cognitive load (Coyne et al., 2009), learning rate (Bondareva et al., 2013; Kardan & Conati, 2013), and skill level (Chen, 2008). The methods for determining these factors also vary and using the use of such information also varies. Methods for gathering user state and deducing these factors include the use of biometric technology (Mehigan et al., 2011; Spada et al., 2008) and psychophysiological response data (Rosch & Vogel-Walcutt, 2013), especially eye tracking (Alsobhi et al., 2015; Barrios et al., 2004; Bondareva et al., 2013; Calvi et al., 2008; Conati et al., 2013; Conati & Merten, 2007; Kardan & Conati, 2013; Merten & Conati, 2006; D'Mello et al., 2012; Gütl et al., 2005; Mehigan, 2014; Mehigan & Pitt, 2013; Mehigan, 2013; Mehigan et al., 2011; Porta, 2008). Development in technologies for measuring these signals and understanding of psychophysiological responses now provide the unique opportunity of adapting eLearning environments in real time.

Whilst learning style can be determined via questionnaire (Surjono, 2011) this interrupts the student with non-learning based assessment. Progressively more research indicates that measures of the students' behaviour and biometric technology can predict learner style, therefore alleviating the need to have student input (Mehigan et al., 2011; Spada et al., 2008). Mouse movement patterns have been shown to have a high correlation with global / sequential learning style (Spada et al., 2008). Eye tracking has also been shown to be a potential way of identifying visual/verbal learner style (Mehigan et al., 2011). Eye movements in areas of interest on the page were related to measures of learner style in that investigation. Similar uses of eye tracking have been used to compare learning behaviours between novice and advanced students when learning SQL (Liu, 2005). This study revealed that advanced students look at the database schema more than novice students. Studies such as this are useful for identifying this difference in order to provide more help for novice students. The concept of adaptive eLearning also extends to mobile learning. The MAPLE framework uses a combination of eye tracking and accelerometer data to determine learner style in both mobile and online environments (Mehigan & Pitt, 2013).

Adaption is not only provided via detection of learning style. Eye tracking can be used to detect many facets of human behaviour. Eye gaze patterns have been used to detect what kind of task the participant is performing (Iqbal & Bailey, 2004) or whether a person is reading or not (Campbell & Maglio, 2001) as well as if they are reading or skimming (Buscher et al., 2008), and their cognitive load (Rosch & Vogel-Walcutt, 2013). Eye movement measures have been shown to be effective at distinguishing between readers with low and high level of understanding as well as predicting English language skill (Martínez-Gómez & Aizawa, 2014). Eye gaze has also been used to investigate parts of text that readers are failing to comprehend.

Results from this investigation indicate that eye gaze features such as number and duration of fixations can be used to determine reading incomprehension (Okoso et al., 2015). Eye tracking can also be used to analyse how multiple-choice questions are answered (Nugrahaningsih et al., 2013; Tsai et al., 2012) and to predict student performance of physics concepts when presented as text or images (Chen et al., 2014).

Eye tracking has been used in multiple ways to provide adaptivity to eLearning. A classic example of the use of eye tracking in eLearning is AdeLE (Adaptive e-Learning with Eye-Tracking). The AdeLE project sets out a structure for how an adaptive eLearning environment could be constructed using eye tracking data such as blink rate and how open the eyelid is (Gütl et al., 2005).

Detection of a student's state are frequently investigated in adaptive eLearning, such as boredom and curiosity (Jaques et al., 2014), emotional state (Calvi et al., 2008), disengagement (D'Mello et al., 2012). An interesting approach to identifying students' engagement comes from the use of type-2 fuzzy logic based system (Paramythis & Loidl-Reisinger, 2003). This novel method gauges degree of engagement to adapt the learning environment. Results show that using the system to adapt material causes a significant improvement in average scores compared to other methods of adaption and no adaption.

Prediction of a student's learning rate is another important way in which a learning environment can react and adapt to the student. Prediction of learning rate has been shown to be effective using eye gaze data (Bondareva et al., 2013; Kardan & Conati, 2012). Similarly, eye gaze has been effective at predicting learning rate and initial experience with information visualisations (Lallé, Toker, Conati, & Carenini, 2015),. Additionally in the area of information visualisation, performance and user's cognitive abilities (Steichen, Conati, & Carenini, 2014) as well as confusion in processing the visualisation (Lallé, Conati, & Carenini, 2016) can be predicted with eye tracking data. Eye gaze is also effective as identifying parts of visualisations that are not conducive for associated tasks (Toker & Conati, 2014). Importantly though, these application are all in the form of prediction of user state to adapt the visualisation to the user, thus being of high relevance to this thesis.

Eye tracking is also used to analyse reading in eLearning environments. One example is iDict, a reading aid designed to help readers of a foreign language that uses eye gaze to predict when a reader is having comprehension difficulties (Hyrskykari et al., 2000). If the user hesitates whilst reading a word then a translation of the word is provided along with a dictionary meaning. Similarly, the Reading Assistant (Sibert et al., 2000) uses eye gaze to predict failure to recognize a word. The Reading Assistant then provides an auditory pronunciation of the word to aid reading. Eye movements have been used in combination with measuring pupil size as a means of gauging mental workload (Lach, 2013).

Adaption of reading material has been shown to be beneficial to young students (Dingli & Cachia, 2014). Adaptive eBooks involves detection of reading difficulty, currently based on measures such as out load reading speed, and dynamically

simplifying the text for the students. The system is designed for year 4 students and an initial study shows that such modifications can improve reading performance. However, the authors' note that the reading detection currently used is in the system is not sufficient and should be replaced, noting also that eye tracking would be a good solution.

In summary there is a broad range of scenarios that these adaptive technologies can be directed at helping students, such as plugging into traditional eLearning environments (Barrios et al., 2004; De Bra et al., 2013), or providing adaption in mobile environments (Mehigan & Pitt, 2013), or accounting for dyslexia (Alsobhi et al., 2015) and foreign language reading (Hyrskykari et al., 2000), and indeed eye tracking has shown to be an effective driver for these adaptions.

2.7 Summary

This literature review covered topics ranging from the physiology of the eye and visual information processing in the human brain to the use of eye tracking in adaptive eLearning. eLearning has extended the reach of teaching and learning from the classroom to a wide and varied audience that has different needs, backgrounds, and motivations. This gives rise to the question of how to make eLearning more effective through adaptivity. Whilst there are existing methods of providing adaptivity, eye tracking has been shown to be an effective way of analysing various human behaviours, particularly reading. Eye tracking is especially useful for analysing the implicit differences between different types of readers. This review sets the scene for the research presented in this thesis. We use this current knowledge to build from and produce advances in the integration of eye tracking technology into eLearning environments.

Chapter 3

Effect of Presentation on Reading Behaviour

"What this means is that we shouldn't abbreviate the truth but rather get a new method of presentation."

— Edward Tufte

The presentations of learning materials affect how we learn. In this chapter, we use eye tracking to investigate how different sequences of text and comprehension questions can affect performance outcomes, eye movements, and reading behaviour for first (L1) English language and second (L2) English language readers. We show that different presentation sequences induce different performance outcomes, eye movements, and reading behaviour. The sequence can affect how a participant reads the text as well as their perceptions of how well they understood what they read. For instance, if questions and text are not shown together, this improves participants' ability to accurately perceive their comprehension and promotes thorough reading. Alternatively, showing questions before the text promotes skimming behaviour. Importantly, the presentation sequence affects both L1 and L2 readers in the same way. We observe L2 reader take longer to read text but have the same comprehension levels as L1 readers, this difference comes primarily from longer fixation durations. The results from this study can be used to design learning materials in eLearning environments to influence how students interact with the learning environment as well as how they learn. The purpose of this investigation is to make informative decisions about designing adaptive eLearning environments. This chapter builds on work presented at OzCHI 2013 (Copeland & Gedeon, 2013a), OzCHI 2014 (Copeland & Gedeon, 2014a), and IHCI 2014 (Copeland & Gedeon, 2014b) and is largely based on work published in IEEE Transactions on Emerging Topics in Computing (Copeland & Gedeon, 2015).

3.1 Introduction

The way in which learning materials are presented to students can have great bearing on the outcomes of comprehension. It has been established that the presentation of images with text increases comprehension (Clark & Mayer, 2011). Moreover, pretesting with multiple-choice questions improves subsequent learning of materials (Little & Bjork, 2012). In this chapter we explore further how presentation of learning materials affects reading and learning behaviour. We investigate the effects that test questions have on learning by presenting questions and text in different sequences. Furthermore, we investigate if the effects of sequence are different for L1 and L2 readers since there is a growing diversity in the audiences of eLearning courses. Henceforth we will refer to the presentation sequences as formats. To make this comparison, the different formats are investigated to assess how eye movements and learning performance are affected. The central question being asked in this chapter is therefore:

Can outcomes of eye gaze analysis be used to optimise the layout of reading materials in eLearning environments for learning outcomes? How does the layout compare for L1 and L2 readers?

We explore this question by conducting a user study to compare four formats. These formats are manipulations to the order in which text and quiz questions are shown to a student. In the user study, participants' eye gaze was recorded, using eye tracking technology, as they read text and answered questions. Eye tracking has been shown to be an effective way of analysing various human behaviours, particularly reading (see review by (Rayner, 1998)). Eye movements are unique in reading and can reveal when readers encounter difficulties in reading (Frazier & Rayner, 1982) as well as text difficulty and comprehension (Rayner et al., 2006). Eye tracking is especially useful at analysing the implicit differences between different types of readers. One example is in comparing first (L1) and second (L2) English language readers, which reflects an increasing diversity in audiences of online learning materials. Kang (2014) found that L1 and L2 readers performed no differently in comprehension tests and that there was no difference in attention distributions when reading or in eye gaze patterns. L2 readers took longer to read the text and longer to find answers cues in the text. However, this study did not look into differences of eye gaze measures, so more is still to be understood about the differences in eye movement and learning behaviours of L1 and L2 readers. For instance, it may be that there are different methods of presentation of learning materials, which are optimal for L1 and L2 readers.

We hypothesize that the format of the text and comprehension questions will: 1) affect L1 and L2 readers in the same way even though there will be differences between the two groups; 2) have an effect on participants' performance, in terms of time and quiz score, and perceived understanding of the text; 3) cause differences in eye movements and induce different reading behaviour.

The background for this analysis is covered in the literature survey chapter, so this chapter is organized into the following sections: user study method; results and analysis; discussion and recommendations; and conclusions and further work.

3.2 Method

3.2.1 Design

Our study used a between-subjects design where participants were shown one of four formats of tutorial and quiz content. The independent factors of the experiments are the presentation type and English as a first language (L1) or second language (L2). Participants were permitted to take as long as they desired to complete the tutorial and quiz, with no time limit imposed.

3.2.2 Materials & Procedure

The user study conducted involved tracking participants' eye gaze as they read a text and answer comprehension questions. The text and questions are taken from a tutorial and quiz that is coursework from a first year Computer Science course run at the Australian National University. There are 9 screens of text, each covering a specific area about the main topic of the tutorial ("Web Search"). Each screen is 400 words long and has an average Flesch Kincaid Grade readability level of 11.5. This indicates that participants need around a 12th grade education level. This is a suitable readability level as the slides are targeted at first-year university students. For each screen there were two comprehension questions; one of the questions was multiple-choice and the other was cloze (fill-in-the-blanks). These two types of questions were used because they can be used to assess different forms of comprehension (Fletcher, 2006). The scores that the participants can receive for each question are 0, 0.5 and 1, corresponding to incorrect, half correct and correct respectively. Participants were not given any time restrictions on reading the text or answering the questions.

Upon completion of the quiz (but before being shown results) participants were asked to subjectively rate their overall comprehension on a scale of 1 to 10 with 10 being complete understanding. The text and questions are presented to participants in four formats to measure the effect of presentation on participants' eye gaze and answering behaviour. The formats are based on the presentation of quiz questions in relation to the text. These formats are described below:

Format A ($\boldsymbol{T \rightarrow T/Q}$). The tutorial text slide (T) Figure 3.1 is first shown to participants followed by a slide with both questions and the tutorial text (T/Q see Figure 3.2). Since there are 9 topics, 18 slides in total are displayed in this part of the study. In this format participants are required to read the text before being able to read the questions relating to it.

Format B ($\boldsymbol{T/Q}$). A slide containing both the questions and tutorial text (T/Q) is shown to participants. An example of this is seen in Figure 3.2. Since there are 9 topics, 9 slides in total are displayed in this part of the study. In this format

participants are no longer required to read the text before they see the questions. Our question is: is there a difference in quiz performance when participants can immediately answer the questions without reading the text?

Format C ($T \rightarrow Q$). The tutorial slide (T), shown in Figure 3.1, is first shown to participants followed by the questions slide (Q) but no access to the text, see Figure 3. Since there are 9 topics, 18 slides in total are displayed in this part of the study. This format can be considered to be a control presentation method. In this format the reference text is removed from the questions slide so the participants are forced to answer the questions from understanding and memory. We expect that the worst comprehension scores will be observed for this format. Format C is the most commonly used in on-line quizzes.

Format D ($Q \rightarrow T \rightarrow Q$). The last presentation format consists of displaying a slide with only the questions (Q) on it, as seen in Figure 3.3, followed by the tutorial text slide (T) Figure 1, and then again presenting them with the questions slide (Q) as in Figure 3. Since there are 9 topics, 27 slides in total are displayed in this part of the study. The reasoning for this format is to mimic a situation where the participants knew what the comprehension questions are but have no access to them as they read. The hypothesis is that participants will read the text differently than for formats A and C.

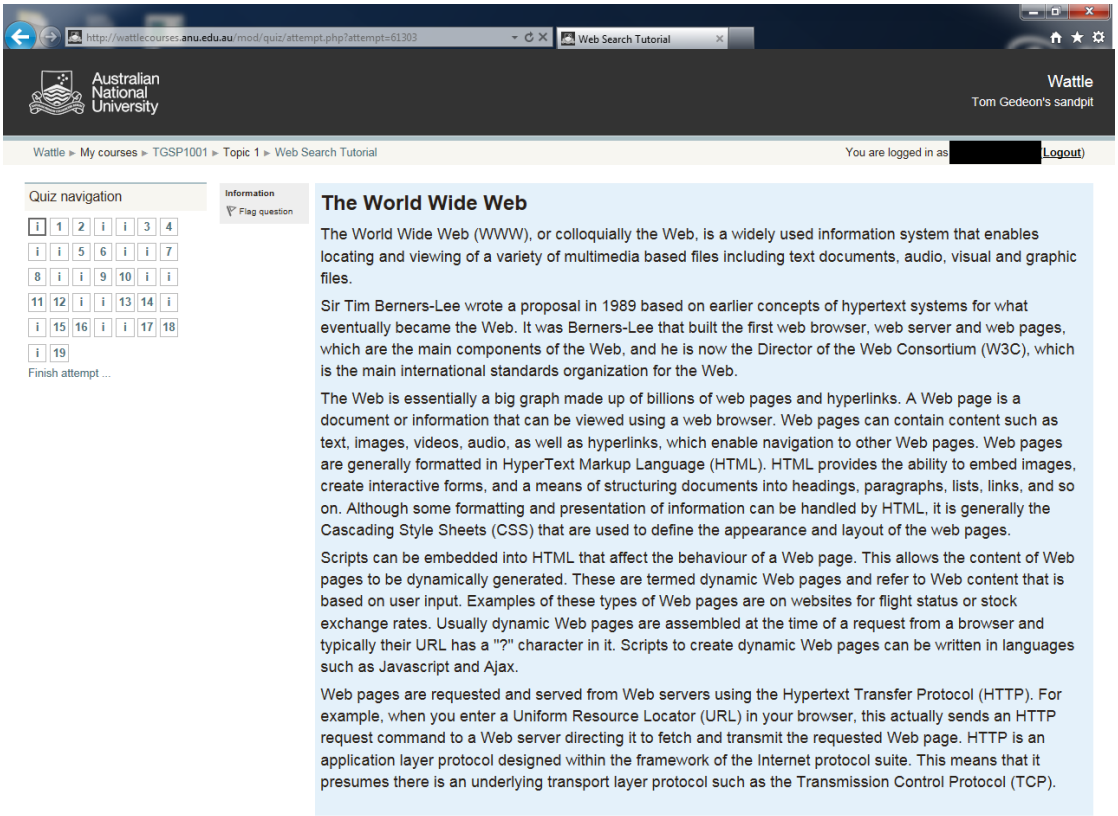


Figure 3.1. Example of text only tutorial page (T).

Figure 3.2. Example of text and comprehension question tutorial page (T/Q).

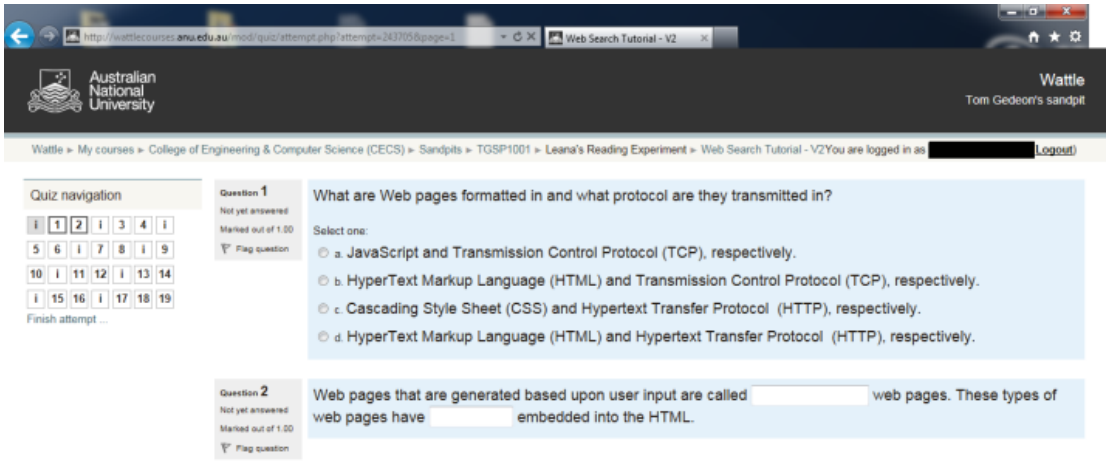


Figure 3.3. Example of comprehension questions only tutorial page (Q).

3.2.3 Participants

The study included 60 participants who were divided equally into the four groups, each of which was shown one of the presentation formats. The breakdown of participants into groups is as follows:

Format A. 15 participants (6 female, 9 male) with an average age of 22.3 years (standard deviation 4.1 years, range 17-31 years). English was not the first language for 4 of the participants.

Format B. 15 participants (6 female, 9 male) with an average age of 22.7 years (standard deviation 6.0 years, range 18-41 years). English was not the first language for 4 of the participants.

Format C. 15 participants (5 female, 10 male) with an average age of 23.5 years (standard deviation 5.3 years, range 18-37 years). English was not the first language for 6 of the participants.

Format D. 15 participants (7 female, 8 male) with an average age of 22.2 years (standard deviation 3.3 years, range 17-28 years). English was not the first language for 5 of the participants.

Figure 3.4. Experiment set up; Participant to the left with the experimenter's laptop and view to the right.

3.2.4 Experiment Setup

The tutorial quiz was accessible via Wattle (a Moodle variant) the online learning environment used at ANU. A copy of the texts used for the experiment, along with the participant information sheet, consent form, and other experiment resources are found in Appendix A. The study was displayed on a 1280x1024 pixel Dell monitor. Eye gaze data was recorded at 60Hz using Seeing Machines FaceLAB 5 infrared

cameras mounted at the base of the monitor. This is shown in the left half of Figure 3.4 along with the laptop the experimenter used to monitor the eye tracking quality in the right half.

This eye tracker has a gaze direction accuracy of 0.5-1° rotational error and measures pupil diameter as well as blink events. The study involved a 9-point calibration prior to data collection for each participant. As the data recorded is a series of gaze points, EyeWorks Analyze was used to pre-process the data to give fixation points. The parameters used for this were: a minimum duration of 60 milliseconds and a threshold of 5 pixels.

3.2.5 Data Pre-processing

The raw eye gaze data consists of x, y-coordinates recorded at equal time samples (60Hz). Fixation and saccade identification was performed on the eye gaze data. From this data many other eye movement measures are derived. The measures used in this analysis are:

Number of fixations: The sum of fixations recorded for each page. The number of fixations can be affected by the reading behaviour, text difficulty, and reading skill (Rayner, 1998).

Maximum fixation duration (seconds): The maximum duration of the longest fixation recorded for a tutorial page. Longer fixations can be an indicator of difficulties in processing particular words or due to linguistic and/or comprehension difficulties (Rayner, 1998).

Average fixation duration (seconds): The sum of the duration of all fixations on a paragraph divided by the number of fixations on that paragraph. This measure has been used to predict reading comprehension (McConkie & Rayner, 1975).

Total fixation duration (seconds): The sum of all fixations on complete text. This measure is useful in global text processing analysis (Hyona et al., 2003) because it measures immediate as well as delayed effects of comprehension.

Number of regressions and regression ratio: The number of regressions divided by the total number of saccades on a paragraph. There is evidence that when reading more difficult text more regressions are observed (Rayner et al., 2006).

Reading analysis: Using our combination of two reading detection algorithms (Buscher et al., 2008; Campbell & Maglio, 2001), this is the percentage of saccades classified as being part of reading (read ratio), skimming (skim ratio), and scanning/searching (scan ratio).

Participants' quiz outcomes are measured to assess how well they performed under different conditions. The measures of participants' performance are:

Subjective comprehension: a self-rated measure between 0 and 10, where 10 is comprehensive understanding of the material.

Comprehension question scores: the multiple-choice questions are graded as 0 (incorrect) or 1 (correct) and the cloze questions are scores as 0 (incorrect), 0.5 (one word was correct) or 1 (correct). The maximum total score for the quiz is 18.

Time taken: the total time it took each participant to complete the tutorial and quiz is recorded.

3.3 Result & Analysis

The first part of this section contains a statistical analysis of participants' performance (score, time taken and perceived comprehension) under each of the experimental conditions. Additionally, L1 and L2 readers are compared under each condition. The second part of this section contains the statistical analysis of the eye movement measures derived from the participants' eye gaze under each of the experimental conditions. Once again, the L1 and L2 readers are compared.

3.3.1 Does format affect performance?

The question of whether format affects reader performance incorporates two hypotheses that will be explored in this subsection. These hypotheses are:

1. The different presentation formats will affect participants' scores, time taken to complete, and perceived understanding, and these effects will be the same for both L1 and L2 readers.
2. Only time taken to complete will be different between the L1 and L2 readers.

The mean and standard deviations for the quiz grade is shown in Figure 3.; the time taken (minutes) to complete the tutorial and quiz is shown in Figure 3.; and the participants' subjective understanding is shown in Figure 3..

To address the above hypotheses a MANOVA is used to determine if there are any statistical differences between the formats and L1/L2 readers. The correlations between the dependent variables are within the acceptable limits for MANOVA outcomes, i.e. the correlations lie between r=-0.4 and r=0.9. To test for normality in the dependent variables the Shapiro-Wilk Test is used, as it is more appropriate for small sample sizes. The quiz scores are normally distributed for all formats (all $p>0.05$). The times taken are normally distributed for the formats A, B and C (all $p>0.05$), it is just the times taken for D ($p=0.026$) which is still relatively normal and should not impact the MANOVA as the assumption is for approximately normal distributions. Whilst the subjective scores for B are normally distributed and the scores for C are very close to being normally distributed, the scores A and D could be a problem. Finally, the homogeneity of variance-variance-covariance matrices is satisfied as the Box's M value of 69.73 ($p=0.165$).

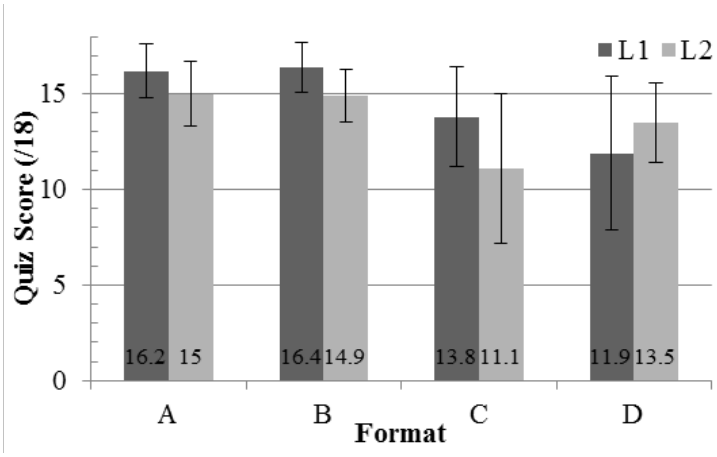


Figure 3.5. Means and standard deviations of quiz scores for each format (A: $T \rightarrow T/Q$; B: T/Q; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$)

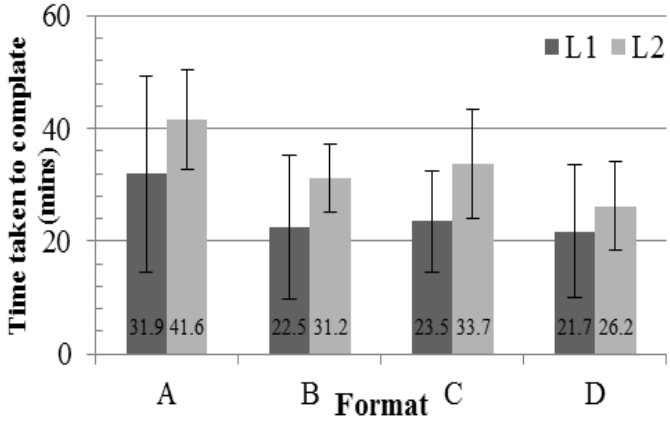


Figure 3.6. Means and standard deviations of time taken to complete the tutorial for each format (A: $T \rightarrow T/Q$; B: T/Q; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$)

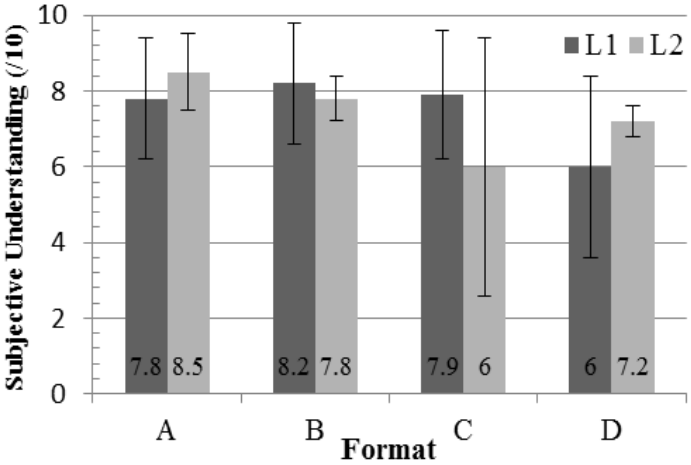


Figure 3.7. Means and standard deviations of subjective understanding scores for each format (A: $T \rightarrow T/Q$; B: T/Q; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$)

3.3.1.1 Effect of format on Performance Measures

The results in Figure 3.5, Figure 3.6, and Figure 3.show that whilst there are some differences between L1 and L2 readers, the effect of format is relatively consistent for both L1 and L2 readers. This indicates that the format does affect performance outcomes for both groups. This is supported by the results from the MANOVA that show there is a statistically significant difference in performance variables based on the format shown to participants, $F(9,121.8)=4.036$, $p<0.0005$; Wilk's $\lambda=0.530$, partial $\eta^2=0.191$. There is no statically significant effect of interaction between the format and reader type. This indicates that format affects both L1 and L2 readers in the same way.

Since statistically significant results have been found, we use ANOVAs to assess if the formats have an effect on the dependent variables. Format has a statistically significant effect on both the quiz grade ($F(3,52)=6.078$; $p=0.001$, partial $\eta^2=0.260$) and on time taken ($F(3,52)=5.552$; $p=0.002$, partial $\eta^2=0.243$), however format did not affect the subjective comprehension score. Tukey's HSD tests are used to make pairwise comparison of the formats. Figure 3.5 shows that Formats A and B have similar quiz scores, as do formats C and D. There is no significant difference in quiz scores between formats A and B or between formats C and D. These two groups correspond to similarities in presentation formats whereby formats A and B show the questions with the tutorial text and Formats C and D do not. Two conclusions can be made from this observation; firstly, the lack of difference between formats A and B illustrates that reading the tutorial text before being presented with the questions does not improve comprehension scores. Secondly, when comparing formats C and D, the knowledge of the questions before reading the text also does not improve quiz results.

However, formats A and B have significantly higher quiz scores than formats C and D, (formats A and C ($p=0.006$), A and D ($p=0.003$), B and C ($p=0.005$), and B and D ($p=0.002$)). For formats C and D the participants did not have access to the content as they answered the questions and therefore had to rely on memory and their understanding of the material.

Format A takes significantly longer to complete than formats B ($p=0.011$) and D ($p=0.002$). There is no significant difference between the other formats. For format A, participants were asked to read the text and then move to the next page with the questions, where they also had the option to re-read the content. The lack of significant difference between formats A and C could be accounted for by the participants reading the text on the text only page before the questions and text page, which is analogous to format C.

The format has no significant effect on subjective comprehension scores. However, for formats C and D there are strong positive correlations between the quiz scores and the subjective comprehension scores ($r=0.9$ and $r=0.8$, respectively). In these formats participants estimate their comprehension level more accurately compared to other formats. Participants shown formats A and B seem unable to estimate their own comprehension levels ($r=0.3$ and $r=-0.1$, respectively). An important part of the learning process is awareness of skill (Dunlosky & Lipko,

2007). Under-estimation of understanding can lead to students wasting time on material already understood instead of using the time to learn more material. On the other hand, overestimation of understanding will results in students not learning what they need to and not realizing their lack of understanding.

For format C, participants are asked the comprehension questions after having read the content and cannot refer back to the text. The participants can seemingly gauge whether they know the answers or not. Interestingly, this effect extends to format D where once again the participants did not have access to the content whilst they answered the questions. However, the difference in this format is that participants knew the comprehension questions before reading the content and so could target their reading goals for answering those questions. For formats A and B the participants have access to the text whilst answering the questions. Participants accordingly do not fully read the content and thus fail to find key concepts in the text. In this case the participants have a false sense of confidence.

3.3.1.2 L1 versus L2 readers

The second hypothesis is that the only difference expected between L1 and L2 readers will be in time taken. The MANOVA shows that there is a statistically significant difference in performance variables between L1 and L2 readers, $(F(3,50)= 5.79, p=0.002$, Wilk's $\lambda=0.742$, partial $\eta^2=0.258)$.

Between-subjects ANOVAs are used to compare the groups for each performance variable. The difference between L1 and L2 readers is statistically significant for time taken ($F(1,52)=13.135$; $p=0.001$, partial $\eta^2=0.202$) but has no significant effect on subjective comprehension or quiz score. This confirms our expectations and is analogous to existing research that has shown that although L2 readers take longer to read, they perform no differently to L1 readers in comprehension (Kang, 2014). We have also found there is no difference in their subjective comprehension.

3.3.1.3 Summary

The interim conclusion made from this analysis is that presentation formats affect students' performance. In concordance with current research it was found that L2 readers took longer to complete the quiz but performed no differently to L1 readers. Additionally, the differences in measures caused by formats are consistent for both L1 and L2 readers. The presentation format can be manipulated in the same way for both L1 and L2 readers to optimize the performance outcomes of students in order to increase their understanding.

3.3.2 Does format affect eye movements?

The overall hypotheses are that presentation format affects eye movements and that the eye movements of L1 and L2 readers will be different. To address these overall hypotheses, the two central differences in presentation formats are analysed separately. That is, first the tutorial text when shown without the questions will be analysed and then the tutorial text when shown with the questions. Finally, aspects

of the answering process that derive from the nature of the presentation format will be analysed, namely, reading intensity of paragraphs.

3.3.2.1 Text Pages

Two types of behaviour are hypothesized for reading the tutorial text without the questions:

1. Participants presented with format C will take more care reading the text, as they know they cannot refer to it again whilst answering the comprehension questions;
2. Participants presented with format D will not read the text thoroughly, rather will skim the text to find the paragraphs where they believe the answers are located and read only those paragraphs thoroughly.
3. L2 readers will be observed to read the text for longer, i.e. more fixations and longer fixation duration.

The final hypothesis is a deeper analysis into the observation that L2 readers have longer read times than L1 readers.

Table 3.1. Comparison of eye movement measures for text only (T) pages (Mean ± Standard Deviation) (A: $T \rightarrow T/Q$; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$).

Format	Type	Num. Fixations	Max fixation dur (s)	Ave fixation dur (s)	Num. regressions
A	L1	241 ± 21	1.1±0.2	0.17±0.02	74±7
	L2	311±35	2.1±0.3	0.25±0.03	83±11
C	L1	245±23	1.3±0.2	0.21±0.02	75±7
	L2	351±28	1.6±0.2	0.23±0.02	106±9
D	L1	178±22	1.0±0.2	0.17±0.02	66±7
	L2	221±31	1.9±0.3	0.26±0.03	66±10

A MANOVA was used to check for statistical significance of eye movement measures between formats and reader type. The correlations between the dependent variables are all within the range of r=-0.4 and r=0.9. Additionally, the majority of the dependent variables are normally distributed according to the Shapiro-Wilk test for normality for both reader type and format. The total fixation duration time was excluded from the analysis, as it did not have a normal distribution. The Levene's test for equality of variances shows that there is homogeneity for all dependent variables (p>0.05). Additionally, the Box's M value of 98.1 (p=0.025) is interpreted as non-significant so we can be satisfied that we have homogeneity if variance-variance-covariance matrices. The means and standard deviations for the eye movement measures are shown in Table 3.1, whilst the means and standard deviations of reading ratios for each format are shown graphically in Figure 3.5.

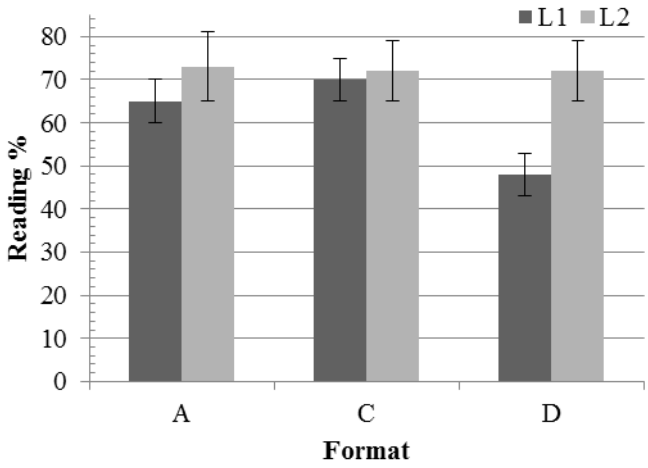


Figure 3.5. Means and standard deviations of reading ratios (% of eye movements detected as reading) for text only page which are in formats A, C and D (A: $T \rightarrow T/Q$; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$).

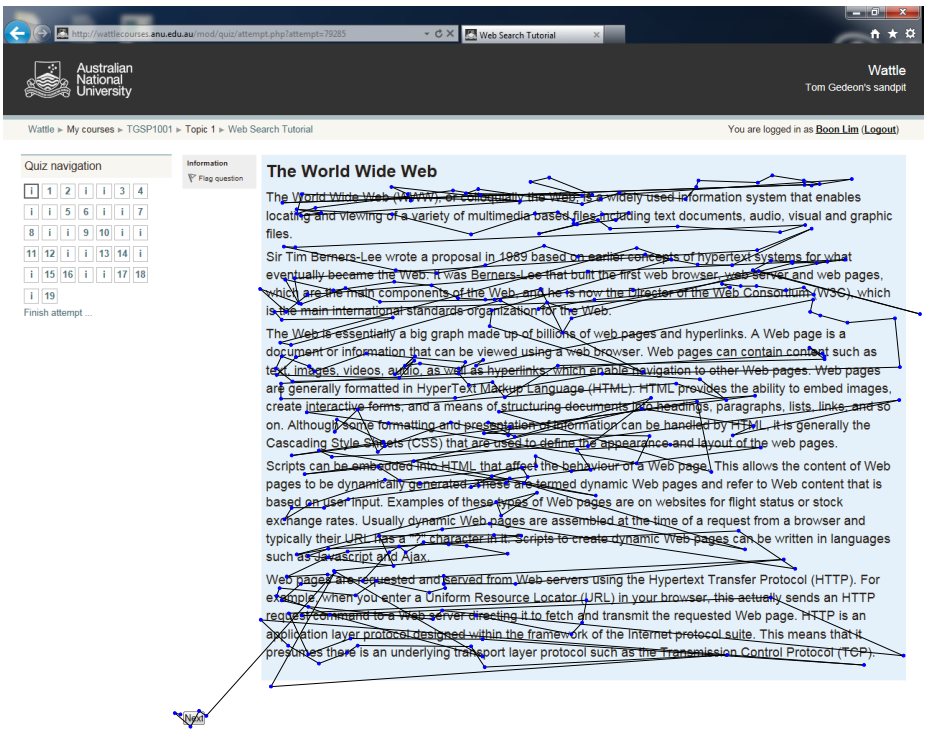


Figure 3.6. Example of fixations recorded from reading text only page in format A ($T \rightarrow T/Q$)

There is a statistically significant difference in eye movement measures based on the presentation format the participant was exposed to ($F(10,72)=3.043$, $p=0.003$, Wilk's $\lambda=0.486$, partial $\eta^2=0.303$). Furthermore, there is a statistically significant difference in eye movement measures between L1 and L2 readers ($F(5,35)=3.623$, $p=0.010$; Wilk's $\lambda=0.659$, partial $\eta^2=0.341$). There was no statistically significant effect of interaction between the format and reader type. Once again, format affects both L1 and L2 readers in the same way.

ANOVAs are used to determine how the eye movements differ for the formats and languages. Format has a statistically significant effect on the number of fixations ($F(2,39)=7.262$; $p=0.002$; partial $\eta^2=0.271$), and number of regressions ($F(2,39)=4.234$; $p=0.022$; partial $\eta^2=0.178$), but no effect on maximum fixation duration, average fixation duration or the read ratio.

Tukey's HSD tests are used to make pair-wise comparisons of the formats. There is a statistically significant difference between number of fixations for Formats A and D ($p=0.034$) and between formats C and D ($p=0.002$). There is a statistically significant difference between the number of regressions for formats C and D ($p=0.030$), but not between formats A and D or A and C. There is no significant difference between formats A and C for any of the eye movement measures.

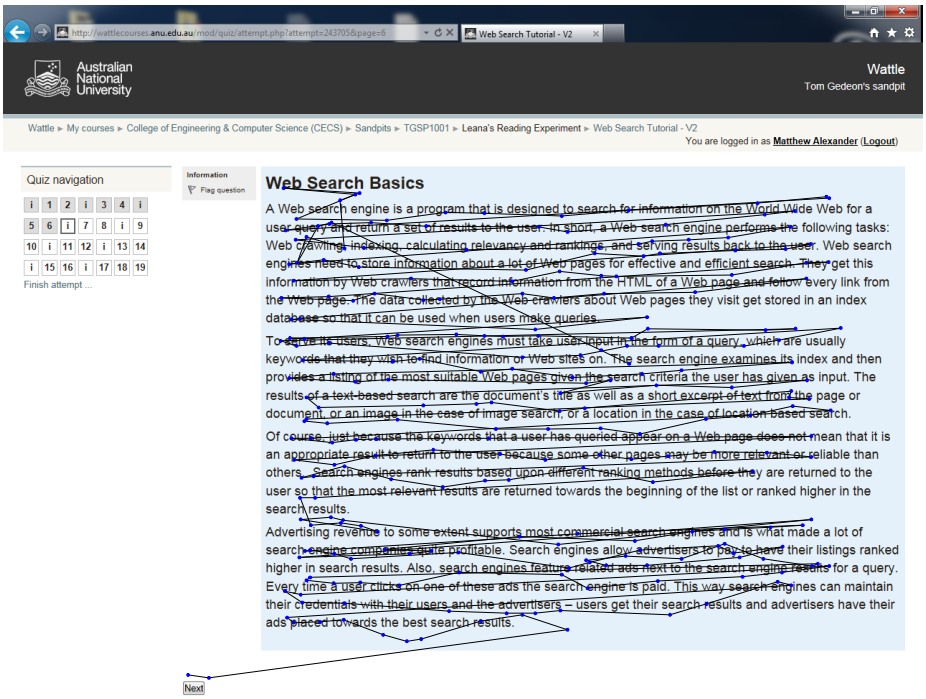


Figure 3.7. Example of fixations from reading text only page in format C ($T \rightarrow Q$)

It was predicted that, for format C, participants would read the text more thoroughly. However, the statistical analysis shows that there is no difference between formats A and C, so participants are actually reading format A as thoroughly as they are reading C. The hypothesis is partially supported as there are significantly fewer fixations recorded for format D, so even though there is no difference in the read ratio there is less overall reading of the text compared to formats A and C. This can be observed visually in the comparison of three different participants' fixations as they read the same text under different formats in Figure 3.6, Figure 3.7, and Figure 3.8. Whilst the eye movements in Figure 3.6 (format A) and Figure 3.7 (format C) are different, they do show coverage of the entire text. These differences can also be put down to individual variance in reading. However, the eye movements shown in Figure 3.8 (format D) are substantially different from those in Figure 3.6 and Figure 3.7, where only the first paragraph is read. All images are indicative of the reading behaviour observed for each of the formats.

Finally, L2 readers have significantly more fixations ($F(1,39)=11.395$; $p=0.002$; partial $\eta^2=0.226$) than L1 readers as well as longer maximum fixation duration ($F(1,39)=13.840$; $p<0.001$; partial $\eta^2=0.262$) and longer average fixation duration ($F(1,39)=11.527$; $p=0.002$; partial $\eta^2=0.228$). Also, L2 readers also have significantly higher read ratios for each format compared to L1 readers ($F(1,39)=4.951$; $p=0.032$; partial $\eta^2=0.113$). This outcome agrees with the observation that L2 readers have longer read times than L1 readers. The analysis of eye gaze shows that this is due to higher numbers of fixations that are also for longer duration.

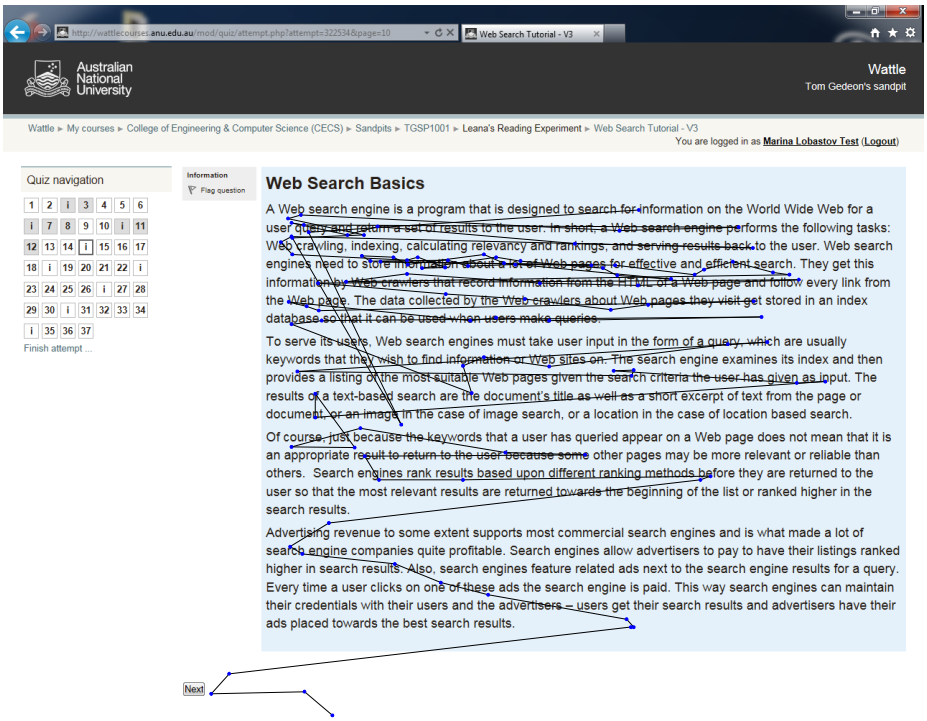


Figure 3.8. Example of fixations recorded from reading text only page for format D ($\boldsymbol{Q \rightarrow T \rightarrow Q}$)

In conclusion, the eye movements and reading behaviours that are observed for the formats A, C and D reflect the participants' overall intentions in reading the text and the goals set for the participants. The purpose of this analysis was to assess the hypotheses that the different presentation formats would affect the eye movements observed and therefore the reading behaviours observed. These hypotheses have been confirmed by this study. The implications of these findings can be used to support design decisions for eLearning environments. That is, if the teacher wants to promote thorough reading, the goals placed on the reader should not be targeted at certain parts of the text as in format D. Instead, thorough reading is observed where the goal was to understand the text overall.

3.3.2.2 Questions and Text Pages

Format A consists of two presentations of the text, first on its own and second with the questions. The hypothesis is that the first read through of the text in format A will help participants answer the questions and they will need less reference to the text compared to Format B. The means and standard deviations for the eye

movement measures are shown in Table 3.2, whilst the means and standard deviations of reading ratios for each format are shown graphically in Figure 3.9.

A MANOVA is used to test for statistical significance of eye movement measures between formats and reader type. The correlations between the dependent variables are all within the range of r=-0.4 and r=0.9. All of the dependent variables are normally distributed according to the Shapiro-Wilk test except average fixation duration, which is therefore excluded from the analysis. Levene's test for equality of variances shows that the homogeneity for all dependent variables (p>0.05). Box's M value of 45.8 (p=0.005) is interpreted as non-significant so we can be satisfied that there is homogeneity in the variance-variance-covariance matrices.

Table 3.2. Comparison of eye movement measures for Questions and Text pages (Mean ± Standard Deviation) (A: $T \rightarrow T/Q$; B: T/Q)

Format	Type	Num. Fixations	Max fixation dur (s)	Total fixation dur(s)	Num. regressions
A	L1	225±37	0.97±0.13	38±8	97±14
	L2	246±61	1.65±0.21	54±14	95±23
B	L1	350±37	1.31±0.13	64±8	149±14
	L2	429±61	1.85±0.21	102±13	167±23

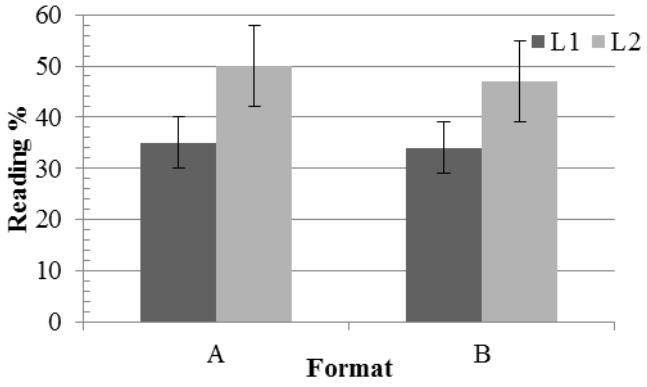


Figure 3.9. Means and standard deviations of reading ratios (% of eye movements detected as reading for text and questions pages) for Formats A and B. (A: $T \rightarrow T/Q$; B: T/Q)

There is a statistically significant difference in eye movement measures based on presentation format, (F(5,22)=3.142, p=0.027; Wilk's λ=0.583, partial η^2=0.417. Also there is a statistical difference between L1 and L2 readers, F(5,22)=3.309, p=0.022; Wilk's λ=0.571, partial η^2=0.429. However, there is no statistically significant effect of interaction between reader type and format. There is no difference in how L1 and L2 readers are affected by the presentation format.

ANOVAs are used to compare each of the eye movement measures separately. Format has a statistically significant effect on number of fixations (F(1,26)=9.279, p=0.005), total fixation time (F(1,26)=10.924, p=0.003), and the number of regressions (F(1,26)=10.827, p=0.003) but not on maximum fixation duration or the read ratio. Thus, format B has more observed fixations and therefore a longer total fixation

time as well as more regressions. This confirms the hypothesis that less eye movements would be observed for Format A. This can also be seen visually in Figure 3.10 and Figure 3.11. Both figures show the eye movements for reading and answering questions on the questions and text tutorial pages. For format A it can be seen that whilst participants did use the text to answer the questions (Figure 3.10) there are fewer fixations and therefore less reading of the text compared with Format B (Figure 3.11).

Figure 3.10. Example of eye movements from reading and answering questions on questions and text tutorial page for Format A ($T \rightarrow T/Q$)

L2 readers have significantly longer maximum fixation durations ($F(1,26)=12.230$, $p=0.002$) and higher read ratios ($F(1,26)=4.350$, $p=0.040$) compared to L1 readers. Additionally, L2 readers have significantly longer total fixation durations than L1 readers ($F(1,26)=5.870$, $p=0.023$). However now there is no difference between the numbers of fixations observed for L1 and L2 readers and there is no significant difference between the numbers of regressions observed for L1 and L2 readers. This is an interesting result as no difference in the number of fixations between L1 and L2 readers indicates that the increase in time taken for L2 readers is due primarily to increased fixation duration.

The conclusion from this analysis is that pre-reading of the text before questions is asked (Format A) decreases the time needed to answer the questions. This means

that the participants are using their knowledge of the text to answer the questions as well as checking the text for the correct answers.

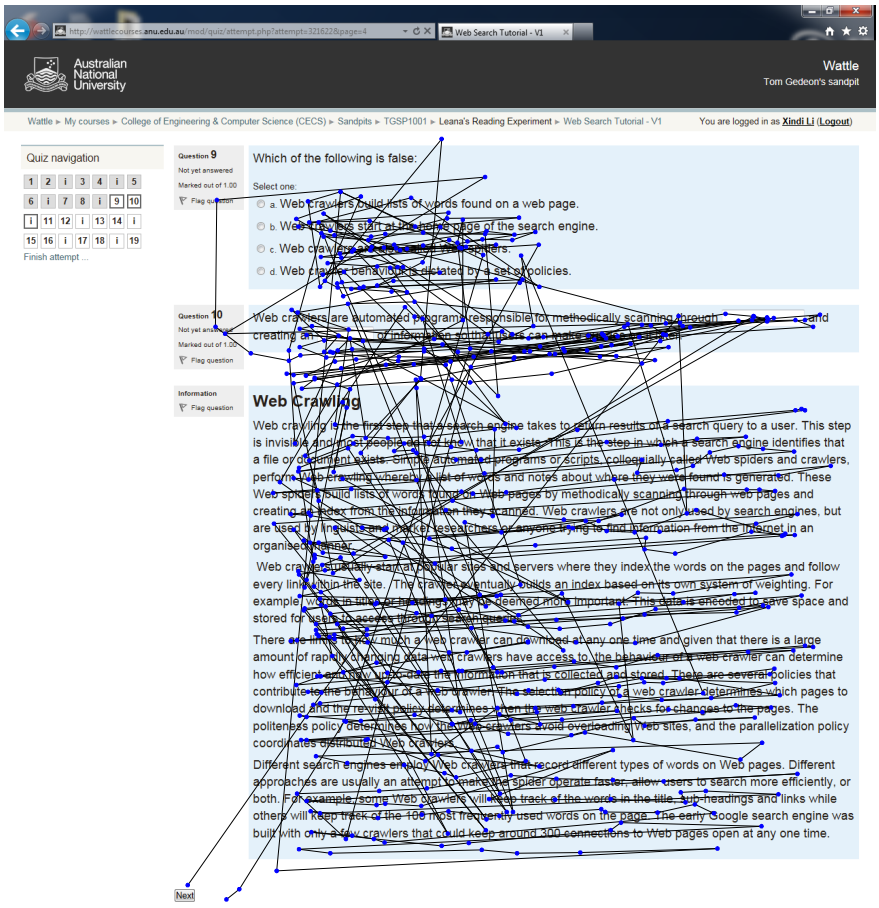


Figure 3.11. Example of eye movements from reading and answering questions on questions and text tutorial page for Format B (B: T/Q)

3.4 Discussion and Implications

The objective of this chapter is to investigate whether the presentation of text and comprehension questions in an eLearning environment the performance outcomes of participants and how those participants implicitly interact with the learning materials. Extending this question further we also investigate these effects on two groups of readers, L1 and L2 readers. The availability of learning materials to a wide and varied audience is becoming more common with the growth of online eLearning environments and online courses, such as MOOCs. More learners are reading materials written in their non-native language. The effects of this need to be explored further, the importance is only growing as accessibility to foreign language materials is becoming easier.

The results generally confirm that whilst L2 readers take longer to read content their comprehension is no different to L1 readers (Dednam et al., 2014; Kang, 2014). Delving deeper into there is a discrepancy in read times, we move to eye movements for insight. When reading text with no questions present, L2 readers are

observed to have higher numbers of fixations for longer durations than L1 readers. The divide between L1 and L2 readers is primarily due to fixation duration once the questions and text are presented together.

The hypothesis that format affects performance outcomes and eye movements was confirmed. The formats elicited distinct eye movement and reading behaviours. The presentation format can therefore be manipulated to promote specific behaviours. In the next subsection, recommendations are made based on these observations. There were some surprising results on the effects of presentation format that will now be discussed.

The scores from format D are somewhat surprising given that pretesting with multiple-choice questions has been shown to benefit subsequent learning (Little & Bjork, 2012). In fact, just memorizing the pre-test questions instead of answering them has been shown to improve recall of information (Little & Bjork, 2012). This is analogous to format D where participants were given the comprehension questions before they read the material to answer them. Participants were told that they should read the questions and were welcome to answer them if they wished. Yet our results showed no improvement in comprehension scores compared with the control presentation format which required participants to rely purely on their memory of the text to answer the questions (format C).

Furthermore, format A had two surprising effects on participants' behaviour. The first was that for this format there was no correlation between participants' quiz scores and their subjective ratings of understanding. This is surprising because for formats C and D, where the participants had to answer the questions without the text being available, there were very strong correlations. The effect of showing the text before asking the questions was believed to, at least, partially mimic these formats thereby partially enhancing the ability to subjectively rate understanding. This however was not the case. The second surprising effect of format A was that it was hypothesized that participants would read format C more thoroughly, in terms of fixations and read ratio than format A. This was not found either; instead participants read the text in format A as thoroughly as participants did in format C.

3.4.1 Recommendations for presenting text and assessment questions

This section outlines recommendations based on the observations from this study for 1) educators designing courseware in eLearning environments, and 2) design considerations for developers of eLearning environments. The analysis has established that the presentation of text and evaluation resources, such as quiz questions, impacts learning outcomes and reading behaviour. The presentation format can be manipulated to optimize the performance outcomes of students, thereby increasing their understanding.

Formats C and D were shown to promote more accurate self-assessment of comprehension, which minimizes both under- and over-estimation of knowledge. Formats A and C were shown to promote more thorough reading of the learning

materials compared to D, therefore in the context of learning this is a more optimal outcome. Given the aims promote thorough reading and accurate self-assessment format C is thus optimal.

The differences in eye movement measures and reading behaviours reflect the overall purpose and goals placed on the reader. If an educator wants to promote thorough reading, the goals placed on the reader should not be targeted with the use of quiz questions. In this case, students only read the parts of the text that they think contains the answers. However, not showing the text with the questions means that the students have to rely too heavily on short-term memory and this impacts their quiz scores. The happy medium is format A where the students are requested to read the text and then move on to answer the compression questions. Of course this raises the question of how to make students read the text before moving on to the questions and text page. This is where eye tracking can be utilized. The eye tracker can be integrated into the learning environment so that it can monitor reading behaviour. Once the student has read the text then the learning environment would allow the student to move on to the questions.

3.5 Conclusion and Further Work

The study presented in this chapter was designed to increase our understanding of how text and comprehension questions presented in eLearning environments affect eye movements and performance outcomes. These effects are investigated for L1 and L2 readers. We found that presentation of text and comprehension questions affect L1 and L2 readers consistently. There is a difference between L1 and L2 readers, where L2 readers take longer to complete the task. However, L1 and L2 readers are otherwise no different. Following on this observation we observe that L2 readers have consistently longer fixation durations and in the situation where reading is the primary task, L2 reader have more fixations than L1 readers.

Importantly, making participants rely on memory to answer assessment questions promotes more accurate subjective ratings of understanding. When participants are asked comprehension questions after reading the content and have no reference back to the text they can more accurately gauge their understanding. When shown the text with the assessment questions participants are unable to gauge their own understanding.

The primary finding is that different presentation sequences of text and comprehension questions affect performance outcomes and eye movements of participants. The order in which text and comprehension questions are presented to students can therefore be manipulated to optimize performance outcomes and / or reading behaviour.

A limitation of this study is that only two types of questions were investigated in this analysis, being multiple-choice and cloze questions. These are commonly used question type but not the only types generally available in eLearning environments, so further research should investigate what effect other question types have on the observed behaviour.

Further exploration of presentation formats on mobile devices would be beneficial given the prevalence of this technology, as this study only considers reading from a computer screen in a university setting.

One might ask why formats in which only question or only text were not used. In the former, we would be able to assess the answering behaviour and performance as a baseline to assess intuition and prior-knowledge. It would be quite informative to test. For the purposes of this experiment we were highly focused on reading behaviour of the text foremost, and questions second, so the case was omitted. However, follow-up should be run because the implications for adaptive eLearning are quite useful in that is we could predict from a student's eye gaze whether they know the answer to a question or are confident in answering a questions, the learning material could be adapted to help them. This information is useful in addition to the answer correctness as it would indicate areas which the student needs help with, or conversely, already excels at. With the latter case it would be interesting to examine baseline reading behaviour in the absence of any test to see what that behaviour look likes, and compare it to the observed behaviour in this chapter.

The next step in our investigation of how students read in eLearning environments involves predicting their reading comprehension from their eye movements. The first point of call is further analysis of this data set. In particular, a specific pattern in eye movements is observed for the questions and text pages. The unique eye movements seen for these pages are analysed further in the proceeding chapter as a method of estimating learning in eLearning environments, and therefore to provide feedback to developers of eLearning environments. Following on from this we investigate prediction of comprehension scores from eye movements. This would allow for the removal of comprehension questions as well as the dynamic change of textual material based on predicted behaviours.

Chapter 4

Answering Questions in eLearning Tutorials

"Google can bring you back 100,000 answers. A librarian can bring you back the right one."

— Neil Gaiman

In Chapter 3 we investigated how different sequences of text and comprehension questions affect eye movements and learning outcomes. Two of these formats, A and B, provided participants with the opportunity to read text whilst answering the questions. The eye movements that occur as a result of these presentation formats are characterised by transitions between the questions and the text to find or confirm the correct answer. We term these eye movements as answer-seeking behaviour. In this chapter we describe answer-seeking behaviour and present a method for measuring and comparing this behaviour. We propose using the degree of answer-seeking behaviour as an implicit measure of question difficulty. The end of the chapter explains how the use of eye movement to predict implicit question difficulty can benefit the design of eLearning environments. This chapter includes work that was presented at CogInfoCom 2013 (Copeland & Gedeon, 2013b) and work presented in IEEE Transactions on Emerging Topics in Computing (Copeland & Gedeon, 2015).

4.1 Introduction

Eye tracking provides the capability of providing feedback about students answering behaviour in eLearning environments. This feedback can then be used to monitor student learning behaviour as well as to improve learning materials. The use of eye tracking to analyse answering behaviour in eLearning environments has been attempted. Results are promising given that it has been shown that eye movement measures can be used to predict student performance on certain

problems such as answering physics problems (Chen et al., 2014). Eye tracking has also been used to analyse how multiple-choice questions are answered giving information on how eLearning environments can be designed in order to exploit such behaviour (Nugrahaningsih et al., 2013; Tsai et al., 2012).

We investigate eye movement measures for analysing reading behaviour and reading comprehension when text and comprehension questions are presented on the same page. The situation of presenting text on the same page as comprehension questions elicits unique eye movement behaviour. In this chapter we move toward answering the question of whether eye movements can be used to predict how difficult students' find a task and if this can be used to provide feedback about that task. Our objective is to identify eye movement measures that will be useful for providing feedback in eLearning. The central question for this chapter is therefore:

Can eye gaze be used to provide feedback about learning behaviour in eLearning environments for L1 and L2 readers?

Primarily, we hypothesise that there will be varying levels of answer-seeking behaviour for each question as well as between participants. It is for this reason we propose the use of this measure for evaluating text and question difficulty. Additionally, we propose using this measure for evaluating students reading and answering behaviour to analyse their ongoing performance.

There are two formats, A and B, under investigation. Given the results from the previous chapter, the eye movements observed will likely be affected by format. In format A there is pre-exposure to the text before being presented with the questions and the text. We hypothesise that this pre-exposure to the text will induce less answer-seeking than is observed when there is no pre-exposure (i.e. Format B).

In this chapter we further analyse a subset of the data from the user study described in Chapter 3. The data used is that collected for formats A and B (A: $T \rightarrow T/Q$; B: T/Q). As this is an extension of the user study presented in the previous chapter we omit the details of the user study methodology and description of data set. For further details on these data sets refer to Chapter 3. The bulk of the chapter is an analysis of answer-seeking behaviour. We present the uses of answer-seeking behaviour for feedback in eLearning environments. Finally, we will conclude and indicate how these results will be used in the future work of developing an adaptive eLearning environment.

4.2 What happens when text is presented with questions?

Formats A and B (A: $T \rightarrow T/Q$; B: T/Q) provide participants with the opportunity to check the text whilst answering the questions. Participants exhibit specific eye movement behaviours as a result. This analysis outlines an investigation of these behaviours. It begins with looking solely at the data from format A as it provides insight into answering behaviour after a participant has read the text. In the case of format B, participants have no knowledge of the text before being presented with

the questions and the behaviour exhibited is to find the answers. We then move to defining this behaviour as a method of providing informative feedback in eLearning.

4.2.1 Eye movements and answering behaviour: towards Answer-Seeking

Format A offers an interesting case where participants have already read the text before they see the questions. We have established in Chapter 3 that participants do read the text less on the questions and text page for format A than they do for format B. The hypothesis is that this would mean that participants would read the second presentation of the text less than the first.

For format A, there are negative correlations between each of the following measures to the score for the multiple-choice question: the number of fixations (r=-0.8), the total fixation time (r=-0.8), and the number of regressions (r=-0.7). The results are similar for the cloze questions, where there are negative correlations to the following measures to the score for the cloze question: the number of fixations (r=-0.8), the maximum fixation duration (r=-0.8), the total fixation time (r=-0.8), the number of regressions (r=-0.8), and the regression ratio (r=-0.8). These correlations indicate that participants tended to do worse on the quiz if they read both the questions and the second appearance of the text more. The definition of more reading is that there are high numbers of fixations and regressions as well as a longer total fixation time.

Additionally, there are correlations between the eye movement measures observed when reading each type of question and the reading behaviour seen when reading the second display of the content. We hypothesise that those participants who re-read the question more are having difficulty answering the question and would therefore exhibit similar behaviour when reading the text for the second time. We found positive correlations between the number of fixations observed for reading the multiple-choice question (r=0.7) and the cloze question (r=0.6) to the number of fixations observed for the second display of reading the content. Similarly, a positive correlation was found between total fixation time (r=0.7 and r=0.6, for multiple-choice and cloze questions respectively) and number of regressions (r=0.8 and r=0.7, for multiple-choice and cloze questions respectively) observed when comparing the eye movement recorded for the multiple-choice and cloze question to the second display of the content.

More reading is indicative of the participant's lack of understanding of either the questions or the content. Time spent reading questions and referencing text for the questions is related to the participant's understanding whereby longer time spent answering the questions indicates less understanding.

The participants who do not understand the question or the content well enough to answer the question, seek to find the answer by re-reading both the question and the content. We term this answer-seeking behaviour. Answer-seeking behaviour is indicative of the participant's lack of confidence in answering the questions. The

participant's confidence is related to his actual understanding of the content, his perceived familiarity with the subject matter, as well as his confidence in his or her abilities to answer the questions correctly.

4.2.2 Answer-Seeking behaviour

There are many reasons for why a participant would seek an answer, namely they do not know the answer and / or they are not confident with the answer. It is beneficial to measure such behaviour so that feedback can then be given based on the existence and the extent of answer-seeking observed.

Figure 4.1. Example of answer-seeking behaviour

This behaviour can be seen visually in Figure 4.1, where there are large jumps between the questions and the text regions, which are followed by extensive reading of the text. Reading only occurs in certain paragraphs of the text, where the participant thinks the answer is. There are also heavy fixations on the questions indicating re-reading of the questions. We propose measuring answer-seeking behaviour by recording the jumps between question and text regions and the reading behaviour recorded after each scan jump. The reading behaviour is detected

and recorded using a combination of reading detection algorithms (Buscher et al., 2008; Campbell & Maglio, 2001).

4.2.3 Effect of format on answer-seeking

We have now established a definition for answer-seeking behaviour and how to measure it. Answer-seeking behaviour can be recorded for formats A and B. We now investigate the hypothesis that pre-exposure to the text will induce less answer-seeking than is observed when there is no pre-exposure (i.e. Format B). The answer-seeking behaviour for each of the formats and each reader type are show in Table 4.1. The results show that there is little difference between the L1 and L2 readers, however more jumps are observed for format B as compared to format A. Furthermore, more answer-seeking behaviour is observed for the cloze questions compared with the multiple choice questions.

Table 4.1. Mean ± standard deviation answer-seeking behaviour for formats A and B (A: $\boldsymbol{T \rightarrow T/Q}$; B: $\boldsymbol{T/Q}$)

Format	Reader Type	Multiple Choice		Cloze	
		Jumps	Reading Saccades	Jumps	Reading Saccades
A	L1	8.3 ± 1.6	68.6 ± 17.2	10.2 ± 1.8	87.4 ± 18.4
	L2	6.3 ± 2.7	62.1 ± 28.5	11.1 ± 3.0	98.6 ± 30.5
B	L1	10.3 ± 1.6	136.6 ± 17.2	15.1 ± 1.8	124.4 ± 18.4
	L2	11.8 ± 2.7	142.7 ± 28.5	19.5 ± 3.0	160.1 ± 30.5

A MANOVA was used to test for statistical significance of eye movement measures between formats. The correlations between the dependent variables are all within the range of r=-0.4 and r=0.9. Levene's test for equality of variances shows that there is homogeneity for all dependent variables (p>0.05) and the Box's M value of 17.092 (p=0.207) is interpreted as non-significant so we can be satisfied that there is homogeneity in the variance-variance-covariance matrices.

There is a statistically significant difference in eye movement measures based on the presentation format to which the participant was exposed to, $F(4.23)=4.199$, $p=0.011$; Wilk's $\lambda=0.578$, partial $\eta^2=0.422$. There is no statistical difference between the L1 and L2 readers and no significant effect of interaction.

ANOVAs are used to determine how the eye movements differ for the formats. Format has a statistically significant effect on the number of reading saccades recorded after scan jumps for multiple-choice questions ($F(1,26)=10.006$, $p=0.004$), and scan jumps for cloze questions ($F(1,26)=7.050$, $p=0.013$), but not on multiple-choice scan jumps or cloze reading saccades after scan jumps.

For format B there is significantly more reading after a jump between the multiple-choice questions and the text than for format A. Yet this is not true for the cloze questions, contrary to the hypothesized behaviour. Additionally, format B has significantly more jumps between the cloze questions and the text than format A and once again this is not true for multiple-choice questions. Format does not affect

the reading behaviour for the cloze questions. This is an interesting outcome, reading the text before having access to the questions does not help participants when answering the cloze questions. Furthermore, reading the text before knowing the questions (Format A) does not reduce the number of jumps between the multiple-choice questions and the text.

The difference in the observed effect of pre-reading the text between the two types of questions is presumably due to the nature of the questions. Multiple-choice questions are more of a pattern matching exercise that promotes reading the multiple options and scanning through the text to try to find a similar phrase. We hypothesised that fewer of these jumps would be required for format A as the participants had already read the text and would presumably have some knowledge about the answers. The fact that there is significantly more reading after these jumps for Format B compared to A is supportive of this hypothesis.

Conversely, cloze questions require the participant to come up with a word to fill in the blank. This is comparative to a search task where the participants are looking to find words in the text. However, since they have to come up with a word, it does require a certain level of comprehension of the text in order to come up with the word so we would expect less jumping between the questions and text and more reading of both the questions and the text. This is what was found; for format A participants had significantly less jumps between the cloze questions and the text than format B. In this case, irrespective of the amount of reading, the fact that there are fewer jumps for format A confirms the hypothesis that reading the text before seeing the questions (format A) helps answer the cloze questions.

This analysis partly confirms the hypothesis that participants would on average show less answer-seeking behaviour when presented with format A compared to B. Although pre-reading the text before seeing the questions does not decrease the amount of answer seeking observed for the multiple-choice questions, it does for the cloze questions, which confirms the hypothesis for that kind of question.

There is no significant difference between the L1 and L2 readers in answer-seeking behaviour. This contrasts with the other analyses in the previous chapter where a clear difference was evident. This is an important finding: as the aim is to level the playing field for readers, and providing questions and text together will do so.

4.2.4 Using Answer-Seeking for Feedback in eLearning

We propose two purposes for measuring answer-seeking behaviour. The first is as a feedback tool for instructors and / or authors concerning the nature of how students read and answer questions. The second is to provide feedback to instructors about how individual students are performing.

4.2.4.1 Feedback about answerability of questions

This discussion will begin by elaborating on the first use mentioned above, feedback about comprehension questions for Format A. The average number of jumps

between question and text, the average reading behaviour and the average score for that question for both formats are shown in Table 4.2. Since the first part of the analysis showed that there was no statistically significant difference between L1 and L2 readers in this behaviour, we have not separately the reader types for this example.

Table 4.2. Answer-seeking behaviour averages per question for format A (A: $T \rightarrow T/Q$)

Quest. No.	Format A			Format B		
	Scan Jumps	Reading after jump	Score	Scan Jumps	Reading after jump	Score
1 (MC)	11.1 ± 9.5	85.9 ± 79.4	0.7 ± 0.5	9.8 ± 5.5	178.9 ± 118.3	0.6 ± 0.5
2 (CL)	12.2 ± 9.2	115.6 ± 118.6	0.9 ± 0.2	20.7 ± 15.9	166.6 ± 136.1	1.0 ± 0.0
3 (MC)	3.3 ± 3.4	23.9 ± 36.7	1.0 ± 0.0	6.5 ± 3.3	92.7 ± 55.7	1.0 ± 0.0
4 (CL)	5.9 ± 2.6	65.5 ± 52.8	1.0 ± 0.0	9.2 ± 7.3	105.7 ± 99.3	0.9 ± 0.2
5 (MC)	4.6 ± 4.6	36.9 ± 40.3	0.8 ± 0.4	12.5 ± 9.2	178.2 ± 96.6	0.9 ± 0.3
6 (CL)	4.3 ± 3.2	32.0 ± 20.8	1.0 ± 0.0	8.8 ± 5.5	50.1 ± 37.9	1.0 ± 0.0
7 (MC)	10.6 ± 8.0	58.6 ± 47.7	0.9 ± 0.4	9.7 ± 3.5	81.9 ± 79.5	0.9 ± 0.2
8 (CL)	19.5 ± 13.6	203.3 ± 173.5	0.9 ± 0.2	22.1 ± 10.1	214.2 ± 134.1	0.9 ± 0.2
9 (MC)	6.1 ± 5.0	73.5 ± 90.9	0.7 ± 0.5	14.7 ± 11.5	178.5 ± 101.0	0.9 ± 0.2
10 (CL)	11.2 ± 8.1	77.5 ± 55	1.0 ± 0.1	20.9 ± 7.6	150.6 ± 70.9	1.0 ± 0.0
11 (MC)	17.3 ± 12.9	148.1 ± 107.9	0.7 ± 0.5	18.4 ± 13.4	189.8 ± 102.8	0.5 ± 0.5
12 (CL)	17.3 ± 11.5	146.7 ± 120.6	1.0 ± 0.1	23.5 ± 13.1	202.7 ± 138.7	1.0 ± 0.0
13 (MC)	4.6 ± 4.1	57.3 ± 60.6	0.8 ± 0.4	9.8 ± 5.0	124.5 ± 79.4	0.7 ± 0.4
14 (CL)	7.9 ± 5.6	65.9 ± 80.4	1.0 ± 0.0	12.3 ± 9.9	90.8 ± 80.3	1.0 ± 0.0
15 (MC)	5.8 ± 7.7	64.8 ± 100.4	0.7 ± 0.5	8.7 ± 6.0	128.3 ± 64.8	0.7 ± 0.4
16 (CL)	7.1 ± 4.2	61.2 ± 46	1.0 ± 0.1	12.5 ± 6.9	107.1 ± 101.4	1.0 ± 0.1
17 (MC)	5.5 ± 4.2	33.7 ± 32.7	0.9 ± 0.3	5.9 ± 3.8	91.4 ± 79.3	0.8 ± 0.4
18 (CL)	9.3 ± 5.3	71.7 ± 42.8	0.9 ± 0.2	16.3 ± 6.3	117.3 ± 50.3	1.0 ± 0.0

There is a sizeable range in the average jumps for each question. As we found above the number of jumps observed for format B are higher than for format A. The same observation can be said the reading behaviour after each jump. For format A, there is a minimum average of 3.3 jumps and 23.9 reading transitions for question 3 and a maximum average of 19.5 jumps and 203.3 reading transitions for question 8. From these observations, question 3 was the easiest question on average to answer as fewest jumps and lowest amount of reading was needed to answer the question. Question 8 was the most difficult question on average to answer as the most jumps and the most reading were needed to answer the question. For format B however, the lowest number of jumps is 5.9 for question 17 but the least reading behaviour after a jump was recorded for question 6 at 50 reading transitions. The most jumps and reading behaviour were not recorded for the same question either where the most jumps of 23.5 was for question 12 and the most reading of 214 for question 8. Although the answer-seeking results for the formats do not exactly match one another, they are roughly similar. Questions appear to have similar relative answer-seeking observed under both formats. This is why answer-seeking should be recorded for both the jumps and reading behaviour after the jumps as it provides additional context to the answering behaviour.

On average, there were more jumps and reading observed for some questions; that is, more answer-seeking recorded for particular questions. This indicates that some questions are more difficult to answer than others. This difficulty could be for several reasons, such as ambiguity in the question or technical difficulty of the question. There is little correlation between the number of jumps or the amount of reading transitions observed to the score obtained for the question (r=-0.1, for both measures for format A and r=0.2, r=-0.3 for jumps and reading for format B). Therefore, performance on the question is not an accurate measure of how difficult the participants found the questions. Instead the answer-seeking behaviour is a measure of how difficult participants found the questions to answer as well as how much attention they gave to the question. We propose the use of answer-seeking behaviour be used in combination with answer correctness to describe how difficult a question is.

The large standard deviations seen in Table 4.2 show that there is a large variation in the observed answer-seeking behaviour. This is expected, as there is a large variation in eye movement behaviour observed between individuals (Rayner, 1998). Furthermore, we are only considering average performance on questions, as we would expect that some individuals would find questions easier to answer than others. This leads us to the discussion of using answer-seeking behaviour to quantify individual student performance.

4.2.4.2 Feedback about student performance

We now investigate the use of answer-seeking behaviour to analyse participant learning performance. We will only use the data from Format A as a case study for this proposed use as this is sufficient for showing its use.

The average number of region jumps between question and content, the average reading behaviour, and the total score for each participant are shown in Table 4.3. Note that in Table 4.3 the participants are listed in ascending order of average number of jumps. This ranking of participants' shows the extent of the variance of answer-seeking behaviour each participant exhibits. Once again we can use this information to extrapolate how difficult the individual participant found the tutorial and quiz.

As we would expect, the L2 participants did not have higher answer seeking than the L1 readers. The L2 readers are distributed between the L1 readers. Participant 15 showed quite a high amount of answer-seeking behaviour whilst the participant 1 showed about a seventh of the number of jumps and reading behaviour. There is a small negative correlation between the average number of region scans and the participants' total score (r=-0.3). This indicates that the participants who displayed less answer-seeking behaviour were not necessarily correct and may be over confident with their answers. Therefore, answer-seeking behaviour does not guarantee that the correct answer is selected by the participant, and neither does the lack of answer-seeking behaviour. This could be a by-product of the laboratory setting of the experiment where some participants, knowing they are being watched, will read more thoroughly and carefully to avoid making errors. Further investigation of this behaviour should be considered in an *in-the-wild* type

study. However, we propose the use of answer-seeking behaviour, in combination with answer correctness, as an implicit measure of how difficult a participant finds the text and quiz.

Table 4.3. Average answer-seeking behaviour per participant for format A (A: $T \rightarrow T/Q$)

Participant ID	Region jumps	Transitions classified as Reading	Total Score	L1/L2
1	2.9 ± 3.2	24.9 ± 34.9	17	L1
2	3.2 ± 3.6	23.7 ± 21.9	18	L1
3	5.0 ± 3.4	67.4 ± 68.3	17	L1
4	5.4 ± 4.7	46.6 ± 55	17	L2
5	6.4 ± 4.7	40.3 ± 38.9	16.5	L1
6	6.8 ± 5.8	47.1 ± 43.1	13	L2
7	6.9 ± 5.9	39.4 ± 43.4	17	L1
8	7.8 ± 6.8	95.8 ± 108.1	14.5	L2
9	8.6 ± 5.6	60.2 ± 51	18	L1
10	10.6 ± 11.4	89.7 ± 103.6	14.5	L1
11	11.4 ± 6.9	136.2 ± 57.9	15	L1
12	12.3 ± 8	56.7 ± 51.9	13.5	L1
13	12.3 ± 9.7	143.7 ± 163.8	16	L1
14	14.8 ± 8.6	139.4 ± 95.1	15.5	L2
15	22.1 ± 14.3	173.9 ± 129.6	16	L1

Once again there is high variation in the observations as shown by the standard deviations. This is a reflection of the differing difficulty of the 18 questions as already discussed and shown in Table 4.2. The standard deviations for each participant can be used to evaluate how consistently difficult that participant found the questions. For example, a low standard deviation indicates low variability and therefore that the participant consistently showed similar answer-seeking behaviour. This result indicates that the participant found each question to be similar in difficulty. This information can be used to construct questions of similar or differing difficulty.

4.3 Using Answer-Seeking Behaviour for Feedback

We have established a definition of answer-seeking behaviour of recording the large jumps between questions and text combined with the amount of reading that is performed in the question and text areas. We propose the use of this measure as an indicator to the instructor of question difficulty as well as the participant's implicit difficulty in completing the quiz. We will now establish the benefits of such information.

There is a range in the answer-seeking behaviour seen for each of the questions. This shows that some questions were harder to answer then others. The use of answer-seeking behaviour as a measure of question difficulty can be used as a

feedback system to an instructor. Such information can be used to gauge the difficulty of questions. This difficulty could be due to factors such as the technical nature of the material, and ambiguity in the material. Conversely, the instructor could see that the question is too easy and change it to be more challenging. This information could also be used to weight questions so that more difficult questions are weighted higher than those that are less difficult.

Furthermore, there is a range of answer-seeking behaviour seen among the participants. Some found the quiz more challenging than others. It is beneficial for learning for students are challenged equally in respect to one another, so that some students aren't being under-challenged whilst others are over-challenged. Under-challenged students may get bored and lose interest in the material whilst over-challenged students may become anxious and disheartened by the material. In either case, there is a negative impact on the learning process. Using answer-seeking behaviour as an implicit measure for a student's confidence in the material can provide the framework for an adaptive online learning environment. Such an environment can use input from the eyes to measure the answer-seeking behaviour and alter the learning material and questions in response to the student's behaviour. That is, if a student is found to be having no difficulty completing a quiz, then the material can be altered to be more advanced and technical. Conversely, if a student is having difficulty then the material can be altered to be less technical and more basic.

4.4 Conclusion and Further Work

In this chapter we have investigated answer-seeking behaviour as a method of evaluating text comprehension for a tutorial and quiz. Answer-seeking behaviour is the eye movement behaviour exhibited when students are presented with questions and text on the same page. Answer-seeking behaviour is characterised by jumps between the questions and the text to find the correct answer, or to reassure the participant that they have the correct answer. We hypothesised that the pre-exposure to the text before being asked the questions would affect the reading behaviour observed when presented with the text and questions, and therefore induce less answer-seeking than is observed when there is no pre-exposure (i.e. Format B). However, we found that pre-exposure to the text does not decrease answer-seeking behaviour for multiple-choice questions, although it does for the cloze questions, which partly confirms the hypothesis.

An interesting point found from the study was confirmation of the hypothesis that the presentation format affected the L1 and L2 participants in the same way. Additionally, there is no significant difference between the L1 and L2 readers in answer-seeking behaviour. This is an important finding as it means that any conclusions regarding how presentation format affects students can be generalized for both reader types and it is not an additional factor that creators of learning materials have to take into account.

Additionally, we hypothesised that there would be varying levels of answer-seeking behaviour for each question as well as between participants. We have

proposed the use of answer-seeking behaviour to describe how difficult a question is to answer and as an implicit measure of how difficult a participant finds the tutorial and quiz. The eventual goal is to create a tool that will provide feedback to instructors about implicit behaviour of students performing a reading task through an online learning environment. For example, if the instructor receives feedback that multiple students are failing to understand specific parts of the text then the instructor can dedicate more time explaining these concepts during face to face teaching time, or could re-word the content to make it easier to understand. Furthermore, the instructor can be given feedback about how students are reading questions and be able to deduce if questions are appropriately worded or are ambiguous and hence causing low scores or confusion. Finally, the information about reading behaviour can also be used to dynamically alter tutorial content to personalize the learning experience where students familiar with or excelling at specific content can be given more advanced content to read compared to students that are not familiar with the content or who find it harder to understand.

As stated in the previous chapter, a limiting factor of the study is the use of only two question types. The answer seeking behaviour from multiple choice and cloze questions varies so it is pertinent that different questions be assessed. Additionally, participants were not asked how difficult they found the questions to answer. This is a limitation of the study that we deal with in Chapter 7 where we investigate participants' perceptions of difficulty.

The results from this study will be used as the foundation for uses of applying eye tracking in adaptive eLearning. Eye tracking can be used to determine learning rates and behaviour during reading so that learning can be adapted to students' needs as well as increasing the quality of the materials in the environment. The next step in the investigation is to predict reading comprehension scores from eye movements, including answer-seeking behaviour.

Chapter 5

Effects of Presentation on Prediction of Comprehension

Predicting reading comprehension from eye gaze data is a difficult task. In this chapter we investigate the effects of presentation format on prediction accuracy of reading comprehension measures. The data from the user study outlined in Chapter 3 is used to explore the problem of predicting reading comprehension from eye gaze using machine-learning techniques. Chapters 3 and 4 established that presentation format affects eye movements and reading comprehension. The hypothesis examined in this chapter is that the different formats will cause different levels of prediction accuracy. The investigation begins by using artificial neural networks (ANNs) to predict reading comprehension scores. To help increase prediction accuracy of the ANN we investigate the use of fuzzy output error (FOE) as an alternative performance function to mean square error (MSE) for training ANNs as a means of improving reading comprehension predictions. The results show that the use of FOE provides more accurate predictions. Additionally, the FOE trained ANN outperforms other comparison machine learning techniques. Nevertheless, we deduce there are complex relationships between eye movements and reading comprehension as three hidden neuron layer ANNs provided the best classification results. We encountered problem with imbalanced data sets that requires further investigation. In this chapter we present research that extends work presented at ICONIP 2014 (Copeland, Gedeon, & Mendis, 2014a) and based on work published in AIR journal (Copeland, et al., 2014b).

5.1 Introduction

In this chapter the data from the user study conducted in Chapter 3 is used to investigate methods for predicting reading comprehension from eye gaze using machine-learning techniques. The task of predicting quantified measures of reading comprehension has been attempted with poor results (Martínez-Gómez & Aizawa, 2014). Additionally, little prior work has been done to predict reading comprehension via machine-learning techniques. Prediction of reading comprehension has classically been made using statistical analysis of eye movement measures that have been derived from the eye gaze signal such as fixation duration (Underwood et al., 1990) and regressions (Rayner et al., 2006). Current applications of eye tracking in reading analysis only take into account assessment of reading behaviour such as using fixation time to predict when a user pauses on a word (Hyrskykari et al., 2000; Sibert et al., 2000) and finding word relevance (Loboda, Brusilovsky, & Brunstein, 2011). Instead, we look at combining eye movement measures to make more complex predictions about reading behaviour. The central question being asked in this chapter is:

Can eye tracking data be used to predict reading comprehension scores in eLearning environments for L1 and L2 language readers?

However, we know that predicting reading comprehension is not trivial so we approach this question in two chapters, this chapter and the next. In each of the two chapters we investigate factors that could affect prediction accuracy. In this chapter we also investigate the effect of text presentation on prediction accuracy of comprehension:

Does presentation of text affect predictions of comprehension?

We explore this question by investigating different methods of increasing prediction accuracy. Initially, we build on previous work that involves prediction of reading comprehension from eye gaze using fuzzy output error (FOE) as the performance function for back-propagation training of feed-forward artificial neural networks (ANNs) as this showed promising results (Copeland, Gedeon, et al., 2014a). We extend this research by exploring different membership function shapes (FMFs) for calculating FOE and compare these results to using mean square error (MSE) as the performance measure for training. The next part of the analysis is comparison of the results from the ANN classification to comparative classification techniques to deduce whether this is the optimal technique. We then move to assess prediction of the different questions types, multiple-choice and cloze. Finally, we perform cluster analysis of the more complex formats.

As this chapter uses the eye gaze data collected from the user study explained in Chapter 3, we will not restate the details of the methodology. Please refer to Chapter 3 for more details about the user study. This chapter begins with a background review of classification and clustering techniques that will be used throughout the chapter, which is followed by an introduction to fuzzy output error; method for

analysis; followed by a discussion and implications for eLearning; finally, the conclusion and further work are outlined.

5.2 Making Predictions

Prediction involves modelling input data to produce an output that reflects the input in some way. Traditionally, types of machine learning problems are dependent on the learning strategy used. These are primarily supervised, unsupervised and reinforcement learning (Russell, Norvig, Canny, Malik, & Edwards, 2003). Supervised learning is where outputs are known and used to train the model. Unsupervised learning is where the outputs are not known so a structure in the inputs has to be found by the learning algorithm. An example of unsupervised learning is clustering. This section provides background information on several prediction techniques, which are used in this chapter. However, this is by no means to full coverage of prediction methods and is rather a short list of some commonly used methods.

The main analysis is the use of artificial neural networks (ANN) to which we compare to decision tree based and k-nearest neighbour (kNN) classifiers as they are commonly used. The advantage of using an ANN is that they allow for prediction of two output values, which suits the problem given that there are two comprehension questions. Additionally, the use of ANNs has shown promise for this type of problem (Copeland et al., 2014a). The problem of classifying reading comprehension from eye movements is difficult; indeed a three-hidden layer topology generates the optimal predictions, namely the [12 6 3] topology for both FOE and MSE (Copeland et al., 2014a).

5.2.1 Decision Trees

Decision trees are commonly used predictive models that map inputs to outputs. Two types of decision tress are classification and regression trees. There are several learning algorithms for constructing decision trees such as CART (Breiman, Friedman, Stone, & Olshen, 1984), ID3 (Quinlan, 1986), and C4.5 (Quinlan, 2014). Decision trees are easy to interpret and fast to learn.

5.2.2 Ensemble Learning

The premise of ensemble learning is the use of many weak learning algorithms in combination to improve predictive power (Rokach, 2010). An ensemble combines a set of supervised learning algorithms, but is itself a supervised learning algorithm as it is trained to make predications. Common weak learners are decision trees and k-nearest neighbour as they are quick to train. However, ensembles can be made from any predictor such as ANN (Hansen & Salamon, 1990). Bagging and boosting are common ensemble techniques (Rokach, 2010).

5.2.3 Boosting

Most boosting algorithms consist of making currently misclassified examples more important in the next round of classification. Therefore, a new weak learner is added that focuses on previously misclassified examples (Freund et al., 1999). Boosting is quite often applied to overcome the problem of imbalanced data sets such as using under-sampling, over-sampling, and other forms of sampling to reduce the imbalance. AdaBoost is a common boosting algorithm used for binary classification (Freund et al., 1999), which has been extended for multi-class situations (Zhu et al., 2009). The boosting algorithm used in this chapter is RUSBoost (Seiffert et al., 2010), which uses a mix of random undersampling (RUS) and boosting to deal with imbalanced data sets, a problem prevalent in the data sets used for this analysis.

5.2.4 Bootstrap Aggregation (Bagging)

Bootstrap aggregation, also known as bagging, is the straightforward strategy of creating multiple predictors for multiple subsets of observations and / or features sets (Breiman, 1996, Rokach, 2010). The bootstrap samples of the data set are chosen at random with replacement from the training set. The result is many diverse classifiers that are aggregated together to find the end prediction of an unseen sample. In the case of regression, the aggregation is an average of the predictors' outcomes and majority vote for the case of classification.

5.2.4.1 Random Forests

Random forests are a special case of bagging where bagging of both observations and feature sets is performed (Breiman, 2001). As described above, bootstrap samples of observations are generated from which decision trees are constructed for each sample. The decision tree algorithm is modified so that at each candidate split in the decision tree learning process a random subset of features is used as the pool of options for the split. If there are features that are strong predictors of the output variable, then many trees will include this feature thereby being correlated. The modification of the bagging method to include feature bagging helps alleviate the correlation of trees thereby making a stronger ensemble.

5.2.5 K-Nearest Neighbour (kNN) Classification

The k-nearest neighbour (kNN) algorithm for classification is conceptually quite easy to understand. The algorithm works by having a set of k training instances for which the test instances are compared (Peterson, 2009). The training instance(s) that are closest to the test instance, as defined by a distance metric such as Euclidean distance, are used to vote for the winning class. In the case of regression, an average is used instead of the mode.

5.2.6 Agglomerative Hierarchical Clustering

Hierarchical clustering is a clustering technique that builds hierarchies of clusters. Agglomerative refers to a bottom-up approach, which all data instances start in

their own cluster that are then merged together into larger clusters, until they are in the same group (Xu & Wunsch, 2008). In order to merge the smaller clusters into the larger clusters a distance metric is used to measure dissimilarity between the clusters, or at the base level, data instances (Rokach & Maimon, 2005). Typical distance metrics include Euclidean, Manhattan, hamming, and cityblock distances. Comparing the distances using a linkage criterion performs merging of clusters. Typical linkage methods are single-, complete-, and average-link clustering (Rokach & Maimon, 2005). This choice of distance metric and linkage criterion affects the clusters that are produced from clustering.

5.2.7 Back-propagation Artificial Neural Networks (ANN)

ANNs are models of biological neural networks. The basic idea of an artificial neuron is that a set of weighted inputs are fed into a neuron and summed together. This sum is called the neuron's activation and if it is above a threshold then the neuron fires. This is analogous to a biological neuron. To calculate if the activation is above a threshold an activation function is used, which can be as simple as a binary threshold function but most commonly sigmoid functions are used (Jain et al., 1996).

Many artificial neurons can be joined together to create networks of artificial neurons. Feed-forward networks are those where there are no loops, and the inputs follow in only one direction, forward. These networks need to be trained to perform the tasks for which they are intended. This usually involves the learning of weights. Different learning algorithms are required for different network architectures (Jain et al., 1996). The most common method of training multi-layered feed-forward ANNs is using the back-propagation algorithm, which is a supervised learning algorithm. These types of networks can be used to perform classification or regression tasks. Back-propagation works by passing a training example through the network, calculating the error from the results output, and using this error to change the weights. Thus, the error is propagated back through the network.

5.3 Fuzzy Output Error (FOE)

Fuzzy Output Error (FOE) (Gedeon et al., 2012) is an extension of FYCLE and SYCLE (Mendis & Gedeon, 2008). FOE uses a fuzzy membership function to quantify the difference between the predicted and the target values, i.e. the error, rather than assign the difference a value of 0, 0.5 or 1, as is done in FYCLE. As opposed to MSE, FOE describes the error in a fuzzy way and then sums the fuzzy errors together to get the total error.

FOE is defined as follows for a data set of n records with matching pairs of target and predicted values for each record 1 to n.

$$FOE = \sum_{i=1}^{n} 1 - \mu(\hat{y}_i - y_i) \ where\ n \in \mathbb{N}. \quad (1)$$

where $\mu()$ is the membership function of a desired classification and its complement describes the error. The membership function is termed the FOE Membership Function (FMF). The FMF is used to describe the output of a fuzzy classification (or a regression) in regards to how close that output is to the target

output. The membership function itself represents the fuzzy set for good classification[15]. The value of $\mu(x)$ gives the degree of membership of the error in the good classification fuzzy set and consequently the complement of $\mu(x)$ gives the error measure. Therefore, $\mu(\hat{y} - y) = 1$ and hence there is no error when there is perfect classification. The more $\mu(x)$ tends toward 0 the higher the error since the difference is larger. The FMF shapes used in this analysis will be trapezoidal or triangular membership functions. FMF's can be created in any shape in order to describe the output of a function.

It is important to note that the difference between target and predicted values is not taken as the absolute value of the difference (i.e. $\hat{y} - y|$). Although this would make the FMF simpler because it would only need one side of a piecewise linear function, not using the absolute value of the difference provides more flexibility in describing the types of error. For example, false negatives may be considered a much worse error than false positives when screening for diseases.

5.3.1 Approximation of FMFs using squashing functions

There are many different ways to construct membership functions as described in (Dombi, 1990), however, commonly piecewise linear functions are used as they are easy to handle (Dombi & Gera, 2005). The problem with these functions is that optimisation of parameters via gradient-based methods become complicated, as they do not have continuous derivatives. One of the solutions to this problem is to approximate piecewise linear functions using combinations of sigmoid functions called squashing functions (Dombi & Gera, 2005, 2008; Gera & Dombi, 2005).

A sigmoid function is an s-shaped function that is commonly used as an activation function of artificial neurons, as well as in economic and biological models. The definition of a sigmoid function is shown in Equation 2.

$$\sigma_{\alpha}^{\beta}(x) = 1/(1 + e^{-\beta(x-\alpha)}) \tag{2}$$

The parameter β controls the steepness of the sigmoid curve, that is, varies the function from a shape either close to linear or more like a step function. The parameter α controls where the centre of the curve, $\sigma(x) = 0.5$, is on the horizontal axis. More precisely, $x - \alpha$ will move the centre to α and $x + \alpha$ will move the centre to $-\alpha$. These two parameters play an important role in how the sigmoid function will be shaped to approximate the piecewise linear membership functions.

To approximate one half of a trapezoidal or triangular function, we integrate the difference between two sigmoid functions on an interval [a, b] (József Dombi & Gera, 2005, 2008). The definition of the squashing function on interval [a, b] is shown in Equation 3.

$$S_{\alpha,\delta}^{\beta}(x) = 1/2\delta \ln\left(\sigma_{\alpha+\delta}^{-\beta}(x)/\sigma_{\alpha-\delta x}^{-\beta}(x)\right)^{1/\beta} \tag{3}$$

[15] Good classification refers to a level of error between the predicted and the desired that is within a threshold that users accept as either correct or close to correct classification.

Where α gives the centre of the squashing function and δ gives the steepness of the squashing function. The parameter δ is referred to as the fuzziness parameter and β the approximation parameter. The larger β is the closer the approximation to the trapezoidal function being modelled.

A piecewise linear membership function can therefore be approximated with the combination of two squashing functions using the conjunction operator. The following equation defines the approximation of a trapezoidal membership function (József Dombi & Gera, 2005, 2008).

$$S_{\frac{1}{2},\frac{1}{2}}^{(\beta)}(S_{a_1,d_1}^{(\beta)}(x) + S_{a_2,d_2}^{(-\beta)}(x) - 1) \tag{4}$$

When $a_1 = d_1 = -½$ and $a_2 = d_2 = ½$ the squashing function approximates a triangular membership function. All FMF shapes are represented in this form throughout the analysis so that gradient descent methods can be used to optimise the error function.

5.3.2 FMF shapes used to calculate FOE

In this chapter we utilise 7 FMF shapes, denoted as FMF1 through to FMF7. FMF1 (Figure 5.1(a)) is designed to be a cross between FYCLE and the shape of an MSE curve. The difference between the predicted value and target value is within ± 0.2 so is not considered an error and therefore considered correct classification. The difference between the predicted and target value is considered to be erroneous after ± 0.2. FMF2 (Figure 5.1 (b)), is designed to be a model of FYCLE. FMF3 (Figure 5.1(c)) is a triangular membership function that is designed to resemble the shape of an MSE curve. The difference between target and predicted values is a lower value for membership in the good classification set the further the difference progresses to -1 or 1.

FMF4 (Figure 5.2(a)) and FMF5 (Figure 5.2(b)) are asymmetrical FMFs that are inverses of each other. They are both a combination of half of FMF1 with the opposite half of FMF2, and were trialled to investigate the effect of asymmetric FMFs, which may have benefit in some applications. The shape of FMF6 (Figure 5.2(c)) is a variant of the FYCLE approximation FMF2. It has a smaller region that defines the difference between the predicted and target values as being completely in the good classification set, i.e. $\mu(\hat{y} - y) = 1$. This region is when the difference is between ± 0.1 instead of ± 0.2. Again, this is to make the error output closer to zero as described above. FMF7 (Figure 5.2(d)) is a variation of FMF1 but is also a combination of FMF1 and FMF3. Again the variation is that there is a smaller region that defines the difference between the predicted and target values as being completely in the good classification set, i.e. $\mu(\hat{y} - y) = 1$. This region is when the difference is between ± 0.05 instead of ± 0.2.

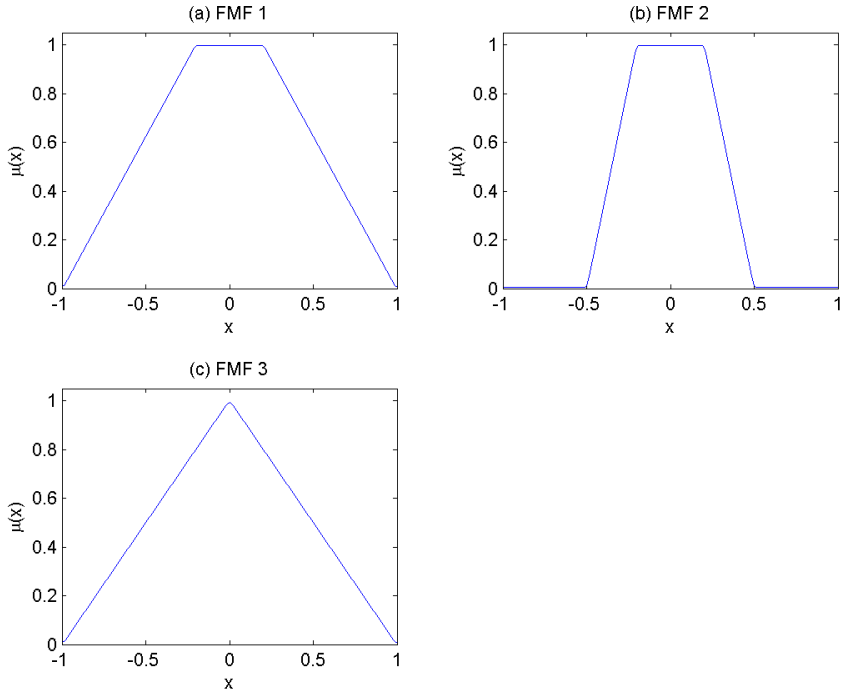


Figure 5.1. Plots of : (a) FMF1; (b) FMF2; and (c) FMF3

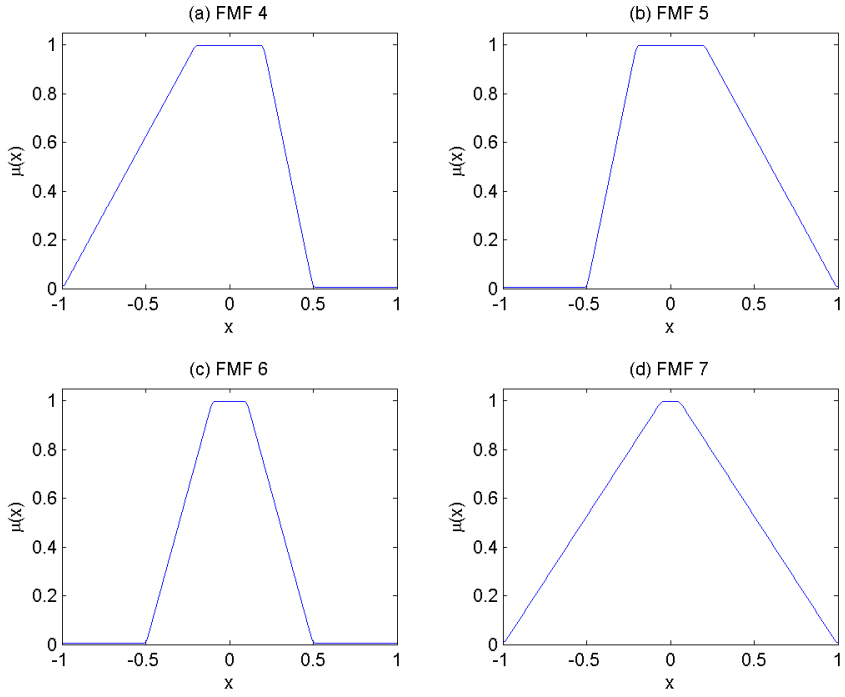


Figure 5.2. Plots of (a) FMF4; (b) FMF5; (c) FMF6; and (d) FMF7

5.4 Description of data sets

For further details on the user study methodology and the associated data sets please refer to Chapter 3. The raw eye gaze data consists of x,y-coordinates recorded

at equal time samples (60Hz). Fixation and saccade identification was performed on the eye gaze data. Eye movement measures are derived from the fixation and saccade data, which are explained in Chapter 3. These measures are:

- Number of fixations,
- Maximum fixation duration (seconds),
- Average fixation duration (seconds),
- Total fixation duration (seconds),
- Number of regressions and regression ratio,
- Average forward saccade length (pixels) and,
- Reading analysis statistics (read, skim and scan ratios).

We include two additional measures not explained in Chapter 3; these measures are:

Regional Analysis: The fixation-to-word and duration-to-word ratios are measured for the paragraphs where the answers are located. These are measures for how long the participant spent in the area containing the answer to the question. The hypothesis is that there is a relationship between the proportion of attention participants give to the answer paragraphs and the answers they provide.

Answer-seeking behaviour: The behaviour of jumping between the questions and the text to find the answers, discussed in Chapter 4.

Table 5.1 summarises the properties of the data sets used in the predictive analysis. Starting from the first row in the table, the features refer to the number of inputs for the ANNs, and other classifiers. These features are the eye movement measures just discussed and vary depending on the presentation method as the inputs are generated from the pages that the participant viewed. This means that for format A, as the participants view the tutorial content page and then the questions and content page, the inputs are generated from both pages for the scores obtained from the questions and content page. Note that since there is a large difference in the ranges for each of the inputs they are normalized to a range of [0,1].

Table 5.1. Properties of each data set (A: $T \rightarrow T/Q$; B: T/Q; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$)

Properties	Format A	Format B	Format C	Format D
Features	36	20	28	40
Size (N)	135	135	135	135
Comprehension scores				
Class Split %: 2/1.5/1/0.5/0	74/7/18/1/0	77/1/22/0/0	44/19/21/7/9	45/8/39/1/6

** The class split refers to the split in marks, i.e. for format A, 81% of participants answered the multiple-choice questions correct and 92% of participants answered the Cloze questions correct.*

The size of each data set is consistent at 135 instances; this is because there are 15 participants for each format who each viewed 9 topics. The final row refers to the

distribution of outputs. The outputs are the total comprehension score from the questions. The classes of comprehension scores are 2, 1.5, 1, 0.5, and 0. The outputs are assigned based on the answers to the multiple-choice and cloze questions. That is, the multiple-choice score can take values of 0 or 1, corresponding to an incorrect or correct answer for the question. Similarly, for the cloze question except that in this case half marks can be achieved so the output that is assigned can take the values 0, 0.5 or 1. The reason for half marks being achieved for the cloze questions is because two gaps requiring a word each for each question, so if a participant got one word correct and not another, they received 0.5 out of 1.

As shown in Table 5.1 the ratio of the number of data instances in each class varies considerably between the formats. We can observe that for formats A and B there are imbalances in the scores for cloze questions, where most people answered the cloze questions correctly for these formats. The cloze question scores for formats C and D are less imbalanced as more participants answered the questions incorrectly. For the multiple choice questions there are slight imbalances in scores for all formats, where the majority of people answered these questions correctly.

5.5 Results and Analysis

The analysis consists of three components; first is the use of ANNs to predict reading comprehension from eye movements, second is the use of common classification techniques to predict reading comprehension, and finally cluster analysis. The results section is organized to reflect this analysis.

5.5.1 ANN predictions of reading comprehension

The focus of the analysis is on finding a satisficing technique for prediction of reading comprehension from eye gaze data. In previous work we considered the use of a novel performance function for training of the ANNs called Fuzzy Output Error (FOE) (Copeland et al., 2014a). This showed promising results that led to improved predictions. We investigate the use of FOE further in this investigation. In the previous work only one FMF shape was investigated (Copeland et al., 2014a). We extend this investigation to look at the use of 7 FMF shapes to calculate FOE, described in Section 5.3.2.

The ANNs were trained using the scaled conjugate gradient algorithm (Møller, 1993) with the performance function set to be either FOE or MSE. From this point on we denote ANNs trained using FOE as FOE-ANN and ANNs trained using MSE as MSE-ANN. The analysis is performed Matlab R2012a using the Neural Network toolbox. FOE was implemented as a custom performance function. The default training method is the Levenberg-Marquardt algorithm (Hagan & Menhaj, 1994) however this training method will not accept custom performance functions. The scaled conjugate gradient algorithm has been shown to perform faster than other methods available (Møller, 1993).

The number of inputs for each presentation format is outlined in Table 5.1 and all networks have 2 outputs. From initial testing it was found that a single layer

network performed poorly for both FOE and MSE. We have chosen two and three layer topologies for the analysis. The following topologies were tested: [10 5], [20 10], [30 15], [12 6 3], [16 8 4], [20 10 5], and [30 20 10]. The notation [X Y Z] indicates neurons in the first hidden layer to the third hidden layer. The analysis revealed that the topology that generates the best predictions is [12 6 3] for both FOE and MSE.

The results from this analysis are reported in Table 5.2, as the average and standard deviations of misclassification rates (MCR) from 10-fold cross validation with standard deviations. The analysis revealed that the topology that generates the best predictions is [12 6 3] for both FOE and MSE. We restrict our presentation of these results to report only average results for all topologies and the optimal topology.

Table 5.2. Misclassification rate (MCR) comparison: FOE versus MSE as the performance function for ANN training (A: $T \rightarrow T/Q$; B: T/Q; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$)

		FOE-ANN		MSE-ANN
Format		FMF	Mean±std	Mean±std
A	[12 6 3]	3	0.15±0.03	0.27±0.08
	Average		0.23±0.12	0.27±0.10
B	[12 6 3]	5	0.12±0.02	0.20±0.10
	Average		0.22±0.09	0.27±0.10
C	[12 6 3]	2	0.49±0.09	0.63±0.10
	Average		0.57±0.12	0.64±0.11
D	[12 6 3]	7	0.57±0.10	0.62±0.11
	Average		0.58±0.11	0.64±0.13

On average the MCR from FOE-ANN is lower than from MSE-ANN as the performance function, for all formats. These results are an improvement on the results from previous work where FMF2 was used to calculate FOE (Copeland, Gedeon, et al., 2014a). We found previously that on average the MCRs for formats A and B were both 0.28. We have improved these results, in particular, for format A when the [12 6 3] topology is used an average across the cross validation results of 85% correct classification is achieved (MCR=0.15). This is a 46% reduction in MCR compared to when MSE is used (MCR=0.27). Similarly, for Format B when the [12 6 3] topology is used an average of 88% correct classification is achieved (MCR=0.12), which is a 39% reduction in MCR compared to when MSE is used (MCR=0.2).

The results from this analysis indicate that prediction of reading comprehension using ANNs is most accurate for Formats A and B. Indeed, there are quite high MCRs for formats C and D. The results can be compared to a simple majority prediction in which the highest class is predicted, where for format A the largest class of output is a score of 2 which is 74% of all scores and a total of 4 output classes. We are able to outperform this by gaining an average of 85% correct

classification. For format B the largest class of output is a score of 2 as well and this accounts for 77% of the scores, yet we are able to obtain an average accuracy of 88%. For format B there are only 3 output classes of which one of the class' only accounts for 1% of the total set of outputs so it is highly unlikely that we are predicting this class effectively and so further research should go into better predicting these minimal classes. However, in both cases we are not simply predicting the majority class of output.

The breakdown of outputs for formats C and D are much more spread out with both formats having 5 classes of outputs. The majority output being 2 for both formats, with of all outputs being 2 being 44% and 45%, for formats C and D respectively. We obtained quite poor classification results for formats C and D. Using FOE as the performance function, for format C we were able to achieve 51% correct classification (MCR=0.49) and for format D we were able to achieve 43% correct classification (MCR=0.57). Given that this is in total a 5-class classification task chance classification is 20%, thus we are achieving double chance classification. However, when taking into consideration the output class breakdown we are achieving above majority prediction for format C, but not for format D, for which classification is actually below the majority class. Further investigation of format D is required.

We hypothesised that it would be easier to predict reading comprehension from format D compared to format C as participants had knowledge of the questions before they read the text. They therefore know what it is that they are looking for in the text. However, from this analysis we cannot confirm this hypothesis. In the next sections we investigate methods for improving these classification results.

5.5.2 Comparison to other classification techniques

The second part of the analysis is the comparison of other classification techniques to the ANN results. The four supervised learning techniques are classification trees, boosted classification trees, random forests, and k-nearest neighbour. The average and standard deviations of the MCRs from 10-fold cross validation are reported for each format are reported in Table 5.3.

The results reported in Table 5.3 are suboptimal compared to using ANNs. The MCR values are double that from using ANNs for Formats A and B. Of the classification techniques used, the random forest ensemble produces the lowest MCR rates for all formats. As in the first analysis, Formats C and D both have poor classification results.

However, the results are that on average the random forest are not performing much better than a majority classifier. For example, for format A on average 71% correct classification (MCR=0.29), but the majority answer output accounts for 74% of the outputs. It could be that these techniques are not optimal for dealing with the particular data sets so further investigation of other machine learning techniques should be considered, such as support vector machines and algorithms for training ensembles that primarily deal with imbalanced data sets.

Table 5.3. Comparison of Misclassification (MCR) results for predicting total comprehension scores for all eye movement measures (A: $T \rightarrow T/Q$; B: T/Q; C: $T \rightarrow Q$; D: $Q \rightarrow T \rightarrow Q$)

Classification Technique	Format			
	A	B	C	D
Classification Tree	0.32±0.10	0.39±0.15	0.76±0.10	0.57±0.13
Boosted Classification Tree	0.47±0.13	0.40±0.19	0.73±0.09	0.78±0.13
Random Forest (Classification)	**0.29±0.08**	**0.30±0.13**	**0.61±0.05**	**0.46±0.13**
kNN	0.37±0.12	0.39±0.12	0.65±0.11	0.55±0.13
*Best FOE-ANN Result**	*0.15±0.03*	*0.12±0.02*	*0.49±0.09*	*0.57±0.10*

Note in bold are the lowest average MCRs for each format from the comparison techniques.

5.5.3 Cluster Analysis

Making predictions on the eye gaze data collected for Formats C and D has proven to be quite challenging. Exploration of these data sets using clustering is performed to see if there are any natural clusters in the data to which we can apply classification techniques and from which we can make conclusions. In both cases agglomerative hierarchical clustering is used. This type of clustering starts with every observation in its own cluster and then merges the groups together until they are in the same group (Xu & Wunsch, 2008). In both cases the distance measure is set to *cityblock* and linkage set to *average*. The eye movement measures from both the text page and the questions page were used in the clustering. The Statistics Toolbox in Matlab R2012a is used to perform the cluster analysis.

5.5.3.1 Format C ($T \rightarrow Q$)

The results of the clustering are shown in Figure 5.3. From the clustering there is evidently an outlying point. The outlying point has 691 fixations recorded for reading the tutorial page and a total fixation time of 176 seconds. This is above what is expected given that the text contains 400 words. If this outlier is removed and the rest of the data set is considered, there are three unequal clusters of data at a high level. Comparisons of the averages of eye movement measures for these clusters are shown in Table 5.4.

The results from the cluster analysis indicate that Cluster 1 represents the readers who spent little time to read the text. We infer that they spend less time to read due to the combination of low numbers of fixations, total fixation times and reading ratios, as well as longer forward saccades and higher skimming and scanning ratios. This cluster also corresponds to higher question scores. This would indicate that this cluster has grouped together the instances where there was prior

knowledge of the subject and so less reading is needed to achieve understanding, and therefore high comprehension scores.

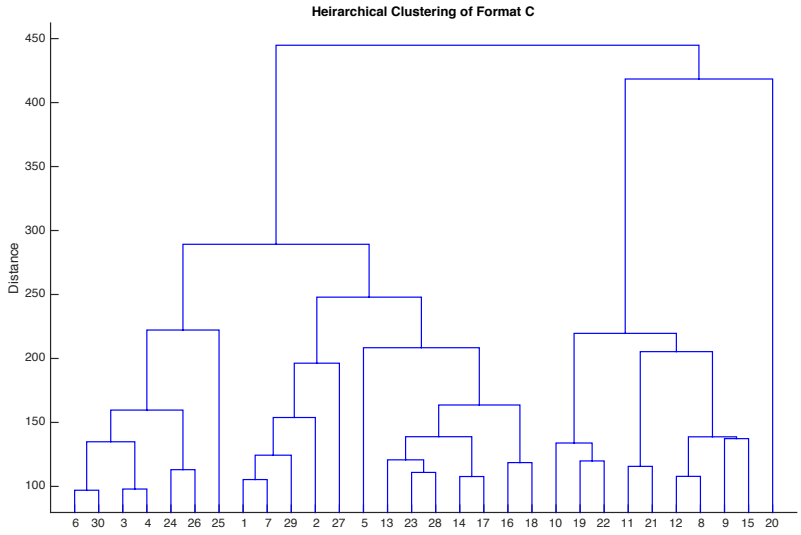


Figure 5.3. Hierarchical clustering of eye movement measures for Format C ($T \rightarrow Q$)

Table 5.4. Comparison of average eye movement measures for clusters obtained from hierarchical clustering of format C data ($T \rightarrow Q$)

	Cluster		
Measures	1	2	3
Cluster Size	41	72	21
Number of fixations	171	295	467
Average fixation duration (s)	0.17	0.24	0.24
Total fixation duration (s)	28.3	69.8	110.9
Regression ratio	35%	28%	32%
Average forward saccade length	123	96	110
Longest reading sequence	48.27	88.99	74.05
Read ratio	58%	78%	71%
Skim ratio	23%	11%	15%
Scan ratio	19%	10%	14%
Multiple Choice Score	0.85	0.74	0.67
Cloze Score	0.78	0.55	0.71
Combined score	1.63	1.28	1.38

Clusters 2 and 3 have similar comprehension scores to each other yet different observed eye movements. Cluster 2 has a higher numbers of fixations and longer total fixation times than observed for Cluster 1 but lower and shorter, respectively, than observed for Cluster 3. This cluster also has the most instances so can therefore be considered to represent the average eye movements for the group. Cluster 3 has

the highest numbers of fixations and the longest total fixation times. This cluster groups together the instances where the readers spent more time on the text.

The outlier was removed from the data set and the remaining data was split into the three clusters described in Table 5.4. Using random forest ensembles the question scores were predicted for within each cluster. The average and standard deviations of MCR are shown in Table 5.5 from 10-fold cross validation.

Table 5.5. Comparison of Misclassification (MCR) results for predicting questions scores for text only pages eye movement measures from Format C using Random Forest Ensemble Classification

Cluster	Combined
1	0.39±0.17
2	0.81±0.13
3	0.61±0.21

The hypothesis was that grouping together like eye movements would create more accurate predictions of the question scores. However, this was not validated.

5.5.3.2 Format D ($Q \rightarrow T \rightarrow Q$)

The results of the clustering are shown in Figure 5.4. The cluster analysis reveals that there are once again three clusters, one of which only contains 3 data instances. Comparing the averages of eye movement measures for these clusters are shown in Table 5.6. The cluster with only three data instances has considerably more observed fixations than the other clusters. Additionally, these instances have considerably lower comprehension scores than the other two clusters. We discount these data instances as outliers for the classification analysis.

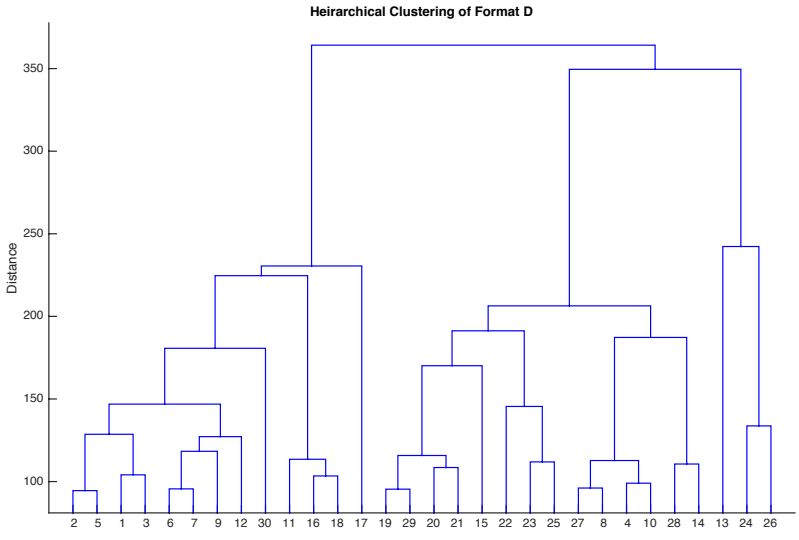


Figure 5.4. Hierarchical clustering for eye movement measures from format D ($Q \rightarrow T \rightarrow Q$)

For Format D we expect fewer fixations and less reading of the text given that participants had knowledge of the reading comprehension questions before reading

the text. As a result, the hypothesised behaviour was for the participant to search the text for the answers and only read the apparently relevant sections of the text. The remaining two clusters appear to fit with this assumption, whereby there is a smaller cluster where more reading of the text occurred. There is however no difference in the scores between Clusters 1 and 3 so the difference really is in the reading behaviour. Given the increase in fixations and reading ratio in Cluster 1 compared to Cluster 3, the behaviour grouped together in Cluster 1 seems to be that of more normal reading behaviour as the characteristics of the measures are similar to Cluster 2 for Format C. In these cases the readers must not have skimmed the text and rather read it in a "normal" manner. Cluster 3 however has dramatically lower numbers of fixations and reading ratio, which is more characteristic of the behaviour that was hypothesised.

Table 5.6. Cluster details for Format D ($Q \rightarrow T \rightarrow Q$)

	Cluster		
Measures	1	2	3
Cluster Size	45	3	87
Number of fixations	286	445	137
Average fixation duration (s)	0.26	0.18	0.17
Total fixation duration (s)	76	81	23
Regression ratio	29%	35%	40%
Average forward saccade length	98	121	143
Longest reading sequence	89	48	30
Read ratio	77%	58%	45%
Skim ratio	13%	22%	28%
Scan ratio	10%	20%	27%
MC Score	0.84	0.33	0.77
Cloze Score	0.60	0.33	0.67
Combined Score	1.44	0.67	1.44

As for the Format C data set, random forest ensembles were used to predict the question scores for Clusters 1 and 3. The average and standard deviations of MCR from 10-fold cross validation are shown in Table 5.7. These results show little improvement from the results obtained without clustering. This indicates that the hypothesis of grouping together similar eye movements is not an effective method for increasing prediction accuracy.

Table 5.7. Comparison of Misclassification (MCR) results for predicting questions scores for text only pages eye movement measures from Format C using Random Forest Ensemble Classification

Cluster	Combined
1	0.39±0.23
3	0.51±0.21

5.6 Discussion

The central question of this investigation is whether eye tracking data can be used to predict reading comprehension. We focus on the prediction of reading comprehension from eye movements recorded from reading text shown in different formats. This allows us to investigate if the presentation of the text and questions affects reading comprehension prediction. The analysis involved three components; the first was an investigation of reading comprehension prediction using ANNs. This was an extension of previous work where we investigated the benefits of using FOE as the performance function for training. The second component was a comparison to other machine learning technique, and finally clustering of two of the data sets, formats C and D.

Generally, the results reflect the fact that the data sets are quite hard to classify, especially the formats C and D data sets. The best classification results were obtained using three layers of hidden neurons, indicating complex relationships between the eye movement measures and reading comprehension scores. In saying this, the results from the analysis show that predictions for formats A and B are the most accurate. Whilst this could indicate that there are relationships between eye movements and the reading comprehension scores in these formats, it is also probably a by-product of the fact that formats A and B are highly imbalanced. Formats C and D on the other had have wider distributions of comprehension scores and are less imbalanced than formats A and B.

Note that the 10-fold cross validation is not done on a participant basis, but rather a data point basis. The implications of this are that the modelled data is for the current participant base and the results may be quite different if completely new participants are tested. This is an area that requires further work, where we test the model on new participants to check for versatility. In practice, we would expect that the trained model would be constantly updated to account for new students and for new scenarios.

The second part of the analysis was an investigation of other methods of classification. FOE-ANNs outperform the other classification techniques used when predicting the combined comprehension scores. The conclusion made from this part of the analysis is that ANNs are the most appropriate classification technique, from the set tested, for predicting reading comprehension scores. However, we note that the satisficing classification technique from the set of four is the random forest ensemble and further investigation of machine learning techniques should be carried out. In particular techniques that can be sued to deal with the imbalance in the data sets.

Martínez-Gómez and Aizawa (2014) found it difficult to predict comprehension of text where no significant result could be found in the regression task of prediction understanding scores. Further analysis could be run on the current data set to make predictions using regression analysis rather than classification analysis. Our work adds to this showing that to some extent eye movements can be used to prediction

reading comprehension scores. However, this still needs further investigation to improve accuracy.

One of the advantages of using FOE is that it is a flexible error function that can be tailored to data sets and problems. Specifying the shape of the FMF used to calculate FOE does this. However, there is no simple way of constructing an FMF. In this analysis we only investigated 7 predefined FMFs, however, a beneficial approach would be to determine the most appropriate FMF shape from the data set. An area of further exploration is how to apply the learning of the FMF shape when using other classifiers such as neural networks.

Notably, it was hypothesised that predictions from Format D would be better than Format C as participants were shown the questions before being shown the text. Participants therefore knew what to look for in the text in order to answer the questions. However, this was not found to be the case. In both formats, participants would have read the text to the point at which they deemed they understood the text. This is subjective and dependent on a number of factors including prior knowledge, familiarity with the subject matter, current state (mood, arousal, etc.) as well as their motivations. This could account for the variability in eye movement measures and the reading comprehension outcomes. Participants' subjective reading comprehension ratings were not recorded for each of the individual texts. Future work, will be in recording this information and exploring relationships between eye movements and subjective comprehension for each text.

The clustering of both formats C and D reveals interesting patterns of eye movements and reading behaviours that are evident in the data. In particular, there was no relationship between reading speed and reading comprehension but there is high variability in reading styles, which is consistent with Underwood et al. (1990). The clustering of both formats C and D demonstrates that there are clear changes in reading behaviour observed for participants. The clusters were composed of data samples from different participants, indicating that participants to some extent changed their reading behaviour to reflect the text. The change is from low reading behaviour (lower numbers of fixations and reading ratios, as well as shorter total fixation time), to medium reading behaviour, up to high reading behaviour, which is indicated from higher numbers of fixations and reading ratios, as well as longer total fixation times.

5.6.1 Implications for eLearning

The analysis shows that there are relationships between eye movements and reading comprehension, albeit complex relationships. These relationships are strongest when the questions are shown along with the text, which in the context of practical use for eLearning is not so significant. Nevertheless, this is an important finding as it shows that there are relationships.

The goal of reading comprehension detection is to incorporate eye tracking into eLearning environments and use the eye tracking data as a form of adaption. Consequently, the content and the presentation of content can be altered to reflect

the student's current state. The product of reading comprehension prediction is twofold; first, if students are given text to learn, instead of explicitly assessing their comprehension, eye tracking could be used to assess their understanding thus reducing time, workload, and potentially stress or anxiety of the students. Secondly, predicting students' comprehension using eye tracking would allow the learning environment to, 1) adapt the questions asked of students about the content, and, 2) alter the learning path to reflect the students' current understanding levels.

The second point can be elaborated upon, as this is the main advantage of predicting comprehension from eye gaze. Take for instance the case where a student has read some learning materials but does not understand it. He is then asked the same comprehension questions as all other students. Not understanding the text possibly increases the student's anxiety about the learning material, causing him to be disheartened. Two solutions arise from this, first is that the questions themselves are modified to be easier, perhaps covering more superficial understanding of the content. Text with more explanation of the content that was not understood could then be given, after which they are assessed on the original comprehension questions. Secondly, instead of asking comprehension questions at all, the text with more explanation could be provided to the student.

If we now consider the converse case where a student has a high level of understanding, as is the case when the student has prior knowledge on a certain topic, this student may become frustrated or bored by being presented with easy content and unchallenging questions. Again, either the questions or the content could be altered to present these students with hard subject matter and questions that require much more thought and insight.

Furthermore, the use of a technology such as eye tracking gives rise to the possibility of monitoring implicit behaviours related to reading and learning in eLearning environments. As seen, especially from the cluster analysis, there are differences in eye movement and reading behaviours within the formats. From previous work, we have shown that differences in eye movements and related measures can be used to measure how difficult or interesting a student is finding certain texts (Copeland & Gedeon, 2013b). This information can then be used by the instructor, or writer, of the learning materials to find the implicit difficulty of the questions and text, to get a ranking of how the students are performing, as well as any other information such as the rate at which students are developing (Copeland & Gedeon, 2013b).

The eye movement measures can also be used to determine if a student is having problems reading materials over a longer term so that remedial assistance can be provided. In the opposite case of a student who consistently skims text due to high levels of prior knowledge and understanding of a given topic, this student can be helped by being either moved up a level in the course or provided with more challenging tasks.

5.7 Conclusion

This chapter is an investigation of predicting reading comprehension from eye movements. The eye gaze data that was collected from the user study in Chapter 3 showed that presentation format affects eye movements and reading behaviour. In this chapter, we explored how those differences cause variations in prediction outcomes of reading comprehension from eye gaze. We found prediction of reading comprehension measures was most accurate for formats of presentation where the text and questions are shown together. For these formats, denoted A and B, we could achieve 85% and 88% correct classification, respectively, using FOE-ANN. Prediction from the formats where questions and text are shown separately to one another proved to be more challenging, where poor classification results were obtained. Further work is required to investigate other machine learning techniques, especially those that could be used to deal with imbalanced data better.

The extension of previous work of the use of FOE as a performance function for training ANNs has shown that FOE-ANN provides better prediction results than the use of MSE-ANN. However, further research needs be carried out to explore the nature of this performance function and the creation of the FMF shapes used to calculate FOE. Additional data sets and problems should be trialled as well to investigate if these results generalise.

Chapter 6

Effect of Text Difficulty on Prediction of Comprehension

"Prediction is very difficult,
especially if it's about the future."
— Niels Bohr, Danish Physicist

Prediction of reading comprehension scores is a difficult task, as we have already seen in Chapter 5. In this chapter we extend the work from the previous chapter by investigating the effect of text difficulty and machine learning techniques on prediction accuracy. To this end, a user study was carried out to collect data from L1 and L2 participants as they read texts with differing degrees of difficulty. The grades of overall difficulty are based on different levels of readability and conceptual difficulty. We hypothesised that text difficulty and reader type would affect prediction quality. We found that neither had a significant effect on the accuracy of the k-nearest neighbour (kNN) classifier used. However, we did improve the classification accuracy to on average 80% for the L1 group and 73% for the L2 group, which is a substantial improvement from the 44% correct classification obtained in the previous chapter for format C. These results were achieved by using genetic algorithms (GA) for feature selection, which were significantly higher than the results produced when no feature selection is performed. We found that readability affects normalised number of fixations (NNF) but not regression ratio. We also found that there is a significant difference between the L1 and L2 readers NNF and tendency to regress. Although the significant difference between the groups is what we would expect, from past research and the findings of this thesis, there was no interaction effect between the reader groups and the text difficulty. This indicates that in this study the readability and conceptual difficulty of the text affect the two groups similarly. This chapter builds on work presented at CogInfoCom 2015 (Copeland et al., 2015).

6.1 Introduction

The eye moves in specific patterns when reading, making it not only possible to detect when a person is reading from their eye gaze (Campbell & Maglio, 2001) but also how they are reading (Buscher et al., 2008), what task they are performing (Iqbal & Bailey, 2004), how relevant they find text (Buscher et al., 2012; Vo et al., 2010), and even their cognitive load (Iqbal, Zheng, & Bailey, 2004). Unsurprisingly, given this list, eye movements also provide insight into how difficult text is to read (Rayner et al., 2006) and comprehend (Martínez-Gómez & Aizawa, 2014; Underwood et al., 1990). As text increases in difficultly, the number of fixations increases, fixation duration increases, saccade size decreases, and regressions increase (Rayner, 2009; Rayner et al., 2006; Staub & Rayner, 2007). Text characteristics have also been shown to affect comprehension for which, in the context of legal documents, making text simpler would benefit vulnerable populations (Scherr et al., 2015). This can be extended to considering the differences of students in eLearning, where some students may be supported by simpler texts.

Eye movement measures have been shown to be effective at distinguishing between readers with low and high level of understanding as well as predicting English language skill (Martínez-Gómez & Aizawa, 2014). Eye gaze has also been used to investigate those parts of text that readers are failing to comprehend, indicating that eye gaze features, such as the number and duration of fixations, can be used to identify reading incomprehension (Okoso et al., 2015). Prediction of reading comprehension derived from eye movements would make the current model of adaptive eLearning more versatile. It would allow for the eLearning environment to change dynamically based upon implicit behaviour. This would result in decreased time for the student to learn the material as well as not contributing to their over- or under-confidence in actual understanding of the learning materials.

However, the task of predicting reading comprehension from eye gaze is not a simple one, as we have established in the previous chapter. Clearly, the current method of prediction is inadequate so we explore methods of increasing prediction quality. It has been shown that the number of fixations increases as text difficulty increases and the number of regressions increases when inconsistencies are introduced into texts (Rayner et al., 2006). Indeed, there are many factors in texts that affect readers' eye movements. With this in mind we postulate that the differences induced by text with differing degrees of difficulty will i cause significant differences in eye movements and therefore will have differential effects on predictions of reading comprehension scores. As in the previous chapter the central research question is:

Can eye-tracking data be used to predict reading comprehension scores in eLearning environments for L1 and L2 readers?

Once again we also investigate factors that could affect prediction accuracy. In this chapter we investigate the effect of text difficulty on prediction accuracy of comprehension:

Does text difficulty affect predictions of comprehension?

We explore this question by investigating factors that influence prediction performance of reading comprehension scores from eye tracking data. These factors are text readability, conceptual difficulty of the text, and whether the reader is a first (L1) or second (L2) English language reader. To perform this investigation, we conducted a user study to collect eye gaze data from participants as they read text with differing degrees of difficulty. We hypothesise that predictive performance is affected by text difficulty and reader type, in particular that, 1) more accurate predictions will be obtained for L1 readers compared to L2 readers, and 2) more accurate predictions will be made when the text is most difficult. We explore two additional methods of increasing predictive accuracy; the first being the inclusion of pupil dilation data into the feature set, and the second being another way of generating the feature set. This involves breaking the task into smaller windows are generating the eye movement measures for these windows rather than the whole task. This technique, in combination with feature selection using genetic algorithms, has been used to improve stress prediction (Sharma & Gedeon, 2013b) as well as for biofeedback (Gedeon, Zhu, Copeland, & Sharma, 2015). Our hypothesis is that breaking the task up into smaller windows and calculating measures for each window will improve predictive accuracy.

This chapter is organized into the following sections: background information; user study method; results and analysis; discussion and implications; and conclusion and further work.

6.2 Feature selection using genetic algorithms

The majority of the background material for this chapter has been covered in the literature review and in the previous chapter. We introduce a new technique in this chapter: the use of feature selection, specifically the use of genetic algorithms to perform feature selection. Feature selection is the selection of a subset of features is selected before modelling occurs. This may be done for several reasons, such as when data sets contain hundreds or thousands, or indeed hundreds of thousands of features, however it may be that many of these features hinder the model's accuracy because they outweigh the useful features (Guyon & Elisseeff, 2003). Removing redundant or irrelevant features and only using the most "useful" features can improve the quality of the model created as well as speeding model creation (Guyon & Elisseeff, 2003; Siedlecki & Sklansky, 1989). An example of this can be seen when random forests are used to model high dimensional data. Poor results are often found since random sampling of the feature set to create the ensemble often results in subsets of only irrelevant features (Amaratunga et al., 2008).

One method of feature selection that has been shown to be effective is genetic algorithms (GA) (Garrett, Peterson, Anderson, & Thaut, 2003; Yang & Honavar, 1998). Before proceeding to discuss the use of GAs to perform feature selection we will introduce the concept of GAs. GAs are search algorithms that are inspired by natural evolution (Whitley, 1994). In particular, are based on the fundamentals of genetic evolution to search the solution space. GAs are often considered a "global"

search tool because they do not usually suffer from the disadvantages of optimisation methods such a gradient decent, e.g. getting stuck in a local minima, however, GAs should be thought of as a search process rather than an optimisation process (De Jong, 1993).

The foundation of a GA is that there is a population of individuals used to search the solution space. Each individual in the population represents a potential solution to the problem, which is represented as a chromosome. A chromosome represents the characteristics of an individual, which refer to the variables of the search problem. The chromosome is composed of a set of genes whose indices are termed loci. Each gene can have one or more values, which are termed alleles. An important step in the design of a GA is to find an appropriate representation of the chromosomes. The genotypes are often represented as simple data types such as a bit string or numerical representation of a chromosome.

The main driving operators for GAs are selection and recombination, through application of a crossover operator, with mutations to add diversity. The focus of genetic algorithms is generally on recombination of existing chromosomes in the population so mutation rates are usually set to less than a 1 probability (Whitley, 2001). The population is usually initialised to a random set of chromosomes and then crossover functions are used to create the next generation of chromosomes. Types of crossover functions include random selection, proportional selection, tournament, and elitism; however there are many other types.

GAs have been used successfully for feature selection for neural network classification of a number of University of California Irvine (UCI) machine learning repository data sets[16] (Yang & Honavar, 1998). These data sets mostly have feature sets below 20, with a maximum range of 60. This is quite a low number of features when physiological data is considered such as pupil dilation, EEG, ECG, GSR, amongst other signals. Selecting features from larger feature sets and features derived from physiological data have also been successful (Garrett et al., 2003; Schroder et al., 2003). In particular, feature selection has been shown to be beneficial in predicting stress during reading tasks from physiological signals, including eye gaze data (Sharma & Gedeon, 2012, 2013a).

6.3 Method

6.3.1 Participants

The eye gaze of 70 participants (47 male, 23 female) was recorded. Participants had an average age of 25 years (9 years standard deviation, range of 18 to 60 years). Of the participants 46 stated that English was the first language they learnt to read in and the remaining 24 stated a language other than English. Participants were mostly (n=44) sourced from a first year Web design and development course offered at the university (ANU). All other participants were sourced from the university more widely.

[16] Available at http://archive.ics.uci.edu/ml/ Last accessed: 29th January 2016

6.3.2 Design

Participants' eye gaze was tracked as they read and completed a tutorial on the topic of "*Digital Images*". The tutorial was taken from a first year computer science course on web design and development offered at the ANU. The tutorial was composed of 9 texts of approximately 240 words (standard deviation of 20 words) in length. An example of the tutorial text is shown in Figure 6.4. Since participants were mostly sourced from the course, they were provided with an incentive to do well on the tutorial to gain marks for the course. Additionally, the task was similar to tasks the majority of participants were used to performing throughout the course, so participants were not given practice texts to read. Reading was self-paced; participants were told to read the text and that after reading the text they would be asked comprehension questions.

	Readability		
Concept	**Easy**	**Medium**	**Difficult**
Easy	A	B	C
Medium	D	E	F
Difficult	G	H	J

Figure 6.1. Description of the text property breakdown

Two variables in the text were altered in each new text; the readability and concept difficulty. Each variable has 3 values; "easy", "medium", and "difficult", giving a total of 9 combinations (see Figure 6.1). The readability was measured using the Flesch-Kincaid Grade Level (Kincaid et al., 1975). Each increase in readability level is about 3 years of education, so we start at a level that all participants should comfortably be able to read and finally move to a grade that signifies postgraduate level studies (see Figure 6.2). The COH-Metrix L2 Readability index is designed to rate the readability of text for L2 readers. The L2 readability index for each text was generated using COH-Metrix 3.0 (McNamara et al., 2013). Since, it has been shown that the L2 readability index is more appropriate for describing the readability of texts for L2 readers (Crossley et al., 2008), we check that the L2 readability indices are consistent with the Flesch-Kincaid grade levels, in that there is a consistent increase in L2 readability indices as there is a decrease in Flesch-Kincaid grade levels. More specifically, the increase in difficulty in readability is consistent for L1 and L2 readers.

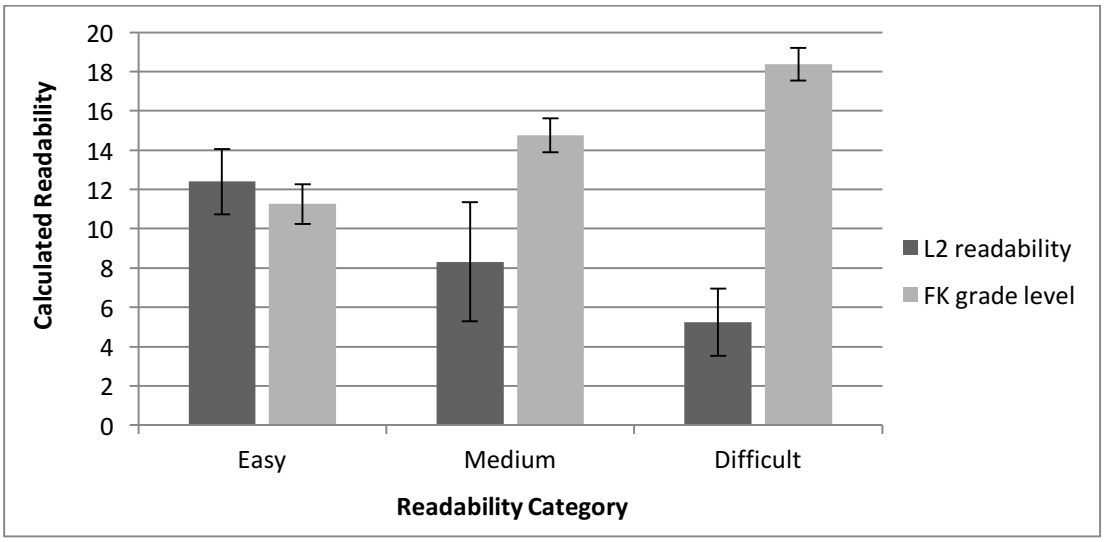


Figure 6.2. The Flesch-Kincaid readability grade level and COH-Metrix L2 readability for each level of readability.

Table 6.1 Example of *chunking* concepts to derive the levels of concept difficulty for Topic 3 - Photo Credibility

Chunk	Conceptual level		
	Basic	Intermediate	Advanced
1	Photography was invented in 1850s, photos have been modified since then	That altering is actually tampering. That deliberately false photos are created	That tampering is done to deceive. That photos were publishing their deceptive images into the public record.
2	Manipulating images used to be difficult, now it is easy. People do it to critique others or make money	That image manipulation is called post-processing. That these images can easily be distributed given our networked society. That they go around the world quickly	That these manipulated images are now part of our knowledge base and daily experience
3	There is something called digital forensics that run some tests on photos, but there is no authenticate process	That image manipulation causes manipulation artefacts.	It is not possible to positively authenticate an image
4	Forensics are important because manipulated photos can affect people and industries	Forensics is a quickly growing field because of these impacts	Despite forensics, digital photographs cannot be guaranteed to be real

An expert and educator in digital images wrote texts as teaching material for web design course the participants were sourced from. Since conceptual difficulty in a discipline is largely a qualitative judgment often best measured by a subject matter

expert, it is difficult to measure with automated tools. Whilst the concept level was defined by the expert's judgments, she based her concept content design on the idea of conceptual chunking (Miller, 1956). Table 6.1 provides an example of how chunking was used to create the different levels of conceptual difficulty. Additionally, the text difficulty was designed using the following schema and principles:

i. systematic chunking, including increased number of concepts presented in each level of expository text, refer to Table 6.1 for an example of this;
ii. consideration of information scaffolding (what participants could be expected to already know); and
iii. expanded knowledge demands presented in each level of expository text (Initiative, 2010).

The 9 texts were shown in groups of 3; each set of 3 texts covered a topic. The 3 topics are "Working with Digital Images", "Copyright and Intellectual Property", and "Photo Credibility". Each text had differing degrees of difficulty, whereby the readability and the concept difficulty was changed. The descriptions of each text's properties are shown in Figure 6.1. Each text is given an alphabetic label. Each participant was given a sequence of texts to read. These sequences are described as paths and that the first text is always A, the second text is always B, E or D (one square in the grid away from A), and finally the last text was one of B, C, D, F, G, H, or J. The paths are as follows: A>E>J; A>E>H; A>E>F; A>B>D; A>B>G; A>D>C; and, A>D>B, and are graphically shown in Figure 6.3.

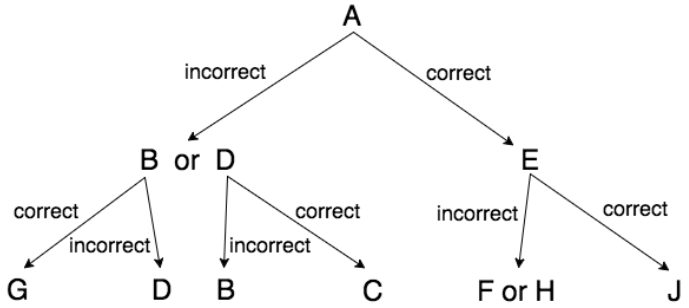


Figure 6.3 Process used to generate the paths

The process used to generate the paths was that of a hypothetical eLearning environment that changed the next text based upon answer correctness from the corresponding comprehension questions. Text A is always the start of the path. In the situation where the student answers the comprehension questions correctly for text A, then the system would present text E to the student. If the student then answered the comprehension questions for text E correctly then the system would be presented with text J. If the student answered the comprehension questions for text E incorrectly, then the system would present with another text, either F or H, each one step away from E. If the student answered the comprehension questions for A wrong, then texts B or D are shown, each one step away from A. If the system presents text B and the student answers are correct then the system presents text D, which is an increase in conceptual difficulty but same level of readability. If the student answers incorrectly, then the system presents text G, which is a decrease in

conceptual difficulty but increase in readability to check for whether increase in readability will influence the student's ability to comprehend the text. If the system presents text D, the opposite texts are shown to B.

The reasons these paths were chosen were to 1) start all participants on common ground, 2) only have subtle increases in text difficulty so that it would not be obvious what the text difficulty is, and 3) not have all increases as linear increases in difficulty, to elicit if one of the text properties has an influence over the subjective ratings.

Participants' prior knowledge on the subject area was not tested however participants were asked to rate how familiar they were with each topic after they read the text. The participants' ratings of familiarity to each topic are shown in Table 6.2. The ratings indicate that across the topics there are consistent percentages of all familiarity evaluations, with the largest percentage is that about 50% of participants have somewhat familiarity to all three topics.

Table 6.2. Participants' ratings of familiarity to each topic.

Familiarity evaluation	Topic 1	Topic 2	Topic 3
Familiar	19%	19%	17%
Somewhat Familiar	51%	52%	49%
Not Familiar	31%	27%	32%

After each text was read, participants were asked two comprehension questions to assess their understanding of the text. This is analogous to format C in the first user study (see Chapter 3). However, after being asked the two comprehension questions participants were then asked four qualitative questions related to the text they read:

1. How well do you think you understood the text?
   (Very well / Well / Somewhat / Not at all)
2. How confident were you answering the questions?
   (Very confident / Confident / Not Confident)
3. How difficult did you find the text to read?
   (Easy / Moderate / Hard)
4. How complex was the concept being explained in the text?
   (Basic / Intermediate / Advanced)

6.3.3 Experiment Setup

The texts and questions were implemented in the online learning environment used at ANU, called Wattle (a Moodle variant). A Moodle quiz module was used to implement the process. The text was presented to the participants as shown in Figure 6.4. A copy of the texts used for the experiment, along with the participant information sheet, consent form, and other experiment resources are found in Appendix B. All participants had knowledge of the learning environment and had

used it prior to the experiment. The study was displayed on a 1280x1024 pixel Dell monitor and the set up was identical to the set up used in Chapter 3.

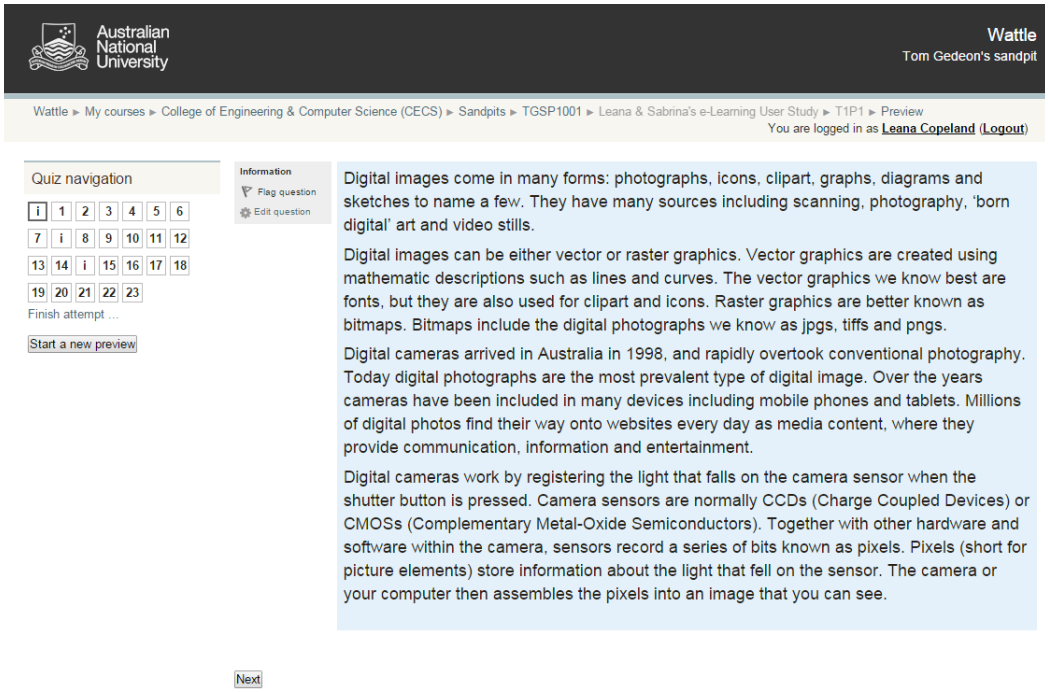


Figure 6.4. Example of text presented in the Wattle online eLearning environment

Eye gaze data was recorded at 60Hz using Seeing Machines FaceLAB 5 infrared cameras mounted at the base of the monitor. This eye tracker has a gaze direction accuracy of 0.5-1° rotational error and measures pupil diameter as well as blink events. The study involved a 9-point calibration prior to data collection for each participant. As the data recorded is a series of gaze points, EyeWorks Analyze was used to pre-process the data to give fixation points. The parameters used for this were a minimum duration of 60 milliseconds and a threshold of 5 pixels.

6.3.4 Data Pre-processing

The raw eye gaze data consists of x,y-coordinates of where the participants' eyes were looking. Fixation and saccade identification were performed on the eye gaze data. From this data many other eye movement measures are derived. Given that there are 70 participants and 9 texts there is a total of 630 eye gaze sets for the prediction analysis. Due to problems in collected data, 12 of these eye gaze sets had to be removed resulting in 618 eye gaze data sets for the prediction analysis. For each piece of text, eye movement measures and pupil dilation measures are calculated.

6.3.4.1 Inputs

Many of the eye movement measures have already been discussed; however, we introduce the use of pupil dilation data in this chapter. The measures used in this investigation have in the most part already been explained and so will not be elaborated upon here (see Chapter 5 for more details about eye movement

measures). The pupil dilation measures are calculated for the left and right eyes separately and then an average is calculated for the two eyes. These measures are:

- Average, minimum, maximum and range of pupil diameter
- Average, minimum, maximum and range of pupil area

The average pupil diameter for a fixation is calculated with the average standard deviation. Pupil dilation has been known to increase with increased cognitive load (Kahneman et al., 1969). Additionally, changes in pupil dilation have been found to reflect learning (Sibley et al., 2011).

A list of the 28 measures used is as follows:

1. Normalised number of fixations
2. Maximum fixation duration
3. Average Fixation duration
4. Normalised total fixation duration
5. Number of regressions
6. Regression ratio
7. Average saccade length
8. Reading ratio
9. Skimming ratio
10. Scanning ratio
11. Left eye - Average pupil diameter
12. Left eye - Average pupil area
13. Left eye - Minimum diameter
14. Left eye - Maximum diameter
15. Left eye - Range of diameter
16. Right eye - Average pupil diameter
17. Right eye - Average pupil area
18. Right eye - Minimum diameter
19. Right eye - Maximum diameter
20. Right eye - Range of diameter
21. Both eyes - Average pupil diameter
22. Both eyes - Average pupil area
23. Both eyes - Minimum diameter
24. Both eyes - Maximum diameter
25. Both eyes - Range of diameter
26. Both eyes - Minimum area
27. Both eyes - Maximum area
28. Both eyes - Range of area

We consider two cases in this investigation; the first is the same as in the previous chapter where the measures are calculated for the entire task and the second is *windowing* the task by dividing the task into quarters and calculating the measures for each quarter, sixth and eighth. In this way, there are a total of 112 features for each piece of text when 4 windows are used, 168 when 6 windows are used, and 224 when 8 windows are used.

6.3.4.2 Outputs

Reading comprehension score: The outcome variables are in the form of the participants' reading comprehension scores. After each piece of text the participant was asked two comprehension questions. Possible scores are 0, 1 or 2 for which the distribution of scores in the sets are described in Table 6.3.

As can be seen there is an imbalance of scores across the texts. The majority class is usually 2, however as the texts become more difficult the majority class shifts towards 1. Furthermore, the imbalance is more prominent for the L1 data set. Note that the differences in the comprehension scores will be investigated further in the next chapter.

Table 6.3. Distribution (%) of comprehension scores for each text and for the L1 and L2 data sets

	L1			L2		
Text ID	0	1	2	0	1	2
A	4	28	69	6	38	57
B	10	41	48	12	53	35
C	5	29	67	20	50	30
D	7	51	42	14	34	51
E	9	27	64	11	57	32
F	27	33	40	9	36	55
G	15	50	35	22	33	44
H	18	45	36	43	43	14
J	22	39	39	50	40	10

6.3.5 Feature selection method

Genetic algorithms (GA) are used to perform feature selection. The chromosome length is the total number of features in the set, i.e. 112 when 4 windows are used with the 28 measures outlined in Section 6.3.4.1. A chromosome is represented as a binary string as implemented by Oluleye et al. (2014). A bit represents each feature on the chromosome, and the bit value indicates whether the feature was used. The initial population was constructed using the algorithm specified by Oluleye et al. (2014). Finally, the fitness function used was the quality of the prediction from the predictors when that feature set was used. Further details regarding the GA parameters are outlined in Table 6.4.

Table 6.4. GA parameter settings for feature selection

GA Parameter	Value/Setting
Population type	Bitstrings
Population size	50
Generations	25
Crossover rate	0.8
Crossover	Arithmetic Crossover
Mutation	Uniform Mutation
Mutation Probability	0.1
Selection	Tournament of size 2
Elite count	2

Note that the number of generations and population size is similar to that reported by Yang and Honavar (1998). Furthermore, preliminary analysis showed that this number of generations provided improvement in prediction accuracy.

6.4 Results

The data collected from the user study allows us to investigate if text complexity and reader type affects predictions of reading comprehension. Recall that our

hypotheses are that, 1) more accurate predictions will be obtained for L1 readers compared to L2 readers, and 2) more accurate predictions will be made when the text is most difficult. In this chapter we investigate two additional methods of increasing predictive power; the first being the inclusion of pupil dilation data to the feature set, and the second being the use of windowing of tasks to increase the feature set and then use of feature selection. Our hypothesis is that breaking the task up into smaller windows and calculating measures for each window will improve predictive accuracy. The first part of this section is the use of no feature selection and no windowing, which is analogous to the previous chapter. We can therefore delve into the factors of text difficulty and reader type. The second part of the analysis is looking at using windowing of the task and GA feature selection.

The classification techniques used in the following analyses are ANNs, KNNs, and random forests. All ANNs used have a 2-hidden layer topology with 10 neurons in the first layer and 5 in the second, using MSE as the performance function. The kNN classifiers use *k* with square root the number of data instances in the set. All analyses were performed using Matlab R2013a.

6.4.1 Prediction without windowing

The first part of the analysis uses eye gaze features that are calculated from the entire task. The classification rates (%) from the random forest ensembles, ANN, and kNN are shown in Table 6.5 respectively. The results were generated from 10-fold cross validation.

Table 6.5. Classification rates (%) from no windowing or feature selection

Text	Text Properties		L1			L2		
ID	Read.	Concept	ANN	kNN	RF	ANN	kNN	RF
A	Easy	Basic	66	58	63	47	46	61
B	Mod.	Basic	41	34	50	39	29	56
C	Diff.	Basic	30	40	30	33	67	44
D	Easy	Int.	47	32	48	35	28	32
E	Mod.	Int.	55	54	53	43	42	45
F	Diff.	Int.	28	57	47	14	43	43
G	Easy	Adv.	50	35	53	50	30	60
H	Mod.	Adv.	30	45	30	40	45	75
J	Diff.	Adv.	30	50	55	20	40	40
Average			*41*	*46*	*48*	*36*	*41*	*51*

The results from this analysis are poor, analogous to what we found in the previous chapter for format C. The results are similar, or worse than, majority class prediction. The best classification comes from the random forest classifier where for the L1 group an average of 48% correct classification was achieved and 51% for the L2 group.

Furthermore, there is no obvious pattern caused by text difficulty on the predictions. Additionally, there is no obvious difference in prediction accuracy for the L1 and L2 data sets. If anything, contrary to our hypothesis, the L2 data set leads to greater accuracy of predictions.

6.4.1.1 Prediction with feature selection

Genetic algorithms are used to perform feature selection. Random forests, ANNs and kNN are then used to predict the reading comprehension score. The correct classification results (%) from GA-RF, GA-ANN and GA-kNN are shown in Table 6.6. Note that feature selection is performed with nest 10-fold cross validation, that is, the cross validation is performed within the GA.

Table 6.6. Classification rates (%) using feature selection and no windowing

Text	Text Properties		L1			L2		
ID	Read.	Concept	ANN	kNN	RF	ANN	kNN	RF
A	Easy	Basic	65	71	68	58	65	67
B	Mod.	Basic	69	70	71	66	65	68
C	Diff.	Basic	81	83	82	50	80	50
D	Easy	Int.	75	85	83	57	66	73
E	Mod.	Int.	75	77	78	65	75	77
F	Diff.	Int.	80	95	80	60	57	50
G	Easy	Adv.	65	80	75	80	80	90
H	Mod.	Adv.	77	77	83	80	90	70
J	Diff.	Adv.	67	85	80	70	90	60
Average			*72*	*80*	*78*	*65*	*73*	*67*

The use of feature selection significantly increases the prediction quality for all three classifiers. This time the best results are obtained when using the kNN classifier which are significantly higher than when no feature selection is used ($t(17)=11.176$, $p<0.0005$). We are now able to achieve classification results above majority class prediction, especially for the L2 data set and the more difficulty tests which do not have as severe imbalance as compared to the simpler texts.

To assess if the text difficulty, or the reader group, have significant effects on the prediction accuracies we use ANOVA. The kNN accuracy rates are normally distributed (using the Shapiro-Wilks normality test, p=0.923). The dependent measure as the kNN accuracy and the L1/L2 reader groups and the readability and conceptual difficulty as the independent variables. Neither text readability ($F(2,4)=1.68$; $p=0.296$), nor conceptual difficulty ($F(2,4)=3.82$; $p=0.118$), nor reader group ($F(1,4)=3.18$; $p=0.149$) have an effect on the kNN accuracy. This implies that there is no relationship between the degree of accuracy from the kNN classifier with either text difficulty or reader group. There is no statically significant effect of interaction between any of the three independent variables.

Our hypothesis regarding text difficulty is that more accurate predictions would be achieved when the text is most difficult; perhaps because more cognitive

resources are committed to the task hence fewer distractions are possible. However, this hypothesis is not validated by the results gained above.

There is a significant correlation between conceptual difficulty and the kNN accuracy (r=0.5, p=0.049). Whilst the correlation is not large there does appear to be a small effect of conceptual difficulty on kNN accuracy.

6.4.2 Prediction with windowing

The next part of the prediction analysis is using eye gaze features that are calculated from windowing the task. By this we mean that we divided the entire reading take up into smaller segments to calculate the eye movement measures. The effect of the number of windows on prediction accuracy is investigated in this section. In this part we conflate the datasets without controlling for text complexity, i.e. "all texts". Feature selection is performed with nest 10-fold cross validation, that is, the cross validation is performed within the GA. The number of windows used to divide the data set into smaller segments are 2, 3, 4, and 6. The results are generated using kNN classifications as shown in Table 6.7.

Table 6.7. Correct classification (%) of reading comprehension for different windows

Text	Text Properties		L1				L2			
ID	Read.	Concept	2	3	4	6	2	3	4	6
A	Easy	Basic	75	73	72	69	67	67	66	66
B	Mod.	Basic	85	66	73	71	76	59	69	62
C	Diff.	Basic	80	75	75	70	70	70	67	67
D	Easy	Int.	71	68	66	66	70	62	62	67
E	Mod.	Int.	79	71	73	75	69	66	75	77
F	Diff.	Int.	77	88	87	82	73	75	83	77
G	Easy	Adv.	83	87	82	78	78	75	72	75
H	Mod.	Adv.	90	70	70	80	80	82	80	78
J	Diff.	Adv.	85	85	75	80	77	65	65	73
Average			*81*	*76*	*75*	*75*	*73*	*69*	*71*	*71*

To assess if the window size or the reader group, have significant effects on the accuracies we use ANOVA again. The kNN accuracy rates are normally distributed (using the Shapiro-Wilks normality test, p=0.333). The dependent measure as the kNN accuracy and the L1/L2 reader groups and window size as the independent variables. The window size does not have a significant effect on the kNN results (F(3,64)=1.88; p=0.141; partial η^2=0.081), but the reader group does have an effect (F(1,64)=11.87; p=0.001; partial η^2=0.156). It is interesting that the L2 participants have significantly lower prediction accuracies than the L1 participants in this case given that there was no significant difference in the previous test, which requires further analysis. Additionally, there is no statically significant effect of interaction between any of the three independent variables. This implies that the window size does not actually affect the accuracy of the result from the kNN classifier, showing that there is no benefit by windowing the task. When 2 windows are used the

performance remains the same as when no window is used. However, after this the classification accuracy starts to decline; adding more windows actually impedes classification performance.

6.4.3 Effect of text difficulty on eye movements

Given that the results are not as we had hypothesised, we now delve into looking at the eye movements themselves and whether they differ significant due to the text difficulty. In this section we investigate if the grades of readability and / or conceptual difficulty affect eye movements. The assumption, based on past research, is that there will be significant differences in eye movements between L1 and L2 readers as well as between the different grades of text difficulty. To address these assumptions, we use MANOVA analysis determine if there are any statistical differences between text properties and reader type. The two eye movements that are analysed in this section are the normalised number of fixations (NNF) and the regression ratios as both are known to be affected by text difficulty (Rayner et al., 2006). The NNF and regression ratios are shown in Figure 6.4 and Figure 6.5, respectively.

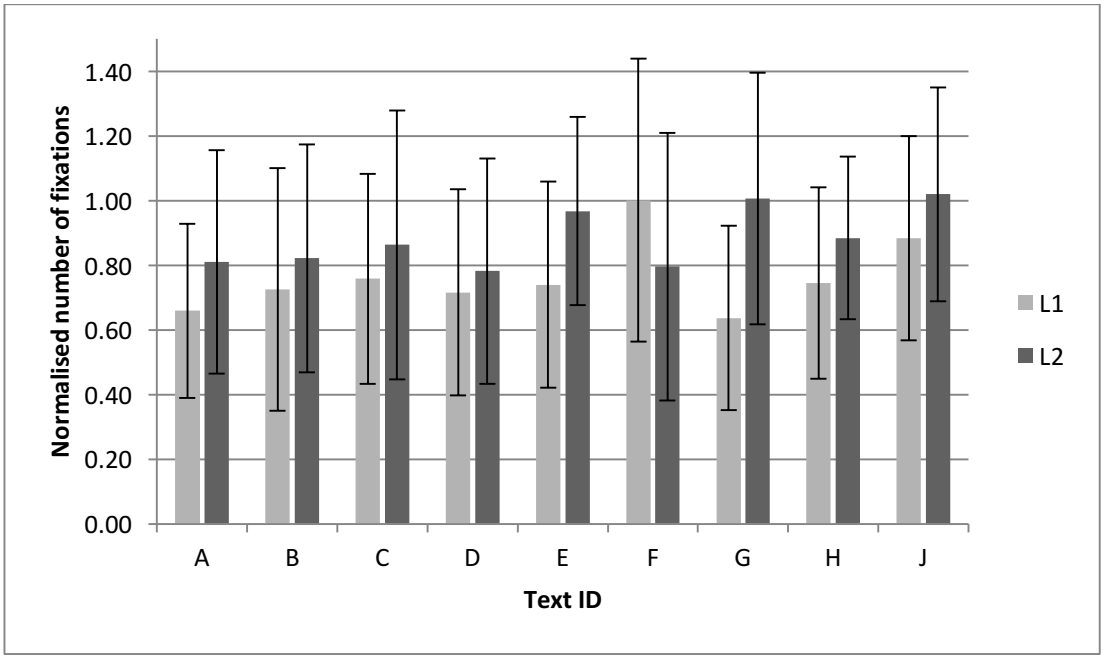


Figure 6.5. Normalised number of fixations (NNF) for each text

The correlation between the NNFs and regression is within the acceptable limits for MANOVA. The Levene's test for equality of variances shows that there is homogeneity for both dependent variables ($p>0.05$). Additionally, the Box's M value of 73.68 ($p=0.04 > 0.005$) is interpreted as non-significant so we can be satisfied that we have homogeneity in the variance-variance-covariance matrices.

There is a significant difference in eye movement measures between L1 and L2 readers ($F(2,592)=6.017$, $p=0.003$; Wilk's $=0.980$, partial $\eta^2=0.020$). Readability affected eye movements ($F(2,593) = 4.074$, $p=0.017$; Roy's $\lambda=0.014$, partial $\eta^2=0.014$), however, conceptual difficulty did not affect eye movements ($F(2,593)=2.299$, $p=0.101$; Roy's

=0.008, partial η^2=0.008). There is no significant effect of interaction between conceptual difficulty, readability, and reader type.

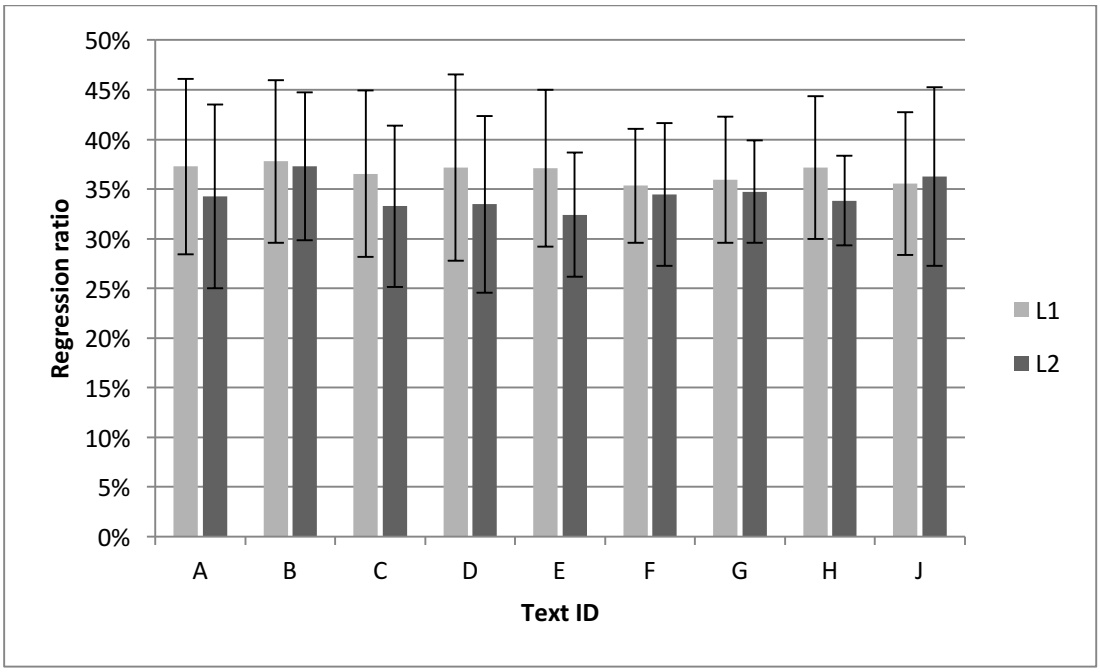


Figure 6.6. Regression ratios for each text

Between-subjects ANOVAs are used to determine how the eye movements differ for each text type as well as between L1 and L2 readers. L1 readers have lower NNFs (F(1,593)=10.972; p=0.001; partial η^2=0.018) and higher regression ratios compared to L2 readers (F(1,593)=5.668; p=0.018; partial η^2=0.009). This result confirms the observations made from inspections of Figure 6.5 and Figure 6.6, that L2 readers have higher NNF values and lower regression ratios.

The ANOVA reveals that the readability only affects the NNFs observed (F(2,593)=3.45; p=0.032; partial η^2=0.012) but not the regression ratio (F(2,593)=0.181; p=0.835; partial η^2=0.001). Tukey's multiple comparison tests were used to further investigate the effect that readability has on NNFs. This reveals that the difference lies in the easy versus the difficult readability ($p<0.005$) however the medium readability results in NNFs that are not statistically different from the easy or difficult text.

Note that the NNF is a ratio of fixations to words. There is generally an uneven distribution of fixations on words whilst reading English (Rayner, 1998). The NNF values for normal reading behaviour are therefore expected to be less than 1. In fact, Carpenter and Just (1983) found that readers fixate on average on 67.8% of words. The closer a value is to 0 the greater the skimming of the text. Values above 1 correspond to more fixations than there are words in the paragraph and are indicative of re-reading of some of the text. The NNF values for each text are shown in Figure 6.5.

From the MANOVA analysis we observe that there is a significant difference between L1 and L2 readers. We can observe from Figure 6.5 that the difference

comes from the L2 readers tending to have higher NNFs compared to the L1 readers. This is expected from past research (Dednam et al., 2014; Kang, 2014). The MANOVA analysis also shows that text readability has a significant effect on NNFs. We can see from Figure 6.4 that as readability difficulty increases, so too does the NNFs. This is what would be expected from past research (Rayner, 1998; Rayner et al., 2006). As we would expect from the MANOVA, there is no similar increase in NNF values as the concept level increases.

Additionally, the MANOVA shows that there is a significant difference between the regression ratios for L1 and L2 readers, averages shown in Figure 6.6. Contrary to what we would expect, the L2 readers have lower regression ratios compared to the L1 readers. That is, the L1 reader tend to regress more in comparison to forward saccades. This could be due to the fact that they simply have fewer forward saccades so this need to be further investigated. As we would expect from the MANOVA there is no relationship between regression ratio and text difficulty.

The analysis shows that there is significant effect of readability on NNF values. However, conceptual difficulty had not effect on either eye movement measure, contrary to our hypothesis. We found that readability affects normalised number of fixations (NNF) but not regression ratio. We also found that there is a significant difference between the L1 and L2 readers NNF and tendency to regress. However, there was no interaction effect between the reader groups and the text difficulty. This indicates that in this study the readability and conceptual difficulty of the text affect the two groups similarly.

6.5 Discussion and Implications

The overall research question for this chapter, and the previous, was whether reading comprehension can be reliably predicted from eye tracking data. In the previous chapter we established that predicting reading comprehension scores from eye movements is not trivial. We explore the question of whether text difficulty affects prediction accuracy. The premise is that text difficulty causes differences in eye gaze (Rayner et al., 2006). Therefore, we hypothesise that increased text difficulty will increase prediction accuracy from eye tracking data. Furthermore, these investigations were performed with respect to L1 and L2 readers. We hypothesized that predictive performance would be different for L1 and L2 readers.

The analysis shows that there are differences in prediction accuracy between L1 and L2 groups. On average, prediction accuracies for the L2 groups are lower than for the L1 group. However, text difficulty was not found to have a significant effect on prediction accuracy. This requires further analysis since, even though there is no statistically significant difference, there is much less of an imbalance in the scores in the more difficult texts. So obtaining similar prediction results to cases with an imbalance in scores indicates that the prediction quality must be improved somehow.

Even though our hypotheses were not validated in this analysis we did improve classification accuracy to on average 80% for the L1 group and 73% for the L2

group, which is a substantial improvement from the 44% correct classification obtained in the previous chapter for format C. These results were achieved by using genetic algorithms (GA) for feature selection, which were significantly higher than the results produced when no feature selection is performed.

The analysis of the eye movements for each text, somewhat, supported this conclusion. We had hypothesised that we would observe a much greater difference in eye movement caused by the difficulty of the text. However, what we see is that there is a significant effect of readability on NNF but not on regression ratio, and the conceptual difficulty does not affect either measure. Further investigation should be carried out to investigate this further, including looking further into the pupil dilation data and whether the pupil is affected more by conceptual difficulty.

We added pupil dilation measure to the feature set as pupil dilation is affected by cognitive load (Beatty, 1982; Iqbal et al., 2004; Kahneman & Beatty, 1966), whereby the pupil dilates under increased load and constricts under decrease load. Furthermore, it has been shown that pupil dilation is also affected by repeated exposure to a task, or more precisely, learning (Kahneman & Beatty, 1966; Sibley et al., 2011). This would appear to be an appropriate candidate measure for reading comprehension prediction. However, average pupil dilation over a task can negate any of the observed changes in pupil dilation caused by the task (Iqbal et al., 2004). This introduces the idea that windowing the task into smaller chunks would improve predictive accuracy. The windowing results in high numbers of features so we also introduced feature selection. Windowing physiological data and then using feature selection has been shown to be beneficial in predicting stress during reading tasks (Sharma & Gedeon, 2012, 2013a). However, the use of task windowing did not provide any significant improvement; it was instead the GA feature selection that provided the substantial improvement. However, we have only considered a very short reading task. Each text had on average 240 words, which took on average 84 seconds to read. This is not a long task when we consider many online collections of learning materials. Further analysis of windowing should be investigated for longer tasks.

In this chapter we investigate ANNs, kNNs, and random forests as predictors. One of the disadvantages of using ANNs and random forests in the need for longer training times, which is made significantly lower when combined with GA feature selection. KNN on the other hand does not suffer from the same problem. The results from the study indicate that in this case kNN is optimal for predicting reading comprehension from eye gaze measures when GA feature selection is used. Sharma and Gedeon (2013a) found that using GA feature selection and support vector machines (SVM) provided high classification rates of stress during reading. The use of GA-SVM should also be considered in this case. Additionally, as noted by Sharma and Gedeon (2013a), even though better classification results were obtained using GA feature selection, the execution time is substantially longer than for other methods. In our investigation we utilised smaller populations and much fewer generations to increase speed. Further exploration using larger populations and more generations should be investigated to determine if there is a general optimal trade-off between accuracy and training time.

6.5.1 Use case

The main goal of reading comprehension detection is to incorporate eye tracking into eLearning environments and use the eye tracking data as a form of adaption. The use of reading comprehension prediction to perform adaption was discussed in Chapter 5, where we outlined how eLearning material can be tailored to a student's current level of understanding. The main advantage of using eye-tracking data over explicit answers to questions is that the comprehension questions could be altogether removed. In turn this speeds up the learning session as well as alleviating stress and anxiety over not knowing answers or questions being too simple. Instead, an individualised dynamic learning path could be followed and students would not even be aware that what they are reading is possibly different from their peers.

Another use of comprehension prediction in eLearning is similar to the concepts put forward by Buscher et al. (2012) and Okoso et al. (2015) whereby part of the documents are labelled based on reading behaviour. In the first case, Buscher et al. (2012) propose using eye tracking to annotate parts of a document that contain many eye movements. Okoso et al. (2015) propose finding parts of documents that are not comprehended. Conflation of the two ideas with the current research leads to the notation of real time detection of comprehension levels and annotation of the learning documents with these levels. The student would not be aware of these annotations but the learning environment could adaptively reshow parts of the text that are not well understood. In this way, when a student clicks a *Next* button to move to the next text in the sequence, the next text could be dynamically selected to also reflect the parts of the text that were not well understood. In this way, it is quite similar to the use proposed in the previous chapter, however, the idea is refined in this case so that instead of the whole text being re-shown to students (and thereby giving them information they already understand), only the subsections of text that are not well understood could be re-shown together with new information.

Alternatively, instead of re-showing the students text plainly, the use of questions and text presentation (as discussed in Chapter 3) could be used to exploit the student's answer-seeking behaviour and reading behaviour to encourage them to read certain parts of the text more thoroughly. That is, instead of giving students another page of text, giving them a page of text with questions as well. The questions themselves could be related only to the parts of the text that were not well understood based on the eye tracking predictions. This could encourage re-reading of these sections to answer the questions. In the case that it does not, the feedback of getting the questions wrong will then encourage re-reading of these sections.

6.6 Conclusion and Further Work

The goal of this chapter was to discover techniques for increasing prediction performance for reading comprehension. We investigated the effects that reader type and text difficulty have on predicting reading comprehension from eye gaze data. In the analysis we were able to achieve 80% classification accuracy for the L1 group and 73% for the L2 group, which is a substantial improvement from the 44% correct classification obtained in the previous chapter for format C. We did not find

that the text difficulty improved predictive accuracy, however, the use of feature selection did provide significantly higher predictions then without. We also experimented with the addition of pupil dilation feature set. The analysis cannot confirm that the addition of this feature was significant. Further work considering the accuracies of the classification using only the pupil dilation feature set should be carried out. This would explore if there are added benefits of using these features. Furthermore, use of only these feature set should be considered to see if this is better than the eye movement measures altogether.

Whilst there was no significant difference in prediction accuracy found due to the text difficulty, we did show that the text readability has a significant effect on eye movements, whereas, conceptual difficulty has no effect on eye movement. This should be investigated further as it was hypothesised that the conceptual difficulty would also affect eye movements. The implications of only the surface variables of the text affecting eye movement is interesting and has important side effects in the context of eLearning.

We also investigated the use of task window and GA feature selection. This showed that whilst windowing provided no improvement, the GA feature selection did improve predictions. Additionally, we found that the best predictor, for this problem type, from the set that we investigated is kNN. Given the results from the windowing, further investigation should be considered where longer documents are read. In our case the documents were quite small; barely longer than the abstract of this chapter, or approximately the length of this paragraph. The hypothesis is that increased length will make windowing useful. However, the current windowing results show a reasonable possibility for labelling paragraphs or sections of text with a level of comprehension. In this way we could move towards real time comprehension detection of small sections of documents. This requires further work, again by investigating longer documents.

Up to this point we have only considered predicting reading comprehension based on eye gaze data. The data collected from this user study allows us to make predictions about the text difficulty. Given the interesting results obtained from the eye movement analysis several questions arise; 1) if text difficulty affects L1 and L2 readers eye movements differently, are their perceptions of text difficulty also affected differently? 2) Since the eye movements are not as we hypothesised, are the participants' perceptions of difficulty more accurate than their eye movements at predicting readability? Finally, 3) given that the readability and conceptual difficulty have different effects on the L1 and L2 groups, is this reflected in their perceptions? In the next chapter we investigate the comparison of text difficulty prediction from eye gaze data versus participants' perceptions and also perform more analysis on the effect of text difficulty on the L1 and L2 readers.

Chapter 7

Perception and prediction of Text Difficulty

"Two people can look at the same thing and see it differently."

— Justin Bieber

Up to this point we have only considered predicting reading comprehension from eye gaze data. Given the results in Chapter 6, several questions arise around whether participants can predict text difficulty. In this chapter, we investigate prediction of text difficulty from eye gaze using machine learning techniques, and compare these to participants' perceptions of difficulty. We show that predictions from eye tracking data are more accurate than the participants' perceptions of both readability and conceptual difficulty. We then show that prediction of participants' perceived ratings of readability and conceptual difficulty from the eye tracking data are also better than prediction of the predefined values. This indicates that the eye gaze measures and pupil dilation data may be more aligned with the participants' perceptions of difficulty rather than the predefined difficulty of the text. Further analysis of participants' perceptions showed that they are poor at predicting predefined text difficulty, especially when the readability and the conceptual difficulty are not the same. The readability and conceptual difficulty of a text interact with each other to distort participants' perceptions of overall text difficulty. Further analysis of text difficulty on participants' perceptions shows that text difficulty does not affect participants subjective understanding but does have a significant effect on comprehension. However, the effect is minimal, where the only significant difference is the scores for the easiest (A) compared to the hardest (J). This suggests that comprehension score alone is not a sufficient indicator of text difficulty. Nevertheless, L1 readers scored higher on comprehension questions compared to L2 readers, contrary to past research, and text difficulty did not affect L2's confidence in answering the questions. The analysis highlights that there are

significant differences in perceptions of L1 and L2 readers, which must be considered when designing texts for education.

7.1 Introduction

Reading online materials is essential as it is a primary way of accessing many forms of information. Much has been done in researching effective ways of presenting learning materials in learning environments (Clark & Mayer, 2011). There has also been headway on investigating the growing diversity of students in eLearning, specifically by linguistic background. It has been established that first (L1) and second (L2) English language readers have different reading behaviour (Dednam et al., 2014; Kang, 2014). However, how we perceive a task does not always match our performance on that task and people often see the same thing differently. For example, it has been shown that people who are unskilled are also unaware of their deficiency and so overrate their abilities in comparison to the appropriate cohort; especially, they think they are above average. Conversely, skilled people tend to know their shortcomings and underrate their abilities in comparison to the cohort (Dunning et al., 2003; Ehrlinger et al., 2008; Kruger & Dunning, 1999). Indeed, task complexity and perceptions of task complexity are distinct but are related to one another (Robinson, 2007). Task complexity affects perceptions of difficulty as well as confidence levels (Robinson, 2007). A complex task does not guarantee that perception of that task will be that it is complex. This is especially true when comparing readers who have different levels of expertise in the language they are reading.

In Chapter 3 we touched on the perceptions of students in eLearning where we recorded their perceptions of their understanding. We found that there was no significant difference between L1 and L2 readers in their perceived comprehension, but there is a difference in accuracy of these perceptions based on the presentation format the participant was shown. That is, when the comprehension questions are shown in isolation from the text, participants were more likely to be able to correctly perceive their understanding as opposed to when the questions were shown on the same page as the text. However, the texts had the same level of difficulty in that study, so we could not investigate how changing the difficulty affects participants' predictions. So whilst L1 and L2 participants had the same perceived understanding in the previous study, this does not imply that both groups found it equally challenging. The user study described in Chapter 6 provides the opportunity to investigate the question of whether participants can predict text difficulty and whether the we can predict text difficulty from participants' eye tracking data. Therefore, the question being investigated in this chapter is:

Can participants predict text difficulty and can we predict text difficulty from their eye gaze?

Text difficulty in this context is a combination of readability and conceptual difficulty. The use of eye gaze has shown potential for predicting task difficulty (Rayner et al., 2006; Victor et al., 2005). In particular, we know that pupil dilation information is related to task difficulty (Engelhardt et al., 2010; Iqbal et al., 2004;

Pomplun & Sunkara, 2003; Zekveld et al., 2014). We therefore hypothesise that both eye gaze and pupil dilation data can be used to predict text complexity. Does text difficulty affect perceptions of participants, and does it do so in the same way for L1 and L2 readers? We hypothesize that changes in text difficulty will be reflected by changes in perceived difficulty but that participants' eye gaze data will be more accurate at predicting the text difficulty than their perceptions.

This chapter is organized into the following sections: background information; prediction results; perception analysis; discussion and implications; and finally the conclusion and further work.

7.2 Background

7.2.1 Defining text difficulty

The definition of text difficulty that we will refer to in this chapter is adapted from the Common Core State Standards (Initiative, 2012) which is an educational initiative in the United States. This standard defines text difficulty as a combination of three components: qualitative, quantitative, and reader and task considerations. The quantitative component is based on the text structure and calculated from a formula using word and sentence structure; examples are the Flesh-Kincaid readability test calculating the education a reader needs to comfortably read a piece of text (Kincaid et al., 1975). Tests with more dimensions include COH-Metrix, which produces measures defining the cohesion of a document as well as the readability for L2 readers (Crossley et al., 2008; McNamara et al., 2013). The qualitative component refers to the levels of meaning and knowledge demands. The last component of text difficulty is not so much related to the text itself but to the reader and the task being performed. That is how motivation, prior knowledge, task and purpose influence the text difficulty (Bunch et al., 2014).

7.2.2 Differences between L1 and L2 readers

The differences between L1 and L2 readers has growing importance given the wide spread and pervasive use of the Internet and World Wide Web. Access to texts that are not written in a reader's native language is now easy and often required especially for study. The impact of this on learners is of great importance for designers of eLearning environments, as they must take into consideration the differences between L1 and L2 readers. There are differences in eye movements as well, for example L2 Afrikaans readers exhibit more fixations and for longer duration than L1 readers (Dednam et al., 2014). This is consistent with what we found for L1 and L2 English readers in Chapter 3.

The differences between L1 and L2 readers can be seen in their reading behaviours. Kang (2014) found that L1 and L2 English readers performed no differently in comprehension tests and that there was no difference in reading attention distributions or eye gaze patterns, but L2 readers took longer to read the text and longer to find answer cues in the text. Notably, L1 readers tend to deal with increases of text difficulty with increased reading efficiency, whereas, L2 reading

efficiency decreases (Dednam et al., 2014). Text characteristics should be considered differently for L1 and L2 readers since they have differential effects on reader type (Zhang et al. 2013).

7.2.3 Prediction of text readability

Text characteristics include word count, syllable count and number of words in a sentence, which are often used to calculate readability. The readability formula used throughout this thesis is the Flesch-Kincaid grade level which is the most widely used readability test, taking into account only the total number of words, sentences and syllables (Kincaid et al., 1975). The such characteristics have been linked to greater difficulty in reading, according to eye movements, which is also linked to lower comprehension (Scherr et al., 2015). Since readability has an effect on reading behaviour and comprehension, it is important to consider how it is calculated. Most automated tools for detecting text difficulty focus primarily on the readability of the text, based on the syntactic nature of the text. That is, the traditional formulas rely on counting words, word length, sentences length, and syllables. The Flesch-Kincaid grade level is one the most widely used readability test, taking into account only the total number of words, sentences and syllables. Whilst this formula is quick and easy to use in practice there are two potential problems with it, firstly it only deals with the surface properties of the text, not accounting for the conceptual difficulty and secondly it is generally aimed at English text for native English readers (Zhang et al. 2013).

COH-Metrix is an important tool in this context as it provides a bridge to overcome these faults. COH-Metrix measures text cohesion at various levels of selected language, discourse, and conceptual analysis to provide a measure of readability from a cognitive view (Crossley et al., 2008; Graesser et al., 2011; McNamara et al., 2014). COH-Metrix has been found to be better at predicting the reading difficulty than traditional readability formulas (Crossley et al., 2008). Additionally, COH-Metrix produces a measure of readability for second language readers, the L2 Readability Index (Crossley et al., 2008; Graesser et al., 2011; McNamara et al., 2013). This has been shown to be useful in assessing the difficulty of texts and highlighting differences between L1 and L2 readers (Zhang et al., 2013). Note that we investigate the L2 Readability Index further in Chapter 8.

7.2.4 Perceptions of task complexity

Perceptions can play an important part in learning and how students approach study. Perceptions of heavy workload with inappropriate assessment promotes surface learning whereas perceptions of good teaching and appropriate assessment promotes deep learning and are a stronger predictor of learning outcomes (Lizzio, Wilson, & Simons, 2002). Additionally, confidence in performing the underlying task influences perceptions of task difficulty. This has been shown in the area of programming studies where computer confidence has a significant effect on perceived task difficulty (Chang, 2005). Importantly, managing perceptions can help alleviate anxieties in learning. Task complexity affects perceptions and confidence (Robinson, 2007). The importance of managing perceptions of L1 and L2 readers is

necessary given that L2 readers have great perceived difficulties with hard texts (Dednam et al., 2014).

7.3 Method

7.3.1 Data collection

The method for the data collection user study was described in Chapter 6 and will not be repeated here. Instead this section will outline prediction of text properties from eye gaze measures. Firstly, we will recap the text difficulty properties, as they are crucial in this investigation. The texts had differing levels of difficulty that are a combination of three different levels of readability, as measured from the Flesch-Kincaid Grade Level (Kincaid et al., 1975), and three different levels of conceptual difficulty, constructed independently by a colleague. This resulted in nine texts with differing difficulty, which is described by the grid system in Figure 7.1.

	Readability		
Concept	**Easy**	**Medium**	**Difficult**
Easy	A	B	C
Medium	D	E	F
Difficult	G	H	J

Figure 7.1. Description of the text difficulty

Note that there is no correlation between the Flesch-Kincaid Grade Level and the conceptual level for each text in each topic (r=-0.1, for all topics). The Flesch-Kincaid Grade Level does not account for any changes in the conceptual level of the text, which was not only expected but also necessary for the above grid system construction of text complexity.

After each piece of text, participants were asked two comprehension questions to assess their understanding of the text and four qualitative questions related to the text they read. These questions are:

5. How well do you think you understood the text?
   (Very well / Well / Somewhat / Not at all)
6. How confident were you answering the questions?
   (Very confident / Confident / Not Confident)
7. How difficult did you find the text to read?
   (Easy / Moderate / Hard)
8. How complex was the concept being explained in the text?
   (Basic / Intermediate / Advanced)

7.3.2 Prediction method

Once again windowing is used along with GA feature selection. The GA parameters and explanation for feature selection are described in Chapter 6. In this analysis we use k-nearest neighbour (kNN) classification and GA feature selection method used by Oluleye et al. (2014). The classification results were generated from nested 10-fold cross validation. All analyses were carried out using Matlab R2013a.

7.3.3 Data Pre-processing for prediction

7.3.3.1 Inputs: Eye gaze and Pupil dilation data

The inputs to the classifier are the same as those defined in Chapter 6. Please refer to Chapter 6 for details regarding the inputs.

7.3.3.2 Outputs

Each text has a readability level of *Easy*, *Medium*, and *Difficult* and a conceptual level of *Basic*, *Intermediate*, or *Advanced*. In this analysis we look at the prediction of both the readability and the conceptual levels. In the final part of the analysis we look at the overall text difficulty which is the product of both the readability and the conceptual difficulty. These outputs refer to the text IDs, as shown in Figure 7.1, A through J.

7.4 Predicting text difficulty

In this section we analyse whether participants are able to predict text difficulty, and compare their perceptions of text difficulty to predictions of text difficult from their eye movements. Each text has a difficulty that is a product of the readability and the conceptual levels, which are first considered separately, and then the combination is considered to assess overall perceptions of text difficulty.

7.4.1 Predictions of conceptual difficulty

The first part of the investigation is prediction of the conceptual difficulty. We control for the readability of the text. Where the *Easy*, *Medium*, and *Difficult* are the three grades of readability. For the kNN classifier this is a three-class problem where the conceptual difficulty can be: *Basic*, *Intermediate*, or *Advanced*. However, the readability also differs for each text by three variables. We consider three cases since we control for the readability of the text. We therefore predict the conceptual difficulty for each of the readability levels. The average correct classification rates (%) from the nested 10-fold cross validation for the GA-kNN classification are shown in Figure 7.2.

Figure 7.2 indicate that the predictions of text difficulty from the GA-kNN are more accurate than participants' ratings of conceptual difficulty, for every level of readability. One explanation for this could be that whilst participants are not *consciously* aware of the correct level of difficulty, they are *non-consciously* aware of the difference since their eye movements reflect the difficulty to a higher degree.

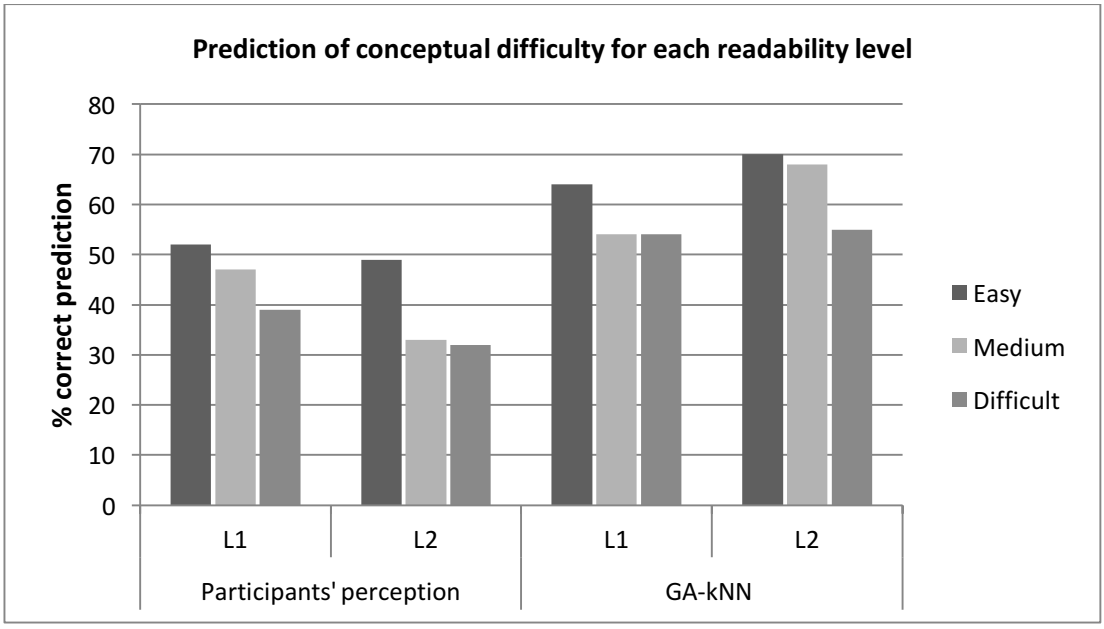


Figure 7.2. Participant versus GA-kNN predictions of conceptual level (for each level of readability)

ANOVA of the predictions shows that there is a significant difference between participants' perceptions of text difficulty compared to the predictions from the GA-kNN ($F(1,2)=33.51$; $p=0.029$). However, there is neither a significant difference between the L1 and L2 reader groups ($F(1,2)=0.02$; $p=0.892$) nor between the different levels of readability ($F(2,2)=0.02$; $p=0.892$).

The participants' predictions of conceptual difficulty are poor; given that there are three classes, chance identification is 33%, participants' ratings are close to chance. We see that for the L1 participants' predictions of conceptual difficulty, as the readability becomes more difficult participants' ratings of conceptual difficulty decrease in accuracy. However, this is not seen as evidently for the L2 participants since the ratings for the concept level is (almost) the same for both the *Medium* and the *Difficult* levels of readability. This could be why we see no significant difference between the levels of readability.

The GA-kNN predictions from the eye tracking data also gets worse for each readability level. This might imply some sort of weak interaction between readability and conceptual difficulty, by which the difficulty in readability is masking the difficulty in conceptual level. The eye tracking data provide slightly better predictions of conceptual difficulty for the L2 group compared with the L1 group. Prediction of the conceptual difficulty is about double chance (64%) for the L2 participants. In this case, whilst the L2 participants are slightly worse at predicting the conceptual difficulty of the text, their eye movements, at least in the easier levels of readability, reflect the conceptual difficulty to a higher extent compared to the L1 participants.

Notably the predictions from the eye tracking data for the L1 group are similar to the L2 participants' ratings where the predictions for the *Medium* and *Difficult*

texts are the same. However, there a clear difference between the predictions for the *Medium* and *Difficult* but not between the *Easy* and the *Medium* texts from the eye tracking data from the L2 participants. Perhaps L1 readers use the same eye movement strategies for *Medium* and *Difficult* text while L2 readers use consistent strategies for Easy and Medium texts.

7.4.2 Predictions of readability

The second part of the analysis is prediction of the readability level of the text. Once again we consider three cases since we control for the conceptual difficulty of the text in this section. We therefore predict the readability for each of the conceptual levels: *Basic*, *Intermediate*, or *Advanced*. The average correct classification rates (%) from the nested 10-fold cross validation for the GA-kNN classification are shown in Figure 7.3.

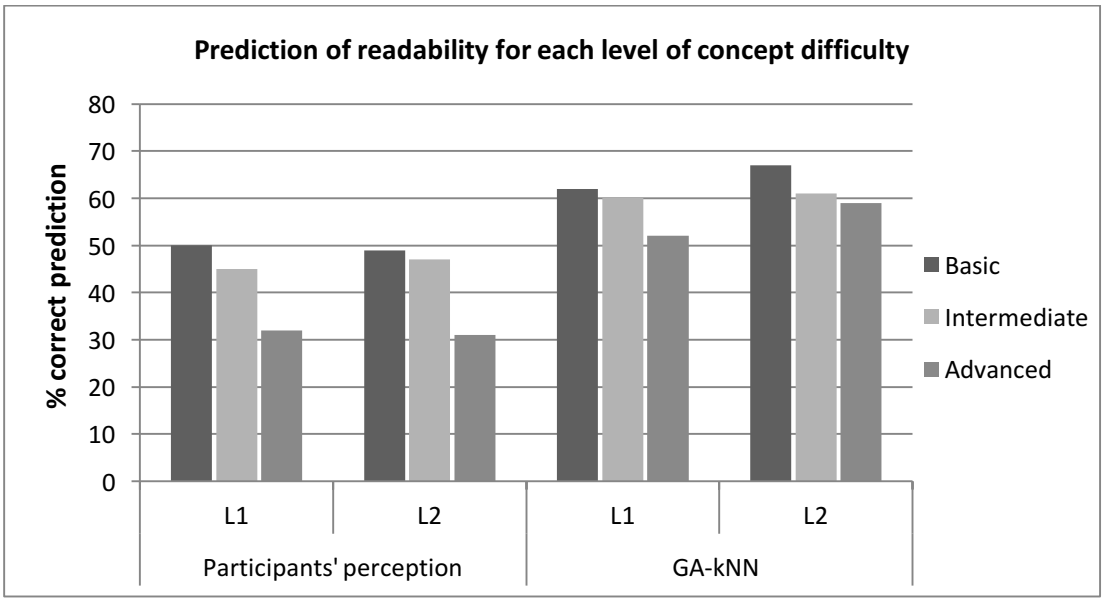


Figure 7.3. Participant versus GA-kNN prediction of readability level (for each level of conceptual difficult)

ANOVA of the predictions shows that there is a significant difference between participants' perceptions of text difficulty compared to the predictions from the GA-kNN ($F(1,2)=170.88$; $p=0.0058$) and between the levels of conceptual difficulty ($F(2,2)=34.79$; $p=0.0279$). However, there is no significant difference between the L1 and L2 reader groups ($F(1,8)=0.25$; $p=0.6332$).

The results of the analysis shown in Figure 7.3 are similar to what was found in Section 7.4.1. Once again, the GA-kNN predictions are more accurate predictions of readability than participants' prediction of readability, for each level of conceptual difficulty. Similarly, the accuracy of predictions of readability decrease as the concept level increases in difficulty, however, this time the difference is significant. This is an interesting finding as it appears that might be some sort of interaction between readability and conceptual difficulty, by which the conceptual difficulty masks the participants' ability to detect difficulty in readability.

The L1 and L2 groups rate the readability level with quite similar accuracy. This is also true for the predictions from the eye tracking data. That is, the readability appears to affect both the L1 and L2 participants in similar ways, no matter what the conceptual level.

7.4.3 Prediction of explicit perception of difficulty

An important question to now ask given the results in Sections 7.4.1 and 7.4.2 is whether the eye tracking data is more related to the predictions of the participants, rather than to the predefined levels of readability and difficulty. In this section we assess this question. The results from prediction of participants' perceptions of conceptual difficulty from their eye tracking data using the GA-kNN, with nested 10-fold cross validation, for each level of readability are shown in Figure 7.4.

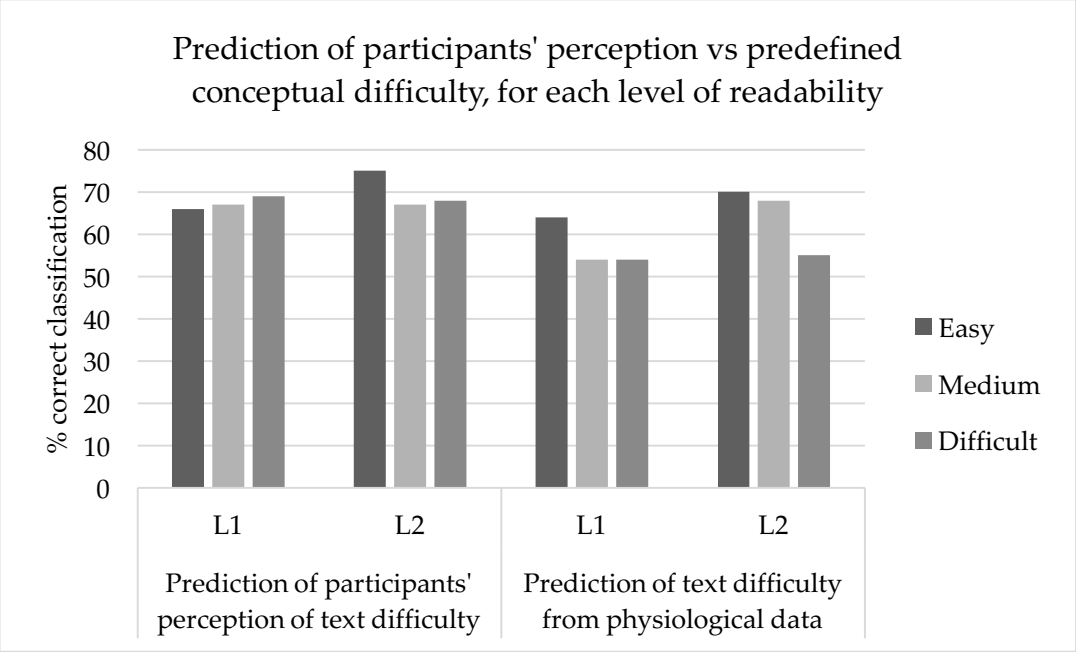


Figure 7.4. Classification of perceived conceptual difficulty versus predefined conceptual difficulty from eye tracking data

ANOVA of the predictions shows that there is a significant difference between prediction of the participants' perceptions of conceptual difficulty to prediction of the predefined conceptual difficulty, both based on their eye tracking data ($F(1,8)=6.09$; $p=0.0389$) but that there is no difference between the L1 and L2 readers ($F(1,2)=2.32$; $p=0.1665$). From Figure 7.4 we can see that the prediction accuracy for the participants' perceptions of conceptual difficulty from their eye physiological data are higher than prediction of the predefined text difficulty, using the same data. This indicates that eye tracking data might be more linked to the participants' perceptions of conceptual difficulty rather than the actual conceptual difficulty. This could explain the lack of significant difference in eye movement measures between conceptual difficulty levels, as discussed in section 6.4.3.

We now consider prediction of participants' ratings of readability level whilst controlling for the conceptual difficulty. The results from prediction of participants'

ratings of readability from their eye tracking data using GA-kNN, with nested 10-fold cross validation, for each level of conceptual difficulty, are shown in Figure 7.5.

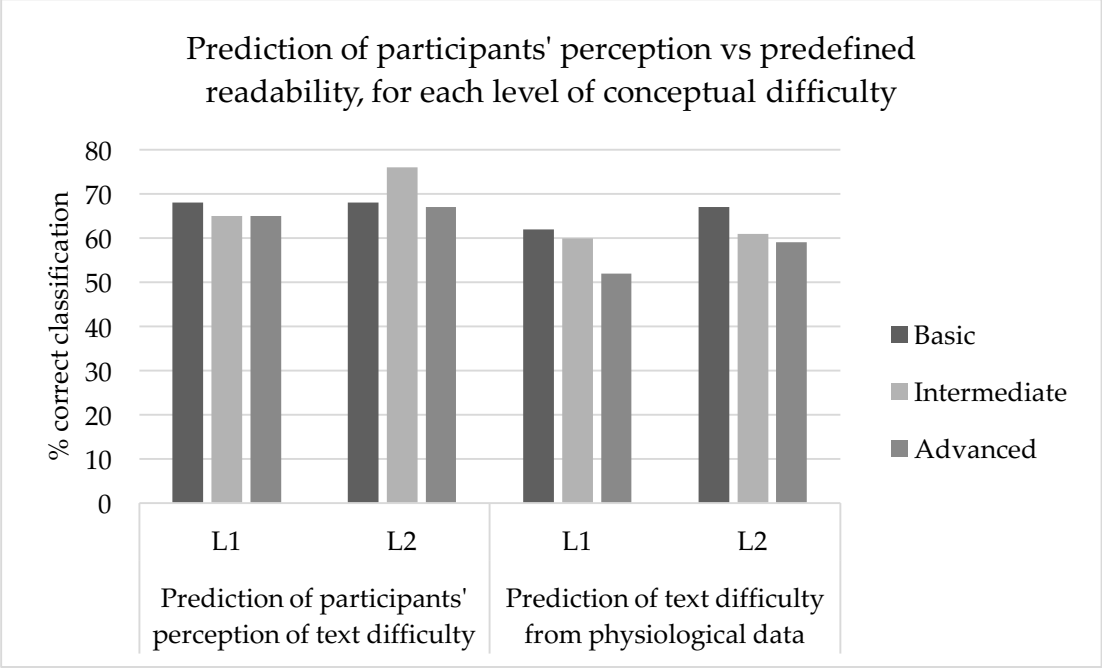


Figure 7.5. Classification of perceived readability level versus predefined readability level from eye tracking data

ANOVA of the predictions shows that there is a significant difference between prediction of the participants' perceptions of readability to prediction of the predefined readability, both based on their eye tracking data ($F(1,8)=10.57$; $p=0.0117$) but that there is no difference between the L1 and L2 readers ($F(1,2)=3.1$; $p=0.1163$). We observe in Figure 7.5, that the prediction accuracy of participants' perceived level of readability are higher compared to the prediction of predefined readability using participants' eye tracking data. Once again, this could indicate that eye tracking data is more linked to the participants' perceptions of readability rather than the actual text difficulty. Whilst we found that readability affected eye NNFs, but it did not affect the regression ratios, and the only difference found in the NNFs was between the easy level of readability and the difficulty level of readability, and not the medium level of readability. The lack of difference between the easy and medium as well as the medium and the difficult levels indicates that the readability did not have a large effect on the NNF. The higher accuracy of participants' perceptions of readability compared to the predefined readability could explain the lack of significant difference in eye movement measures between all readability levels, as discussed in section 6.4.3.

7.4.4 Prediction of overall text difficulty

One of the questions raised from the analysis of eye movement data in Chapter 6 is that given readability and conceptual difficulty have different effects on the L1 and L2 groups, is this difference reflected in their perceptions? In this section we will explore this question by delving deeper into participants' explicit perceptions, as

well as further investigating the use of eye tracking data to predict overall text difficulty.

7.4.4.1 Prediction of text difficulty

In this section we investigate if there are interactions or effects of readability and conceptual difficulty on perceptions. A Chi-square test shows that the text difficulty affects perceptions of conceptual level for both groups ($\chi(16)=51.7$, $p=0.001$ for L1 readers, and $\chi(16)=53.3$, $p=0.001$ for L2 readers). Text difficulty also has an effect on perceived readability level for L2 readers, but not on L1 readers ($\chi(16)=19.4$, $p=0.247$ for L1 readers, and, $\chi(16)=47.0$, $p=0.001$ for L2 readers). These results highlight two key points; the perceptions of text difficulty are different between the L1 and L2 readers and changes in text difficulty are reflected in perceived difficulty.

Table 7.1. Expected versus reported text difficulty for L1 readers

Actual Text Difficulty			Perceived Text Difficulty (%)								
ID	Read.	Conc.	A	B	C	D	E	F	G	H	J
A	Easy	Basic	**45**	13	1	17	19	3	0	0	1
B	Mod.	Basic	36	**10**	0	19	26	5	0	2	2
C	Diff.	Basic	38	5	**0**	19	19	5	0	14	0
D	Easy	Int.	37	5	2	**10**	37	2	2	5	0
E	Mod.	Int.	22	5	0	18	**38**	5	2	5	4
F	Diff.	Int.	7	13	0	20	27	**7**	13	7	7
G	Easy	Adv.	35	0	0	20	35	5	**0**	0	0
H	Mod.	Adv.	32	5	0	23	27	5	0	**5**	5
J	Diff.	Adv.	11	6	0	17	44	0	6	17	**0**
Average			*29*	*7*	*0*	*18*	*30*	*4*	*3*	*6*	*2*

Given that the perception scale is the same as the scale used to rate the texts by the author, we can put these ratings together to come up with the same grid references as shown in Table 7.1 for L1 readers and Table 7.2 for L2 readers. We compare the expected to the reported percentages of text difficulty. A Chi-square test for independence shows that there is a strong relationship between the text difficulty and the perceived text difficulty for the L1 readers ($\chi(64)=92.7$, $p=0.01$) and the L2 readers ($\chi(64)=99.1$, $p<0.005$). However, it is clear that participants, both L1 and L2, are poor at perceiving the predefined text difficulty. This is signified by the main diagonal (in bold) and contains very low percentages. Instead the difficulty affected the participants' perceptions in other ways, which we will elaborate upon in the section below.

As just stated, L1 readers are poor at perceiving the actual text difficulty. With the combined variables of text difficulty, L1 participants correctly classify 47% of the texts, which is well above chance prediction of 11%. As the Chi-square test showed, text difficulty and perceived text difficulty are not independent, so we would expect that participants perform above chance. Most L1 readers rate texts as A, D or E. That is, mostly participants think the texts have an easy level of readability with either an

easy or intermediate level of conceptual difficulty, or a moderate readability with intermediate conceptual difficulty. The interesting issue about this is that A and D have an easy readability but different levels of conceptual difficulty and E has both intermediate readability and conceptual level. However, L1 readers seem unable to distinguish the readability levels from the conceptual difficulty.

Table 7.2. Expected versus reported text difficulty for L2 readers

Actual Text Difficulty			Perceived Text Difficulty (%)								
ID	Read.	Conc.	A	B	C	D	E	F	G	H	J
A	Easy	Basic	**36**	15	0	6	40	1	0	1	0
B	Mod.	Basic	12	**6**	3	6	59	6	0	3	6
C	Diff.	Basic	0	0	**0**	0	50	20	0	10	20
D	Easy	Int.	14	14	3	**11**	43	0	0	9	6
E	Mod.	Int.	7	19	0	4	**52**	0	0	15	4
F	Diff.	Int.	18	9	0	0	27	**9**	0	18	18
G	Easy	Adv.	22	0	0	0	56	11	**0**	0	11
H	Mod.	Adv.	14	0	0	29	29	0	0	**0**	29
J	Diff.	Adv.	10	0	0	0	30	0	0	20	**40**
Average			*15*	*7*	*1*	*6*	*43*	*5*	*0*	*8*	*15*

Furthermore, L1 readers appear to see texts as being simpler than they are, as many of the ratings are in the lower half of the diagonal. More specifically, when the readability and conceptual levels are at opposite extremes to one another (texts C and G) we see interesting interactions that reveal much about the nature of the interaction between readability and conceptual levels. That is, when the readability was difficult and the concept basic (text C) the majority of L1 readers rated the text with intermediate and advanced concept levels and varying degrees of readability. Conversely, when the readability is easy but the concept is advanced (text G) no L1 participant rated the text with advanced conceptual level and instead the majority rate it with intermediate concept level and differing degrees of readability. Very few L1 participants rated the texts as having difficult readability (C, F and J). There are also low ratings for B, G and H. Essentially participants are poor at perceiving the most complex texts as well as the interactions between the readability and conceptual difficulty of the text. There is an interaction that masks the two variables resulting in a rating somewhere in the middle. More specifically, these texts are rated as moderate in readability and intermediate in conceptual difficulty (E). Participants extrapolate the difficulty as being somewhere in the middle of the two variables. This poses an interesting question, how distinguishable is readability from conceptual level to the reader?

With the combined variables of text difficulty L2 participants can correctly classify 45% of the texts, similar to the L1 group, this is well above chance, as we would expect. Again, we see that many L2 readers rate texts as E, both intermediate readability and conceptual level, which is similar to the L1 readers. However, the perceptions of L2 readers are somewhat different from L1 readers. There is a spread

of complexity ratings for the texts, with many more L2 readers rating texts as more difficult than they are. The knock on effect of this is that the L2 readers are capable to perceive the most difficult text J compared to L1 readers. As hypothesised, L2 readers tended to rate texts with higher difficulty compared to the expected difficulty.

L2 readers do however show the same behaviour as the L1 readers whereby the perceived complexity of the text is mostly conflated so that there is a significant over estimation of texts being rated as both *Moderate* readability and *Intermediate* conceptual level (E). More generally, L2 readers mainly rated texts as A, E and J, where the readability and conceptual levels are the same levels.

To summarise, the key highlights of the perception analysis are:

1. L1 and L2 readers have different perceptions:
    a. L2 readers tend to overestimate difficulty of the text
    b. L1 readers tend to underestimate the difficulty of the text
2. Both groups over estimate complexity as E where the readability and conceptual difficulty are both in the middle of the scale
3. Both groups tend to conflate the levels of readability and conceptual difficulty, thus under estimating all texts surrounding the main diagonal, especially as complexity ratings C and G.
4. Their eye movements are a reflection of both the predefined and explicitly perceived text difficulty

The two extremes in the readability and conceptual level do not mix well when they are inverses of one another, (texts C and G). The interaction between the two variables results in an underestimation of one variable and overestimation of the other. If the desired effect is to make a concept appear harder or easier, then the readability can be changed to achieve this. An example of this is underestimation of text B. Conveying a basic concept in text with difficulty readability will cause perceptions of the conceptual difficulty to be overestimated.

7.4.4.2 Predictions of predefined text difficulty from eye tracking data

To contrast the results of the explicitly perceived text difficulty, we investigate the predictions of predefined text difficulty from the eye gaze and pupil dilation data. GA-kNN classification is used once again to predict the predefined text difficulty, denoted as A through to J. The average correct classification rates (%) from nested 10-fold cross validation are summarised in Table 7.3 for the L1 readers and Table 7.4 for the L2 readers.

For the L1 data set we obtained an average correct classification of text difficulty of 49% from 10-fold cross validation. This is roughly the same compared with the explicit perceptions of participants, where the correct classification rate is 47%. The GA-kNN also has a tendency to predict texts as being simpler than they in fact are which is consistent with the perception analysis. Given the lack in substantial difference between eye movements based on text difficulty as shown in Chapter 6 (section 6.4.3), we would not expect the predictions from the eye gaze and pupil

dilation data to be substantially more accurate compared to participants' perceptions.

Table 7.3. Average correct classification rates (%) of text difficulty for the L1 group from GA-kNN classification from eye tracking data

Actual Text Difficulty			Predicted Text Difficulty (%)								
ID	Read.	Conc.	A	B	C	D	E	F	G	H	J
A	Easy	Basic	**90**	3	1	3	2	0	1	0	0
B	Mod.	Basic	43	**41**	2	9	2	2	2	0	0
C	Diff.	Basic	43	29	**19**	5	5	0	0	0	0
D	Easy	Int.	46	3	0	**46**	5	0	0	0	0
E	Mod.	Int.	47	16	0	9	**27**	0	0	0	0
F	Diff.	Int.	53	7	13	13	0	**7**	0	0	7
G	Easy	Adv.	60	5	10	5	5	0	**15**	0	0
H	Mod.	Adv.	59	14	5	5	9	0	0	**9**	0
J	Diff.	Adv.	56	11	0	11	0	0	6	0	**17**
Average			*55*	*14*	*5*	*12*	*6*	*1*	*3*	*1*	*3*

Table 7.4. Average correct classification rates (%) for the L2 group of text difficulty group from GA-kNN classification from eye tracking data

Actual Text Difficulty			Predicted Text Difficulty								
ID	Read.	Conc.	A	B	C	D	E	F	G	H	J
A	Easy	Basic	**96**	3	0	1	0	0	0	0	0
B	Mod.	Basic	47	**29**	0	18	6	0	0	0	0
C	Diff.	Basic	50	10	**40**	0	0	0	0	0	0
D	Easy	Int.	63	3	0	**26**	6	0	0	0	3
E	Mod.	Int.	25	11	0	14	**46**	0	0	0	4
F	Diff.	Int.	64	0	0	9	9	**18**	0	0	0
G	Easy	Adv.	67	22	0	11	0	0	**0**	0	0
H	Mod.	Adv.	57	29	14	0	0	0	0	**0**	0
J	Diff.	Adv.	20	40	0	0	20	0	0	0	**20**
Average			*54*	*16*	*6*	*9*	*10*	*2*	*0*	*0*	*3*

For the L2 data set we obtained an average correct classification of text difficulty of 50% from cross validation. Compared to the explicit perceptions of participants, which have a classification rate of 45%, this is a slight improvement. Similar trends in prediction accuracy for each text are seen where the easier texts are best predicted.

7.5 Effects of text properties on understanding and confidence

We now move to analysing whether text difficulty affects participants' comprehension, perceived understanding, and confidence in answering questions, and if so, is it in the same way for L1 and L2 readers? We hypothesise that harder texts will be associated with lower comprehension scores, and that L2 readers will have lower comprehension compared to L1 readers for the harder texts. Comprehension is a quantitative measurement so we will also look into the qualitative data and consider what text difficulty does to participants' confidence and perceptions of understanding the text. Task difficulty is known to effect perceptions of difficulty as well as confidence levels (Robinson, 2007) so our hypothesis is that harder texts will be associated with lower ratings of confidence and perceived understanding.

7.5.1 Comprehension versus perceived understanding

Participants were asked to rate their understanding level in answering the questions. These ratings are described in Figure 7.6 for L1 readers and Figure 7.7 for L2 readers. The ratings of understanding were recorded on a Likert scale of *Very Well*, *Well*, *Somewhat*, and *Not at all*. Using Chi-square test for independence we observe that text difficulty does not affect the ratings of subjective understanding for L2 readers ($\chi(24)=23.59$, $p=0.485$) or L1 readers' ratings ($\chi(24)=35.81$, $p=0.06$).

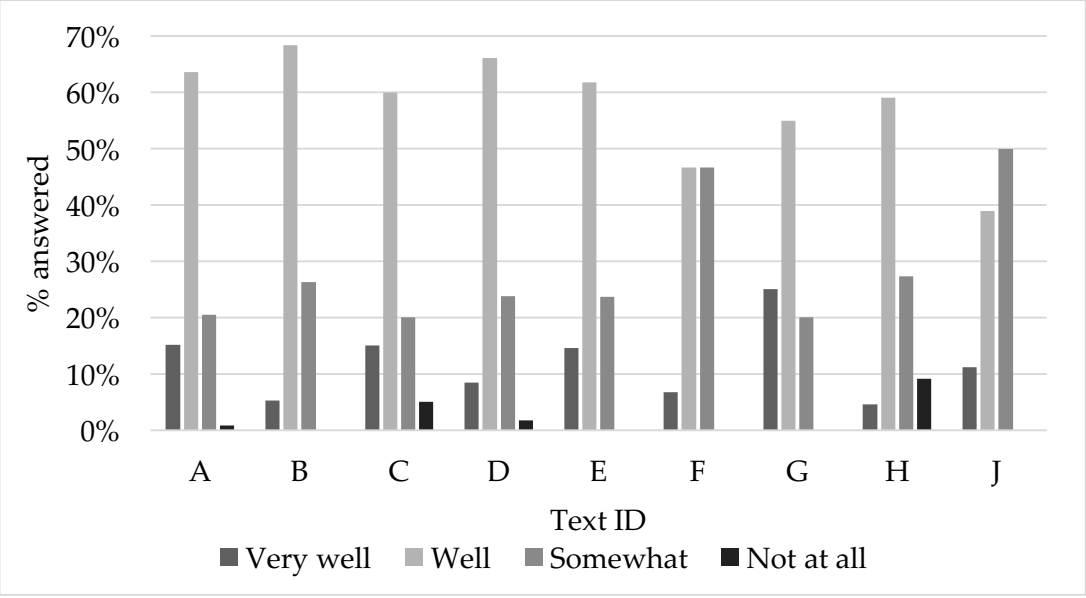


Figure 7.6. L1 readers' subjective understanding on the text

L1 readers rarely answered that they did not understand the text. Most L1 readers rated that they understood the text *Well*; for all but 2 texts over 50% of participants rated their understanding as *Well*. What we do see is that there is a pattern whereas the text gets harder this rating goes down and we see an increase in rating understand as *Somewhat*.

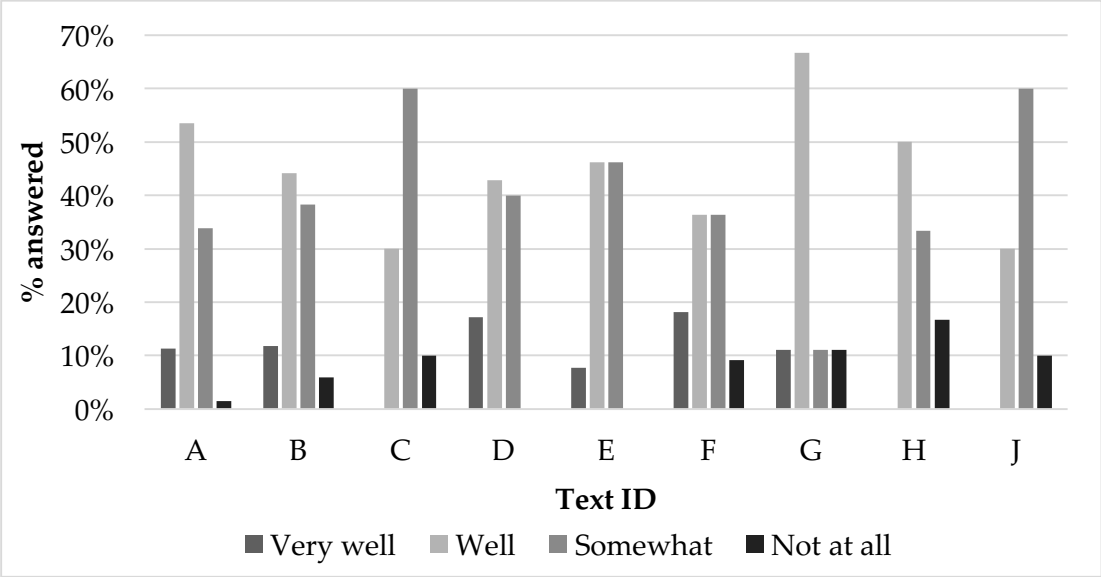


Figure 7.7. L2 readers subjective understanding ratings

L2 readers much more often answered that they did not understand the text well, with 4 texts with 10% or more of participants rating that they did not understand the text. L2 participants are more likely to rate their understanding as *Somewhat* however there are still many answers of *Well*.

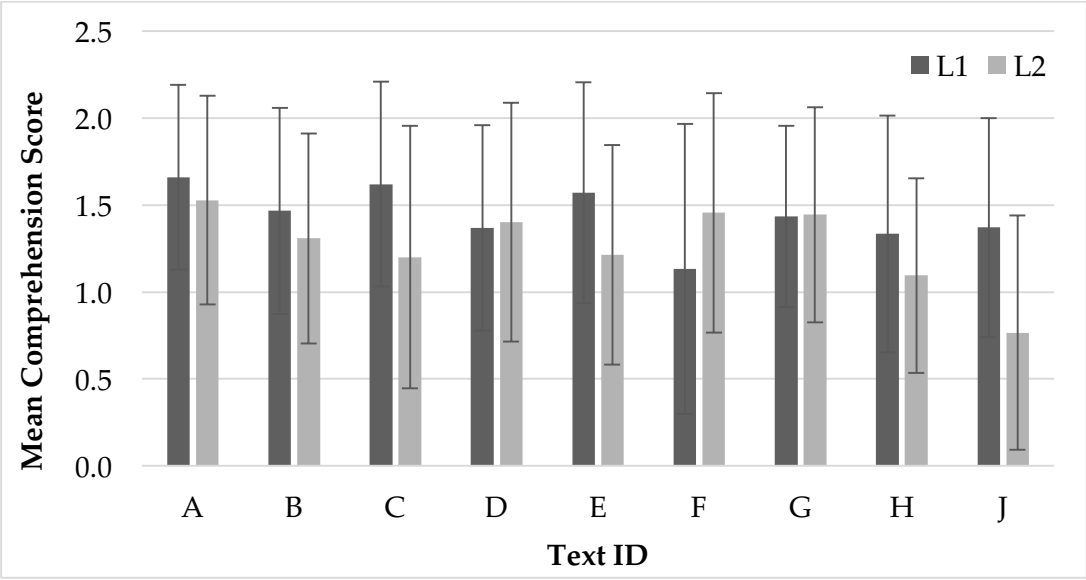


Figure 7.8. Average comprehension score per question

Participants' measured comprehension is shown in Figure 7.8. For the majority of questions, L1 participants received higher scores than the L2 participants. On average, L1 participants scored 1.51 (SD=0.65) on the comprehension test and L2 participants scored 1.36 (SD=0.61). ANOVA analysis shows that whilst the difference is small L2 readers did have statistically significantly lower comprehension scores compared to L1 readers ($F(1,599)=6.56$; $p=0.011$; partial $\eta^2=0.011$). The text difficulty also has a statistically significant effect on the

comprehension scores ($F(8,599)=3.46$; $p=0.001$; partial $\eta^2=0.044$). There is no significant effect of interaction between the reader type and text difficulty.

This result is contrary to what we would expect from past research as well as what we found throughout this thesis, that L1 and L2 participants should perform the same in comprehension tests (Kang, 2014). The results from this investigation indicate that when the degree of difficulty of text is altered then differences in comprehension between L1 and L2 readers occur. L2 participants' perceptions matched the comprehension levels to an extent given that the L2 readers had higher ratings of *Somewhat* understanding.

Tukey's HSD tests were used to perform pairwise comparisons of the texts to further investigate the effects on text difficulty on scores. The pairwise comparison shows that there is a significant difference between texts A and J ($p=0.006$) and there is a weak difference between texts A and D ($p=0.056$). However, the difference between texts A and J is consistent with the findings that text difficulty, in particular readability, affects eye movements, and that the eye movements are somewhat related to comprehension. Both A and J are at either ends of the spectrum of text difficulty. The lack of difference between other texts suggests that there is a spectrum of change with only the ends being statistically significantly different.

7.5.2 Confidence levels

Participants were asked how confident they were with their answers to questions. The ratings are shown in Figure 7.9 for L1 participants and Figure 7.10 for L2 participants. Text difficulty did not have an effect on L2 readers' confidence levels ($\chi(16)=7.85$, $p=0.95$) although it did affect L1 readers' confidence levels ($\chi(16)=30.8$, $p=0.015$).

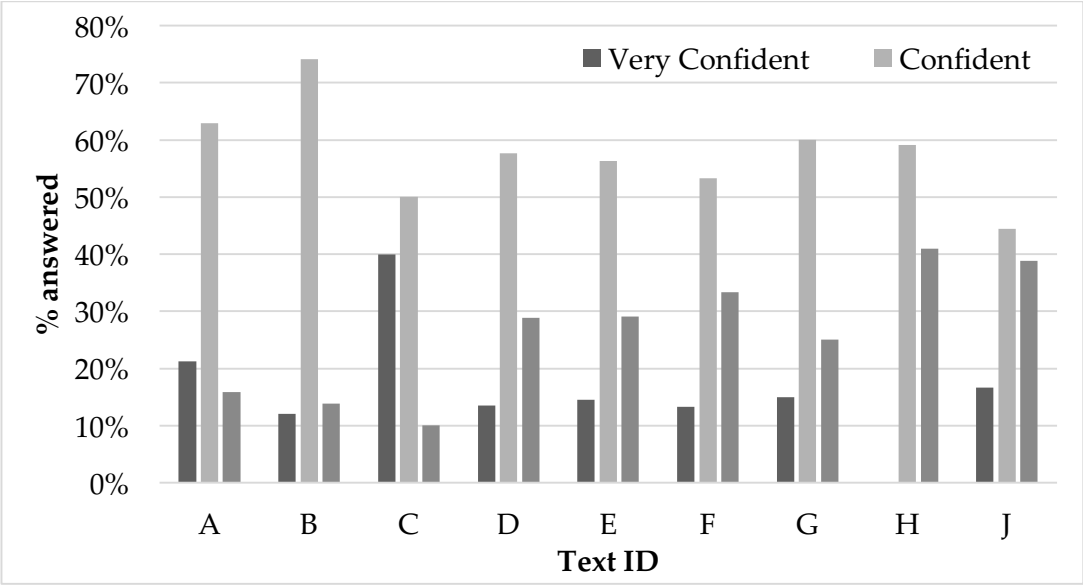


Figure 7.9. L1 participants' confidence ratings

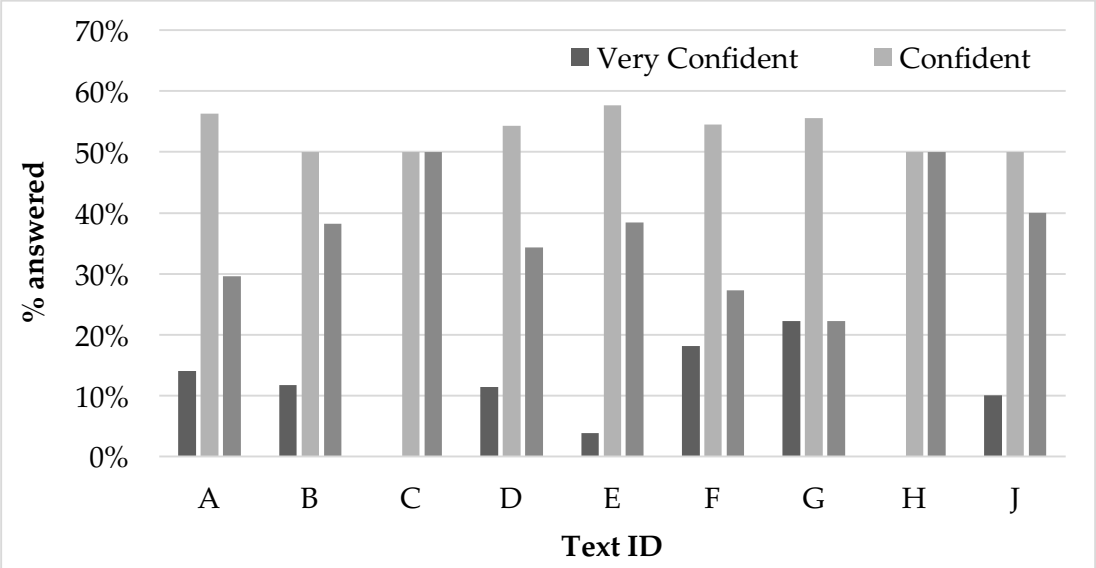


Figure 7.10. L2 participants' confidence ratings

For the L1 participants, whilst we can observe that there is a slight trend of the *Not* confident rating as the text becomes more difficult, the majority of participant state that they are *Confident* in answering the questions no matter what the difficulty is. Therefore, there does not seem to be a clear trend that text difficulty affects confidence ratings of L1 participants.

Whilst again the majority of L2 readers state they are *Confident* similarly to L1 readers, there is a larger number of L2 readers that state they are *Not confident* compared with the L1 readers. Therefore, a larger subset of L2 readers compared to L1 readers had less confidence and thus overrated difficulty, which is consistent with past research (Chang, 2005). Given the lower levels of confidence this accounts for the L2 readers, on average, perceiving the texts as being more difficult than L1 readers.

7.6 Discussion and Implications

In this chapter we analysed participants' perceptions of text difficulty, which showed that that participants' perceptions of text difficulty do not align with the predefined text difficulty. In particular, participants are poor at predicting text difficulty. The GA-kNN predictions, from participants' eye tracking data, of readability and conceptual difficulty were significantly higher that the participants' predictions. However, using the participants' eye tracking data the results from predicting the participants' perceived conceptual difficulty and readability were significantly higher than predictions of the predefined values. This indicates that participants eye tracking data may be also aligned with their perceptions of difficulty rather than just the predefined difficulty.

Further analysis of participants' perceptions of text difficulty shows that the readability and conceptual difficulty of the text interplay to cause deviations of perceptions from the predefined difficulty. For both L1 and L2 readers, the perceptions of text difficulty are worst when the readability and the conceptual

levels do not match. This is observed most obviously when the readability is easy and the concept is advanced (G) or the readability difficult and the concept is basic (C). Participants seem unable to distinguish the properties in these cases. There is an interaction that masks the two variables resulting in a rating somewhere in the middle. The result is that L2 readers mostly rated texts as moderate in readability and intermediate in conceptual difficulty (E), and L1 readers mostly rated texts as easy, with both readability and conceptual difficulty (A). For L2 readers, this could be due to participants not really knowing what the levels of difficulty are and therefore extrapolating or approximating as being somewhere in the middle. This highlights the second hypothesis of the study that L2 participants will have an inflated perception of text difficulty compared to the L1 participants. The results from the study support this hypothesis but also the results show that in general L1 readers underestimate text difficulty. Finally, we hypothesised that the eye gaze and pupil dilation data would be better at predicting the text complexity than the participants' perceptions. The analysis did not provide evidence that this was the case.

The second part of this chapter investigated the effect that text difficulty has on comprehension and confidence. Past research has shown that whilst L1 and L2 readers have different eye movements during reading, they have the same comprehension levels. However, this only dealt with text at a constant level of difficulty. The results from the study indicate that L2 readers have significantly lower comprehension scores to L1 readers. Moreover, text difficulty was found to have a significant effect on comprehension, but the difference is only between the very easiest and the very hardest of the texts, A and J. There is no other clear effect that text difficulty affects comprehension. This indicates that the comprehension scores from the texts are not entirely reliable indicators of difficulty. Given that our analysis showed that predictions for participants' perceptions of readability and conceptual difficulty from eye tracking data are higher than the predictions the predefined levels, this introduces the idea that eye tracking data could be used as an indicator of difficulty.

Finally, we hypothesised that harder texts would be associated with lower ratings of confidence and perceived understanding. This is because task complexity is known to affect perceptions of difficulty as well as confidence levels (Robinson, 2007). The results show that while our hypothesis was incorrect about subjective understanding ratings, there is a difference between L1 and L2 readers, where L2 readers express lower subjective understanding than L1 readers. L2 readers are also less confident than L1 readers, no matter what text is read and hence text difficulty has no significant effect on L2 readers. This is an interesting finding and could be due to skill level of the readers, so that the L2 readers are too challenged by all of the texts and therefore no effect can be seen. However, this is not true for L1 readers where text difficulty was found to significantly affect confidence, where the harder the text becomes, the less confidence readers have.

7.7 Conclusion and Further Work

In this chapter we investigated how eye tracking data and participants' perceptions compare at predicting text difficulty. We take into consideration both the readability and the conceptual difficulty of the text and assess how L1 and L2 readers differ. Participants were poor at predicting text difficulty however we found that using GA-kNN to predict readability and conceptual difficulty, from their eye gaze and pupil dilation data, is significantly more accurate. However, prediction of participants' perceived ratings of readability and conceptual difficulty from the eye tracking data are significantly better than prediction of the predefined values. This indicates that the eye gaze measures and pupil dilation data may be more aligned with the participants' perceptions of difficulty rather than the predefined difficulty of the text.

Whilst comprehension scores were found to be effected by the text difficulty. the effect is minimal, where the only significant difference is the scores for the easiest (A) compared to the hardest (J). Combining both findings indicates that comprehension score alone is not a sufficient indicator of text difficulty but that eye tracking data could be used in combination to determine the overall difficulty.

Finally, L1 readers scored higher on comprehension questions compared to L2 readers, and text difficulty did not affect L2's confidence in answering the questions, highlighting that there are significant differences in perceptions of L1 and L2 readers and not just their reading behaviour. These difference need to be considered when designing texts for education.

Further research into the use of physiological signals could reveal more accurate predictions of text difficulty. In particular, using cognitive load as a measure of text difficulty could provide more accurate text difficulty predictions from eye gaze and pupil dilation data. Cognitive load has been successfully predicted from both eye gaze and pupil dilation data (Rosch & Vogel-Walcutt, 2013), which is useful because cognitive load has been used to predict task difficulty (Waniek & Ewald, 2008). Furthermore, looking into the use of physiological signals such as electrocardiogram (ECG), galvanic skin response (GSR), electroencephalogram (EEG) could prove to be useful in predicting text difficulty as these signals have been used prediction of stress whilst reading documents of difficult degrees of difficulty and stressfulness (Sharma & Gedeon, 2012, 2013a, 2013b).

The results from this chapter indicate that the ranking of text difficultly might be insufficient. We propose the use of eye tracking data to classify texts according to complexity measures that reflect students' perceived difficulty of the text. We will explore this further in the next chapter.

Chapter 8

Deriving text difficulty from eye gaze

The eye gaze data from the user study in Chapter 6 was used to investigate the differences between L1 and L2 readers' reading behaviours as well as whether eye gaze measures can be used to derive text difficulty. The investigation involves clustering eye movement measures from participants using kmeans clustering. The results indicate that whilst there are clusters of different reading behaviours for different levels of text difficulty, such as skimming and thorough reading, the L1 and L2 groups were not found to be distinct from each other. Instead, there is a tendency for L2 readers to exhibit more thorough reading compared to skimming. The previous chapters established that eye movements are related to both the predefined and readers' perceptions of the text difficulty and that the readability and conceptual difficulty interplay to cause deviations from expected text difficulty. This raises the question of whether the ratings of text difficulty are adequate for defining the actual difficulty of the text. The average eye gaze measures for each text were clustered using k-means. The clusters show that there are distinct reading behaviours and that the average eye gaze measures can be used to rate the texts based on the derived reading difficulty for the L1 and L2 groups. These findings can be used to provide feedback to the author for the purpose of adapting learning material. As in previous chapters, this feedback will be in two forms; first on an individual basis to provide feedback regarding reading and thereby aid personalised learning, and secondly, on a cohort basis to provide feedback about reading difficulty of particular texts.

8.1 Introduction

Individual students have different prior knowledge and expertise as well as different levels of reading abilities. From the Chapter 7 that we can see that an individual's perception of text difficulty is likely to be affected by several factors, such as reading skill, prior knowledge, motivation, and arousal or interesting in a given topic. Alternatively, the definition of text difficulty may not be flexible enough to deal with the differences between the L1 and L2 groups as well as within those groups. In this case the problem becomes how to determine a robust method of determining text difficulty. One method is to ask students how difficult text is for them to read. However, this method does not support real time changes, is disruptive to the learner, and people are poor at perceiving their abilities (Kruger & Dunning, 1999). Inexperienced people are unaware of their lack of expertise resulting in them overrating their abilities in comparison to a cohort, whereas accomplished people tend to know their shortcomings and underrate their abilities in comparison to the cohort (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999). In Chapter 3 we found that this can be somewhat mitigated by presentation method, consequently students' perceptions of difficulty cannot be relied upon to gain insight into the level of difficulty of learning material. The use of physical and physiological data can be used to predict cognitive load (Rosch & Vogel-Walcutt, 2013) which in turn can be used to dynamically change an eLearning environment in real time (Coyne et al., 2009). In particular, eye tracking has been used to measure cognitive load during reading, where longer reading times indicate greater cognitive load (Rosch & Vogel-Walcutt, 2013), which is consistent with the finding that eye movements reflect comprehension processes (Rayner et al., 2006).

The main goal of this chapter is to investigate ways of providing feedback regarding text difficulty based upon eye gaze data, similarly to how answer-seeking behaviour was used to provide feedback in Chapter 4. This chapter's approach is similar in that eye gaze measures will be clustered and analysed to provide feedback about how individual students read as well as how groups of students read certain texts, in order to provide a measure of difficulty derived from reading behaviour. This follows on from the results from Chapters 6 and 7, which raised the important question of whether the way in which text difficulty has been defined in this thesis is in fact suitable. Therefore, the research question of the chapter is:

Can eye gaze data be used to differentiate between L1 and L2 readers and to determine derived difficulty of text?

The importance of differentiating between L1 and L2 readers comes from the results from Chapter 7 that the two groups have different perceptions of text difficulty and therefore different measures of text difficulty. Additionally, the two groups are known to have different eye gaze behaviours during reading. Given that we propose using eye gaze to measure derived difficulty, it is imperative that the two are differentiated before calculating measures of derived difficulty. This also

provides us with the opportunity to further investigate the differences in eye gaze and reading behaviour between L1 and L2 readers.

The potential of measuring derived text difficulty is that in an adaptive environment the text could be tailored to students' needs and reading behaviours. It is not just important to identify that participants understand text but it is also crucial to know the level of difficulty at which they are either understanding, or not understanding, so that the content can be changed accordingly. This is touched upon in Chapter 4 where we propose the measure of answer-seeking behaviour to identify how difficult text and the related comprehension questions are, and where a ranking of how hard the participant found the questions is provided. The work in this chapter differs in that we attempt to measure text difficulty without using questions or asking students to state how difficult they found the text using a rating system. We hypothesise that for simpler texts there will be a spectrum of eye movements where L1 and L2 readers are not easy to differentiate. However, as the text gets more difficult, clusters of the L1 and L2 readers' eye movements will become more distinct. Additionally, we hypothesise that each text will induce different average eye movement measures that can be used to find the average reading behaviour of readers of that text to then use as a derived measure of text difficulty.

This chapter does not include a background section as the literature has been covered in previous chapters. The rest of the chapter is organised to firstly cover the analysis of differentiating L1 and L2 readers; then to investigate how clustering of eye movement data can be used to provide a measure of text complexity; finally these result will be discussed in relation to their implications for adaptive eLearning.

8.2 Method

The eye tracking data used in this analysis was recorded in the user study conducted in Chapter 6; refer to section 6.3 for further details. The analysis in this chapter is primarily through k-means clustering of the data using Matlab R2016a. The first part of the investigation looks at clustering the eye tracking data recorded from three texts, A, E, and J, to see if there are natural clustering between L1 and L2 readers and thus distinct reading behaviours. The second part of the analysis looks at the use of eye movement data in rating the texts on derived difficulty.

We used the silhouette method to evaluate the quality of the clusters. Using Matlab's *evalclusters()* function we found the optimal number of clusters for the given data set. We then used the optimal number to cluster the data using k-means clustering. After this the average silhouette width for the total data set was calculated and reported. The average silhouette width provides an evaluation of the clusters to support the choice in number of clusters, where the close to 1 the average silhouette width, the stronger the clustering structure should be (Rousseeuw, 1987).

8.2.1 Eye movement measures

The eye gaze measures that we analyse in this section are the same as used throughout this thesis: normalised number of fixations (NNF), maximum fixation duration (MFD), average fixation duration (AFD), normalised total fixation duration (NTFD), regression ratio, and average forward saccade length (AFSL). Refer to Chapter 6 for more details on these measures.

8.3 Differentiating L1 and L2 readers

Throughout the thesis there has been an assumption that L1 and L2 readers are distinct. However, many of the analyses have shown that a difference does not exist. An example of this is that there is no difference in answer-seeking behaviour between the L1 and L2 groups. In other cases, differences exist and are statistically significant; however, just because the groups are statistically different does not imply that they do not have some overlap. That is, there may be some L2 readers that are similar to L1 readers and some L1 readers that are similar to L2 readers.

In this section, cluster analysis is used to investigate if there are distinct clusters of eye movements between the L1 and L2 readers. Clustering of the eye movement data from three texts A, E, and J is performed. We hypothesise that for the simplest text, A, instead of having distinct clusters, a spectrum of eye movements will be observed. However, as the text gets more difficult, as in texts E and J, the clusters will become more distinct as the differences between the L1 and L2 participants grow.

8.3.1 Easy Text (A)

The eye gaze data recorded from reading text A is clustered in this section. The optimal number of clusters was determined to be 2, see Table 8.1, where the largest average silhouette width is 0.790, which is for 2 clusters.

Table 8.1. Average silhouette widths for clustering of A

Number of clusters	Ave. silhouette width
2	0.790
3	0.732
4	0.741
5	0.735
6	0.691
7	0.680
8	0.695
9	0.697
10	0.695

To assess whether the clustering has separated the L1 and L2 readers, we will examine the contents of the clusters. The average measures for the two clusters are shown in Table 8.2. What we observe is that there is not a clear distinction between the L1 and L2 readers' data points, for the simplest text, A. However, in saying this, the majority of the L2 data points 85% (61 of 72 points) are in cluster 1. This cluster appears to be a clustering of what we can consider as more thorough reading compared to cluster 2. This can be concluded from the higher NNF, longer MDF, AFD, and NTFD, and shorter forward saccades in cluster 1 compared to cluster 2.

Table 8.2. Eye movement averages from clusters for text A

Measure	Cluster 1	Cluster 2
Number of L1 readers	86 (59%)	48 (81%)
Number of L2 readers	61 (41%)	11 (19%)
Total number in cluster	147	59
Mean normalised number of fixations (NNF)	0.81	0.46
Mean maximum fixation duration (MFD)	1.63 s	0.97 s
Mean average fixation duration (AFD)	0.24 s	0.16 s
Mean normalised total fixation duration (NTFD)	0.2	0.08
Mean regression ratio	0.33	0.45
Mean average forward saccade length	102.08	165.7
Mean comprehension score	1.59	1.69
Mean readability	1.5	1.39
Mean conceptual difficulty	1.45	1.39

Note that the points within each clusters are note individual participants but the texts that the participants read. So whilst there are 70 participants, each participant read 3 versions of text A totalling 210 texts read, however due to removal of corrupt data 4 of these were removed totalling 206 texts analysed in this section.

There is generally an uneven distribution of fixations on words whilst reading English (Rayner, 1998). The NNF values for normal reading are therefore expected to be less than 1. In fact, Carpenter and Just (1983) found that readers fixate on an average of 67.8% words. The closer the NNF is to 0, the more this indicates skimming or scanning of the text, whereas values above 1 correspond to more fixations than there are words in the paragraph which is indicative of re-reading of some of the text. The mean NNF for cluster 1 is 0.81 which is above the expected value just stated. The mean NNF for cluster 2, however, is considerably lower than the expected value being on 0.49. This would indicate that cluster 1 is a clustering of reading behaviour that is above average reading behaviour and cluster 2 is a clustering of reading behaviour that is well below the average reading behaviour.

The majority of the L2 data points lie within cluster 1, and the L2 data points make up almost half of the cluster. We conclude that the majority of L2 participants were reading above the average reading behaviour. However, the majority of the L1 data points are also in this cluster, so the behaviour of above average reading is not unique to the L2 readers. Therefore, we can also conclude the L1 readers are more likely to skim, or have well below average reading behaviour, compared to the L2 readers, given that 81% of the data points in cluster 2 come from L1 readers.

Finally, the clustering did not find any distinct differences between the participant's ratings of readability and conceptual difficulty, which are all, on average, between easy/basic and moderate/intermediate. There are no differences in the comprehension scores between the clusters either ($t(204)$=-1.18, p=0.238). This indicates that the main difference between the clusters is the reading behaviour, which can be described as thorough reading, for cluster 1, and skimming for cluster 2. Moreover, neither type of reading behaviour is characteristic of reading groups, however it is more likely that L1 readers will skim compared to L2 readers.

8.3.2 Moderate Text (E)

We now cluster the eye movement measures from text E. The optimal number of clusters was determined to be 3, see Table 8.3, where the largest average silhouette width is 0.799, which is for 3 clusters.

Table 8.3. Average silhouette widths for clustering of E

Number of clusters	Ave. silhouette width
2	0.779
3	0.799
4	0.773
5	0.724
6	0.752
7	0.650
8	0.691
9	0.696
10	0.718

The contents of each cluster are shown in Table 8.4, which indicate that there is no specific differentiation between the L1 and L2 data points. However, in saying this, the differences between the L1 and L2 data points has grown. As for text A, cluster 1 is comprised almost equally by L1 and L2 data points. In cluster 2 we can see that the distribution of L1 and L2 points is now skewed towards L1 points, as with text A. Finally, in the extra cluster, 3, only L1 points comprise this cluster. This indicates that whilst there is not a clear distinction between the L1 and L2 data points, there are some differences in their eye movements, and this distinction becomes more prominent when the text increases in difficulty.

Looking deeper into the reading behaviour represented in the clusters, we once again see that cluster 1 is representative of thorough reading, and most L2 readers are part of this cluster, as for text A. The mean NNF for this cluster is 0.91, which means that on average 91% of words were fixated, which is above the standard fixation rate. Combined with increased fixation durations and smaller forward saccade lengths, this is indicative of increased text difficulty, which is observed for text E. Cluster 2 has reduced reading compared to cluster 1, with values that are indicative of average reading behaviour. Finally, cluster 3 is similar to cluster 2 for

text A, where we can observe skimming behaviour. This cluster is made up solely of L1 data points. This is consistent with what we found for text A, where L1 readers are more likely to be the readers to skim.

Table 8.4. Eye movement averages from clusters for text E

Measure	Cluster 1	Cluster 2	Cluster 3
Number of L1 readers	29 (58%)	21 (75%)	5 (100%)
Number of L2 readers	21 (42%)	7 (25%)	0 (0%)
Total number in cluster	50	28	5
Mean normalised number of fixations (NNF)	0.92	0.71	0.34
Mean maximum fixation duration (MFD)	1.78	1.25	0.73
Mean average fixation duration (AFD)	0.26	0.20	0.15
Mean normalised total fixation duration (NTFD)	0.25	0.14	0.06
Mean regression ratio	0.32	0.40	0.47
Mean average forward saccade length	95.4	140.3	201.2
Mean comprehension score	1.41	1.5	1.6
Mean readability	1.76	1.64	2
Mean conceptual difficulty	1.92	1.71	2

Again there is little difference is the participants' ratings of readability and conceptual difficulty, which are still between easy/basic and moderate/intermediate. ANOVA of the comprehension scores shows that there is no significant difference between the clusters ($F(2,204)=1.4$, $p=0.238$), even though there seems to be a slight trend is the comprehensions being higher as the clusters go up.

8.3.3 Difficult Text (J)

Finally, we cluster of the eye movement measures recorded from reading text J. The optimal number of clusters was determined to be 4, see Table 8.5, where the largest average silhouette width is 0.825, which is for 4 clusters.

The contents of each cluster are shown in Table 8.6. Whilst there are 4 clusters for this text, one of the clusters contains only 1 data point, which is an outlier for the data set. There is, once again, no clear distinction between the L1 and L2 data points, as for the previous texts. In fact, there is an almost even distribution of L2 data points between the 3 other clusters. That is, unlike in the previous sections, there is no cluster (other than cluster 4) that contains only or close to only L1 data points. This indicates that the harder the text gets, the harder it is to differentiate L1 and L2 readers. Perhaps the text becomes too difficult and as a result the L2 readers are unable to cope with the task.

As with the previous texts, the clusters show different reading behaviours that range from thorough reading (cluster 1) to skimming (Clusters 3 and 4). Moving from text A to E there is an increase in the NNF and fixations durations, as well as a

decrease in forward saccade length, all indicative of more reading. We see that cluster 1, which is representative of the most thorough reading, has an average NNF with a much higher than expected, indicated that words are fixated on more than once. This is the highest average NNF that we observe and since J is the most difficult text to read we would expect more thorough reading compared the A or E.

Table 8.5. Average silhouette widths for clustering of J

Number of clusters	Ave. silhouette width
2	0.776
3	0.757
4	0.825
5	0.822
6	0.720
7	0.769
8	0.683
9	0.747
10	0.768

Table 8.6. Eye movement averages from clusters for text J

Measure	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of L1 readers	7 (58%)	7 (70%)	3 (60%)	1 (100%)
Number of L2 readers	5 (42%)	3 (30%)	2 (40%)	0 (0%)
Total number in cluster	12	10	5	1
Mean normalised number of fixations (NNF)	1.10	0.94	0.63	0.46
Mean maximum fixation duration (MFD)	2.22 s	1.13 s	1.05 s	0.58 s
Mean average fixation duration (AFD)	0.32 s	0.20 s	0.16 s	0.15 s
Mean normalised total fixation duration (NTFD)	0.33 s	0.19 s	0.1 s	0.07 s
Mean regression ratio	0.29	0.39	0.42	0.53
Mean average forward saccade length	86.7	114.5	148.5	213.06
Mean comprehension score	1.0	1.2	1.3	2
Mean readability	1.8	1.8	2.2	2
Mean conceptual difficulty	2.3	2.3	2	2

The remaining clusters represent gradually less reading behaviour, where cluster 2 is similar to cluster 2 from text E, which is still indicative of thorough reading. Cluster 3 is the closest of normal reading behaviour with an average NNF

of 63 (compared the stated average of 67.8%). However, given that this is the most difficult text to read we would not expect to have many people reading it as a *normal* text. Finally, the outlier point indicates skimming behaviour. Given that this data point has a high comprehension score (2 out of 2), the skimming behaviour could be due to the fact that the reader had a high level of prior knowledge in the area and therefore did not need to read the text thoroughly.

For the remaining clusters there appears to be little difference in the average comprehension scores, which are on average quite low, and little difference between the other subjective ratings. ANOVA of the comprehension scores, for all of the clusters, shows that there is no significant difference between the clusters ($F(3,27)=0.94$, $p=0.438$), even though there seems to be a slight trend is the comprehensions being higher as the clusters go up. Again, there appears to be little difference between the ratings of readability and conceptual difficulty between the clusters.

8.4 Deriving text difficulty from eye gaze

From the first part of the analysis in this chapter, we observe that there are differences in reading behaviours for each text. It is clear from the previous chapter that participants are poor at identifying the predefined difficulty of the text. The results indicate that this could be due to the fact that *difficulty* is different for everyone, and therefore everyone has different perceptions of difficulty. This is evident given that students have different prior knowledge and expertise, as well as different levels of reading abilities.

Given that eye movements are not a complete reflection of perceptions or predefined text difficulty; we now question whether the predefined definition of text difficulty is suitable for adaptive eLearning? That is, are these definitions flexible enough to deal with the differences between not only the L1 and L2 groups, but also the differences within these groups? Since the eye movements are affected by both perceived and predefined text difficulty, but not completely governed by either, this suggests that eye movements are reflections of the *derived* difficulty of the text. We suggest that each text will have different average eye movement measures. These can be used to find the average reading behaviour of that text and then use as a measure of text difficulty.

As described in Chapter 6, there are 27 texts used in total for the study; from 3 topics, and each topic containing 9 versions of text, based upon the grid system, labelled A through J. For more information, refer to the Method (section 6.3) in Chapter 6. For each of the 27 texts the average eye movement measures, as described in section 8.2.1, were calculated for the L1 and L2 groups. These measures were clustered using k-means clustering for the L1 and L2 groups separately.

8.4.1 L1 derived text difficulty

The optimal number of clusters for the L1 text averages was determined to be 4, see Table 8.7, where the largest average silhouette width is 0.787, which is for 4 clusters.

The average eye movement measures and outcome measures for each cluster are shown in Table 8.8. There is a spectrum of mean eye movement measures across the clusters indicating that there are different average reading behaviours observed for different texts. Starting with cluster 1, this cluster has the fewest texts within it, but is also the cluster that is associated with thorough reading. For the L1 readers, we can see that they did not find a lot of the texts difficult to read as the majority of the texts are associated average reading behaviours.

Table 8.7. Average silhouette widths for clustering of average eye movement measures for each text

Number of clusters	Ave. silhouette width
2	0.670
3	0.768
4	0.787
5	0.724
6	0.697
7	0.643
8	0.628
9	0.694
10	0.754

Table 8.8. Averages of measures for each clusters for L1 readers, based on text averages

Measures	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Total number in cluster	3	10	7	7
Mean normalised number of fixations (NNF)	0.95	0.77	0.75	0.65
Mean maximum fixation duration (MFD)	1.34	1.25	1.35	1.08
Mean average fixation duration (AFD)	0.22	0.20	0.21	0.18
Mean normalised total fixation duration (NTFD)	0.22	0.17	0.16	0.12
Mean regression ratio	0.33	0.37	0.37	0.39
Mean average forward saccade length	105.7	126.1	116.4	138.0
Mean comprehension score	1.1	1.3	1.6	1.7
Mean perceived readability	1.6	1.6	1.5	1.5
Mean perceived conceptual difficulty	2.2	1.8	1.6	1.5

There does not appear to be a large difference between clusters 2 and 3, which are indicative of reading behaviour that is slightly above the average. Cluster 2 has slightly short fixation durations, but more fixations and therefore longer total

fixation duration, and longer forward saccades, compared to cluster 3. This would seem to indicate that the main difference between the two clusters is that the texts in Cluster 3 have more concentrated reading compared to cluster 2, which is why we see longer fixations and shorted forward saccades. Cluster 2 has the most texts within it and given the nature of the eye movements this is most likely the *normal* reading behaviour of participants for this set of texts.

Finally, cluster 4 contains 7 texts which have reading behaviour that is diminished compared to the other 3 clusters. Whilst we cannot describe the average reading behaviour as explicitly skimming, the reading behaviour is below the expected level. Therefore, the texts in this cluster are easier to read.

We use MANOVA to determine if there are any statistical differences between the clusters. The correlations between the dependent variables are within the acceptable limits for MANOVA outcomes, i.e. the correlations lie between r=-0.4 and r=0.9. To test for normality in the dependent variables the Shapiro-Wilk Test is used, as it is more appropriate for small sample sizes. All variables are normally distributed (p>0.05). Levene's test for equality of variances shows that there is homogeneity for all dependent variables (p>0.05) Finally, the homogeneity of variance-variance-covariance matrices is satisfied as the Box's M value of 111.46 (p=0.038>0.001).

There is a statistically significant difference in average eye movement measures between the clusters, $F(18,51)=7.42$, $p<0.0005$; Wilk's $\lambda=0.530$, partial $\eta^2=0.701$. ANOVA shows that the clusters have a statistically significant effect on all measures, NNF ($F(3,26)=3.81$; $p=0.024$; partial $\eta^2=0.332$), MFD ($F(3,26)=4.55$; $p=0.012$; partial $\eta^2=0.372$), AFD ($F(3,26)=9.75$; $p<0.0005$; partial $\eta^2=0.560$), NTFD ($F(3,26)=7.33$; $p=0.001$; partial $\eta^2=0.489$), regression ratio ($F(3,26)=7.63$; $p=0.001$; partial $\eta^2=0.499$), and AFSL ($F(3,26)=129.7$; $p<0.0005$; partial $\eta^2=0.944$). This is not surprising given that the eye movement measures were used to create the clusters. However, this does indicate that we can use this eye movement measures to rank these texts into distinctive groups based on reading behaviour.

Perhaps more informative is an analysis of how the clusters are related to predefined readability and conceptual difficulty as well as the resulting measures of comprehension and perceived readability and conceptual difficulty. Firstly, Chi-square test for independence shows that there is no evidence of relationship between clusters and predefined readability ($\chi^2(6)=3.685$, $p=0.719$) and predefined conceptual difficulty ($\chi^2(6)=7.371$, $p=0.287$). The clusters, and therefore reading behaviours, are not related to the predefined readability or conceptual difficulty. This is what we hypothesised, and expect based on previous analysis in this thesis.

Considering now the comprehension scores and perceived readability and conceptual difficulty MANOVA is used. The correlations between the dependent variables are within the acceptable limits for MANOVA outcomes, i.e. the correlations lie between r=-0.4 and r=0.9. To test for normality in the dependent variables the Shapiro-Wilk Test is used, and all variables are normally distributed (p>0.05). There is a statistically significant difference in average resulting measures

for each text between the clusters, $F(9,51)=2.20$, $p=0.037$; Wilk's $\lambda=0.530$, partial $\eta^2=0.233$. Interestingly, the differences lie in the comprehension scores ($F(3,23)=3.556$; $p=0.03$; partial $\eta^2=0.317$) and the perceived conceptual complexity ($F(3,23)=4.01$; $p=0.012$; partial $\eta^2=0.373$), but not on the perceived readability ($F(3,23)=0.336$; $p=0.\ 799$; partial $\eta^2=0.042$). Whilst the perceived conceptual complexity appears to be associated with the clustering of eye movements, there is no relationship to the predefined levels of complexity.

Table 8.9. Texts within each cluster, for L1 averages for text

Cluster	Characteristic reading behaviour	Texts in cluster
1	Thorough	T1-C, T3-F, T1-J
2	Average	T1-A, T2-A, T3-A, T1-E, T1-F, T2-G, T1-H, T2-H, T3-H, T2-J
3	Average, more concentrated	T2-B, T2-C, T1-D, T2-D, T3-D, T3-E, T3-J
4	Below average	T1-B, T3-B, T3-C, T2-E, T2-F, T1-G, T3-G

NOTE: T1 refers to topic 1, T2 refer to topic 2, and T3 refers to topic 3.

Table 8.9 shows the texts that are within each cluster. The clusters give a measure for the average reading behaviour observed for the text and can be used as feedback to the author or designer of eLearning material to obtain the derived difficulty of the text. That is, texts with low levels of reading are simpler to read, also have less perceived conceptual difficulty, and therefore less thorough reading is observed. We can see that for cluster 1, the text associated with the most thorough reading behaviour on average, the all of these texts have the highest concept difficulty. Yet these texts are only a subset of all texts with the same level of conceptual difficulty, and these texts all have different levels of readability. In this way, the clustering may be surprising to the author as the reading behaviours for the texts are not associated in the ways we would expect to the predefined readability and conceptual difficulty. Since it has been shown that as text becomes more difficult to read, eye movements are seen to reflect the difficulty. This implies that the predefined difficulties are not entirely associated with the reading difficulty.

8.4.2 L2 derived text difficulty

We now consider the L2 averages for each text. The optimal number of clusters for the L2 text averages was determined to be 2, see Table 8.10, where the largest average silhouette width is 0.991, which is for 2 clusters. However, when we inspect the clusters further this clustering results in an outlier text in its own cluster and the rest of the texts clustered together. For this reason, we move to using 3 clusters to describe the texts, as the average silhouette width for 3 clusters is 0.802, which is indeed higher than the average silhouette width for the optimal number of clusters for the L1 text averages.

Table 8.10. Average silhouette widths for clustering of average eye movement measures for each text

Number of clusters	Ave. silhouette width
2	0.991
3	0.802
4	0.621
5	0.613
6	0.737
7	0.726
8	0.687
9	0.638
10	0.517

Table 8.11 shows the average eye movement measures for the texts within the 3 clusters. Examining the contents of the clusters for the L2 averages for the texts shows that the outlying text in a cluster of its own is a text that on average the eye movements that signify skimming behaviour. That is, L2 participants only seemed to skim one text, rather than the 7 texts that the L1 participants are seen to skim. The text that the L2 participants skim is unexpected; instead of being a text with easy readability and easy conceptual difficulty (e.g. text A) it is text H (from Topic 3), which has difficult readability and intermediate conceptual difficulty, therefore being one of the most difficult tasks to read. This text also does not correspond with the texts that the L1 readers had below average reading behaviour for.

Table 8.11. Averages of measures for each clusters for L1 readers, based on text averages

	Cluster 1	Cluster 2	Cluster 3
Total number in cluster	1	9	17
Mean normalised number of fixations (NNF)	0.60	0.85	0.91
Mean maximum fixation duration (MFD)	0.87	1.55	1.92
Mean average fixation duration (AFD)	0.17	0.24	0.29
Mean normalised total fixation duration (NTFD)	0.10	0.22	0.28
Mean regression ratio	0.44	0.36	0.33
Mean average forward saccade length	212.7	112.1	99.3
Mean comprehension score	1.0	1.2	1.3
Mean perceived readability	0.0	2.3	2.5
Mean perceived complexity	1.0	2.2	2.3
Mean COH-Metrix L2 readability	11.3	7.6	9.0

The remaining two clusters show that L2 participants do indeed have more thorough reading behaviour, on average, compared to the L1 participants. We see that cluster 3 has the most texts within it and yet this cluster has similar eye movement averages to the thorough reading cluster for the L1 participants. Cluster

2 has the remaining 9 texts within it with the average eye movement measures being less than those in cluster 3 but still above the normal level we would expect.

Table 8.12. Texts within each cluster, for L1 averages for text

Cluster	Characteristic reading behaviour	Texts in cluster
1	Outlier – below average	T3-H
2	Average/thorough	T1-A, T1-B, T2-B, T3-B, T3-A, T1-E, T1-G, T1-H, T2-F, T2-J
3	Thorough	T2-A, T1-C, T2-C, T3-C, T1-D, T2-D, T3-D, T2-E, T3-E, T3-F, T1-F, T2-G, T3-G, T2-H, T1-J, T3-J

NOTE: T1 refers to topic 1, T2 refer to topic 2, and T3 refers to topic 3.

As above, we use MANOVA to determine if there are any statistical differences between the clusters. The correlations between the dependent variables are within the acceptable limits for MANOVA outcomes, i.e. the correlations lie between r=-0.4 and r=0.9. To test for normality in the dependent variables the Shapiro-Wilk Test is used, as it is more appropriate for small sample sizes. All variables are normally distributed ($p>0.05$). Levene's test for equality of variances shows that there is homogeneity for all dependent variables ($p>0.05$). Finally, the homogeneity of variance-variance-covariance matrices is satisfied as the Box's M value of 43.5 ($p=0.114>0.001$).

There is a statistically significant difference in average eye movement measures between the clusters, $F(12,38)=21.208$, $p<0.0005$; Wilk's $\lambda=0.017$, partial $\eta^2=0.870$. ANOVA shows that the clusters have a statistically significant effect on all measures except NNF; NNF ($F(2,26)=1.795$; $p=0.188$; partial $\eta^2=0.130$), MFD ($F(2,26)=6.07$; $p=0.007$; partial $\eta^2=0.336$), AFD ($F(2,26)=4.43$; $p=0.023$; partial $\eta^2=0.270$), NTFD ($F(2,26)=3.66$; $p=0.041$; partial $\eta^2=0.234$), regression ratio ($F(2,26)=13.64$; $p<0.0005$; partial $\eta^2=0.532$), and AFSL ($F(2,26)=403.6$; $p<0.0005$; partial $\eta^2=0.971$). This time the clusters do not completely differ statistically, as compared to clustering from the L1 participants eye movements. This indicates that the NNF is not a good measure for predicting reading difficulty as the values must not vary enough between the texts. However, as above, it is not surprising that there are significant differences between the clusters given that the eye movement measures were used to create the clusters. This does show that we can use measures such as fixation duration and forward saccade length to classify the texts based on their derived difficulty.

Analysis of how the clusters are related to predefined readability and conceptual difficulty as well as the resulting measures of comprehension and perceived readability and conceptual difficulty. Firstly, Chi-square test for independence shows that there is no evidence of relationship between clusters and predefined readability ($\chi^2(4)=3.13$, $p=0.535$) and predefined conceptual difficulty ($\chi^2(4)=5.88$, $p=0.208$). As with the L1 averages, the clusters, and therefore reading behaviours, are not related to the predefined readability or conceptual difficulty.

Considering now the comprehension scores and perceived readability and conceptual difficulty MANOVA is used. The correlations between the dependent variables are within the acceptable limits for MANOVA outcomes, i.e. the correlations lie between r=-0.4 and r=0.9. To test for normality in the dependent variables the Shapiro-Wilk Test is used, as it is more appropriate for small sample sizes. All variables are normally distributed (p>0.05). Levene's test for equality of variances shows that there is homogeneity for all dependent variables (p>0.05). Finally, the homogeneity of variance-variance-covariance matrices is satisfied as the Box's M value of 11.78 (p=0.131>0.001).

MANOVA shows that there is a statistically significant difference in average resulting measures for each text between the clusters, $F(6,44)=2.26$, p=0.001; Wilk's λ=0.386, partial η^2=0.378. In contrast to the analysis on the L1 text clusters, the difference lies in the perceived readability ($F(2,26)=5.85$; p=0.009; partial η^2=0.328) and not in the comprehension scores ($F(3,26)=0.842$; p=0.443; partial η^2=0.066) or the perceived conceptual complexity ($F(3,26)=1.11$; p=0.345; partial η^2=0.345). Whilst the perceived conceptual readability appears to be associated with the clustering of eye movements, there is no relationship to the predefined levels of readability. This finding implies that readability of the text has a greater effect on L2 readers' eye movement than on L1 readers', and conversely, conceptual difficulty has a great effect on L1 readers' eye movements than L2 readers'.

Interestingly, there are correlations between the predefined conceptual difficulty and the L2 perceived readability (r=0.5, p=0.017) and L2 perceived conceptual difficulty (r=0.6, p<0.0005). In both cases, as the predefined conceptual difficulty gets harder, the L2 participants perceptions of both readability and conceptual difficulty get higher. However, the predefined readability has no significant correlations to either the perceived readability or the perceived conceptual difficulty. An interesting correlation is between the perceived conceptual difficulty and the perceived readability (r=0.8, p<0.0005). This implies that the two have a strong relationship, and even though there is no significant difference in perceived conceptual complexity between clusters, overall the two perceptions are related.

Finally, we investigate the readability of the texts further as this is an important factor on L2 readers' eye movements. The COH-Metrix L2 Readability Index is designed to rate the readability of text for L2 readers. We introduced this index in Chapter 6, but now investigate how these values differ in the clusters. The properties for each text were generated using COH-Metrix 3.0 (McNamara et al., 2013). It has been shown that the L2 readability index is more appropriate for describing the readability of texts for L2 readers (Crossley et al., 2008). However, for these texts it is not a consistent indicator for the derived difficulty based upon the clustering of eye movements, as there is no significant difference in the L2 indices between clusters 2 and 3 ($t(24)=-0.7495$, p=0.460).

There is a correlation (r=-0.6, p=0.002) between the L2 readability indices and the Flesch-Kincaid grade level for each text, so the two are somewhat related even though the L2 Readability Indices take into consideration more than the lexical structure of the text. There is also a correlation between the L2 readability indices

and predefined conceptual difficulty (r=-0.7, p<0.0005), however, there is no significant correlation to the predefined readability. This implies that the L2 readability indices are more related to the predefined conceptual difficulty than readability. However, when we consider the perceived variables of the text we see that the L2 readability indices have no significant correlation to the perceived readability and conceptual difficulty. This indicates that the use of the eye movements to calculate the derived text difficulty is useful as it takes into consideration more than the superficial nature of the text and is versatile for dealing with both L1 and L2 readers.

There are, however, strong correlations between the Flesch-Kincaid grade level and the perceived readability (r=0.5, p=0.008) and perceived conceptual difficulty (r=0.7, p<0.0005). As the Flesch-Kincaid grade level goes up so too does the participants perceptions of readability and conceptual difficulty. This implies that the Flesch-Kincaid grade level is perhaps still useful for assessing readability for L2 readers, as there are a relationship between this measure and the perceptions of readability and conceptual difficulty.

8.5 Discussion and Implications

The goal of the chapter was to investigate the clustering of eye movement measures to provide feedback based upon reading behaviour. The purpose was to first investigate the distinction between L1 and L2 groups as well as reading behaviours of participants for different texts based upon eye movements. Leading on from this, the average eye movement measures for each text were clustered to rate each text's derived difficulty. This provides feedback about how texts are read by the cohort to the author of the text.

Not all readers have the same reading skills, whether they are L1 or L2 readers. There can be variance within each group, not just between the groups. Some L2 readers may actually behave like L1 readers because they have been reading the language for so long. For this reason, we hypothesised that for the simplest text, A, there would a spectrum of eye movements and no clear distinction between the L1 and L2 groups. This was indeed what we observed. However, while there was no clear distinction between the two groups there are trends in the reading behaviours where the majority of the L2 readers tended to read thoroughly. However, there are many L1 readers that also read thoroughly, there is just a lower proportion of L1 readers that read thoroughly.

We further hypothesised that, as the text became more difficult, the clusters would become more distinct as the reading differences between the L1 and L2 participants should grow. This was not validated, and the distinctions between L1 and L2 readers declined as complexity increased. However, similar patterns in reading behaviour were observed, in that there are participants who read more thoroughly than others. So whilst the analysis did not reveal natural clusters between L1 and L2 readers, it did reveal that there are differences in reading behaviours. Furthermore, L2 readers are likely to read more thoroughly than L1 readers in easy to moderate texts, but not necessarily when the text is very difficult.

As text difficulty increases there is an observable increase in thoroughness of reading behaviour and time spent looking at the text, as we would expect. However, we only examined three texts in the first half of the analysis; the three that we know from Chapter 6 to have demonstrable differences in eye movements and from Chapter 7 to have perceptions more aligned to the predefined difficulty. Yet we saw from the analysis of the other texts that the expected differences in eye movements did not exist. Additionally, the perceptions of texts that lay outside of the main diagonal (A, E, and J) were poorest. This raises the question of whether the degree of difficulty assigned to the text based on conceptual difficulty and readability actually reflects of the true difficulty.

The problem becomes how to identify a robust method of determining text difficulty. Asking students to rate texts on difficulty is one method of obtaining a rating of the perceived text difficulty. However, this requires explicitly asking students to rate the texts on difficulty, which is time consuming and inconvenient to students. Calculating the difficulty of text based on students' behaviour, as measured by physiological and physical responses, would solve this problem. Eye gaze measures have been successfully used to indicate cognitive load (Rosch & Vogel-Walcutt, 2013). The proposition is that cognitive load of the learner should be neither too high nor too low, as this degrades learning outcomes (Paas et al., 2004). Additionally, eye gaze measures have been correlated with task complexity in visual tasks such as navigation (Waniek & Ewald, 2008) and search (Crosby et al., 2001). Coupling with the aforementioned effects of the predefined and perceived text difficulty on eye gaze measures, the use of eye gaze measures to calculate derived text difficulty is appealing.

We propose the use of eye gaze measures to calculate the derived difficulty of text for the purpose of providing feedback to the author of the text. This is similar to the use of answer-seeking behaviour to provide feedback regarding the derived difficulty of questions, discussed in Chapter 4. The categorisation of text from students' perceptions and eye gaze reveals that the students did not reflect the expected difficulties of the texts, on average. This supports the hypothesis that eye gaze measures can be used to provide a more accurate rating system for the derived reading difficulty of the text. Furthermore, there are differences between the L1 and L2 readers that make it necessary to categorise the texts for the two groups separately.

It is clear from this chapter, and the previous, that not all students neither rate nor perceive the text in the same way, even within their language groups. Students have different prior knowledge and expertise as well as different levels of reading abilities. The true power of this method of finding the derived difficulty is to be able to individually determine how difficult a student finds a text. In an adaptive environment this provides a wealth of information about the student as well as the materials. Additionally, in accordance with the cognitive load theory for the design of eLearning materials, it also provides the ability to deliver the correct level of difficulty to students, thus personalising the learning path.

8.6 Conclusion and Further Work

In this chapter we first investigated the differences between L1 and L2 readers using eye gaze measures. This revealed that there is no clear distinction between the two groups. L2 readers are more likely to read normally or thoroughly than to skim text. L1 readers have a broad spectrum of reading behaviours, and so reading behaviour does not distinguish the groups. Furthermore, we used cluster analysis of eye gaze measures to assess the derived reading difficulty of text. This is a useful tool for authors of eLearning materials and also allows consideration of individual differences between students. Given that we found that there are substantial differences between readers even within the same language group, it is necessary for any adaptive eLearning system to account for this and ensure that the appropriate level of difficulty of text is shown to the student. Therefore, the contributions of this chapter are a necessary part of the design of any adaptive text based eLearning.

Deriving text difficulty from average reading behaviour is useful for providing more information to the author about the derived difficulty of the text and for adapting a learning environment on a cohort basis. Leading on from this point is the idea of individually detecting a student's reading behaviour and adapting the system to their reading ability. That is, using the individual's eye movements to derive how difficult he or she finds the text to read. Whilst this was not investigated in this chapter it is an area of further research as the implications are important for adaptive eLearning.

However, the participants were all sourced as computer science students and the topic of the texts is a computer science topic. Changing the content matter to be something completely different from what they are used to, (such as biology or chemistry), would increase the difficulty even more. In this case would the eye gaze measures observed be different from what was observed in this study? Furthermore, in this situation would L1 readers exhibit reading behaviour that is more similar to the L2's reading behaviour? Further work should also be carried out to observe the effects of dynamically assessing the text difficulty whilst monitoring their cognitive load as well as reading comprehension.

Chapter 9

Discussion and Implications

Increasingly students are turning to online resources; however, the one-size-fits-all approach predominantly used in this medium is not effective for catering to the needs of all students. There has been much work on effective ways of presenting learning materials in learning environments (Clark & Mayer, 2011). Eye tracking is a useful method of investigating the reading process (Rayner, 1998) as well as cognitive load (Rosch & Vogel-Walcutt, 2013). In this chapter we look back at the studies and results presented throughout this thesis and tie them together to finally discuss the main research question of the thesis:

Can eye tracking be used to make eLearning environments more effective for first and second language English readers?

To do so, this chapter is divided into 2 sections; the first discusses how the results compare to the current literature, the second discusses application into an eLearning environment. To discuss the application, we propose the architecture of an eLearning environment that provides dynamic text selection and presentation based on eye movements. The students' eye gaze would be used to predict their comprehension level and the text difficulty altered to reflect this. This can be used to influence how students interact with the learning environment as well as how they learn the material, streamlining the learning process and optimising learning outcomes. The latter half of the discussion is based on work presented at IHCI 2014 (Copeland, Gedeon, & Caldwell, 2014) and throughout this thesis in the implications parts of the chapters.

9.1 Eye tracking in eLearning

The two propositions behind the research question are firstly, educational materials are being offered through online and electronic media more frequently. Universities are now frequently offering online and / or off-campus courses and degrees where students have little or no face-to-face interaction with their instructors or other students. The need for additional forms of student monitoring are necessary to detect when a student is under or over-performing so that they can either be given remedial help or advanced material. Even for university courses that deliver educational material traditionally, absenteeism from lectures is more prevalent and has been shown to negatively affect learning (Romer, 1993; Woodfield et al., 2006). However, the use of online learning can actually be beneficial for dealing with not only this problem, but also the problems encountered with large class sizes and dispersed students by providing consistency and accessibility in delivered materials (Welsh et al., 2003).

Secondly, online eLearning extends teaching and learning from the classroom to a wide and varied audience that has different needs, backgrounds, and motivations. Yet eLearning for the most part is one-size-fits-all. For these reasons there is a growing importance in designing effective eLearning materials that take these differences into consideration. One way of achieving this is by developing personalised education that is adaptive to students' individual needs. We focus on analysis of text materials and the comparison of first (L1) and second (L2) English language readers, as students with different language backgrounds are an increasing diversity in audiences of online learning materials.

We discussed in the literature review (Chapter 2) that adaption can be provided through various methods. The use of physiological and physical responses allows for real time adaption based upon cognitive load (Rosch & Vogel-Walcutt, 2013). Eye tracking in particular is becoming more precise, less expensive, and is not invasive or obtrusive for the student. This offers the possibility of using eye tracking as a common input to computer systems, and thus a potentially effective way of providing adaptive eLearning. Indeed, the use of eye tracking in adaptive eLearning is not new and has been shown to provide benefit in learning (Barrios et al., 2004; Calvi et al., 2008; D'Mello et al., 2012; Gütl et al., 2005; Mehigan et al., 2011; Porta, 2008). In particular, eye tracking has a long history of being used to analyse reading behaviour (Rayner, 1998). Furthermore, eye tracking is especially useful at analysing the implicit differences between different types of readers, such as linguistic background (Dednam et al., 2014; Kang, 2014).

The user studies presented in this thesis utilised eye-tracking technology to investigate how participants interact with an online eLearning environment, Wattle[17] (a Moodle[18] variant). L1 and L2 English language participants were sourced in order to investigate the differences between groups under different scenarios. Whilst it is known that L1 and L2 readers have different eye movements and

[17] https://wattle.anu.edu.au/ Last accessed: 22nd January 2016
[18] https://moodle.com/hq/ Last accessed: 22nd January 2016

reading behaviours (Dednam et al., 2014; Kang, 2014), there are several areas that had not been investigated. One of these is whether the two groups interact with eLearning environments in the same way, such as, how they answer questions in a tutorial, as investigated in Chapters 3 and 4. L2 readers take longer to read but perform at the same level when the materials are targeted to suit the right level of education. This result is expected given related research (Kang, 2014).

However, in Chapters 6 and 7 we see that once text is made more difficult, L2 readers perform worse than L1 readers in comprehension. However, the differences in eye movements between the different texts, and between the reader groups, were not as we expected. This warranted further investigation of how the eye movements were affected. In Chapter 7 the perceptions of text difficulty were investigated. L1 and L2 readers have different perceptions of text difficulty where L2 readers tend to overestimate the difficulty of a text and L1 readers underestimate difficulty. Neither group were good at perceiving difficulty. Using eye tracking data, better predictions of both the readability and conceptual difficulty of the text were achieved compared to participants' perceptions. Further analysis of participants perceptions indicates that both groups tend to conflate the levels of readability and conceptual difficulty. This results in both groups overestimating texts with the same levels of readability and conceptual difficulty and underestimating the other texts, especially when the readability is notably higher than the conceptual level and vice versa.

This raises the question of whether the predefined measure of text difficulty is adequate. Throughout this thesis we have rated readability using the Flesch-Kincaid grade level (Kincaid et al., 1975). However, this measure only deals with the surface properties of the text, not accounting for the content problems and is generally aimed at English text for L1 readers, not L2 readers (Zhang et al., 2013). Yet it is important that text features be considered differently for L1 and L2 readers since they have differential effects on reader type (Zhang et al., 2013). The analysis in Chapter 7 showed that the intended text difficulty might not be perceived this way. Perceptions are powerful predictors of learning outcomes (Lizzio et al., 2002) and alleviate anxieties about learning (Chang, 2005). However, it has been shown that people are poor at assessing their own skills (Kruger & Dunning, 1999) so the perceptions of students cannot be relied upon to measure the implicit difficulty of a text.

Instead, the text difficulty could be predicting by the cognitive load of the reader when reading that text, which can be determined using eye gaze (Rosch & Vogel-Walcutt, 2013). Eye gaze measures have been correlated with task complexity (Crosby et al., 2001; Waniek & Ewald, 2008). In Chapter 8, clustering of average eye movement measures per text showed that texts have significantly different average reading behaviours. Some texts are associated with low levels of reading behaviour (skimming) whereas others require higher levels of reading. Importantly though, the predefined text difficulty did not guarantee the amount of reading, so the results may surprise the author of the text and be helpful in creating and classifying text for eLearning. Additionally, the clustering of average eye movements from the text is different for L1 and L2 readers.

Whilst investigating eye movements for this use we also examined further the differences between L1 and L2 readers, exploring whether there are discernible differences between the two groups. Whilst the analyses in this thesis have shown that there are differences between L1 and L2 readers in eye movements, there are some notable similarities. The comprehension and eye movements from the two groups are affected equally by the presentation of text and comprehension questions (see Chapter 3). Furthermore, there is no difference between L1 and L2 readers in their answer-seeking behaviour (Chapter 4). Given these results, it is perhaps unsurprising that the results from Chapter 8 show that there is no clear distinction in eye movements between L1 and L2 readers. What we observe is instead a spectrum of eye movements that range from thorough reading to skimming reading. L2 readers are more likely to be amongst the readers who read thoroughly when the text is easy to moderate in complexity. However, there are many L1 readers who read thoroughly so this is not a discriminating factor. When the text gets really difficult to read and understand, the differences between the L1 and L2 readers became less clear and L2 readers tended to revert to normal reading rather than thorough reading. The L2 readers tend to not deal with the difficulty as well as the L1 readers who instead switch to more thorough reading when the difficulty is notably increased.

Differences between the reader groups such as this are important to take into consideration when designing learning materials for students. It is also important when deciding what texts should be given to students in an adaptive eLearning environment. Indeed, what is appropriate for an L1 reader may not be appropriate for an L2 reader and vice versa. This leads back to the proposition that eLearning environments can be adapted to the learner. It has been shown that adaptive learning environments result in significant improvement in learning outcomes compared to no adaption (Dingli & Cachia, 2014; Lach, 2013; Paramythis & Loidl-Reisinger, 2003).

Learning environment adaption can be based on different qualities of the learner such as the current understanding, emotional state such as stress (Calvi et al., 2008; Porta, 2008), learner style (Mehigan et al., 2011; Spada et al., 2008; Surjono, 2014), cognitive load (Coyne et al., 2009), and skill level (Chen, 2008). Methods of determining adaption, i.e. learner style or emotional state, also vary from using questionnaires (Surjono, 2011) to the use of biometric technology (Mehigan et al., 2011; Spada et al., 2008) and physical and physiological response data (Rosch & Vogel-Walcutt, 2013), especially eye tracking (Barrios et al., 2004; Calvi et al., 2008; D'Mello et al., 2012; Gütl et al., 2005; Mehigan et al., 2011; Porta, 2008). There are a broad range of scenarios that these adaptive technologies are directed at helping students, such as plugging into traditional online learning environments (Barrios et al., 2004; De Bra et al., 2013), or providing adaption in mobile environments (Mehigan & Pitt, 2013), or accounting for dyslexia (Alsobhi et al., 2015) and foreign language reading (Hyrskykari et al., 2000).

Building on all of this past research we are able to take the results from the studies presented in this thesis and add to the current knowledge base of adaptive eLearning. The contribution is solely in the domain of text-based learning materials.

Eye tracking can certainly be used to make learning via reading more effective in the context of eLearning.

In Chapter 3 we made the observation that different presentation sequences of text and comprehension questions affect performance outcomes and eye movements of participants. The order in which text and assessment questions are presented to students can therefore be manipulated to optimize performance outcomes and / or reading behaviour. That is, the sequence in which you present information and then assess it can have a large bearing on students' reading behaviour, learning performance, and perceived performance. In particular, making students rely on memory to answer comprehension questions promotes more accurate subjective ratings of understanding.

One of the major questions we investigate in this thesis is whether eye gaze can be used to predict reading comprehension measures in eLearning environments. The outcome of this is far more tangible than the previous question and needs far less explanation as to the benefits of such predictions. Being able to predict how well a reader understands text provides the benefits of removal of some comprehension assessment in place of using the implicit measure of eye tracking. It has been established that whilst eye movements are useful for investigating reading comprehension (Okoso et al., 2015; Rayner et al., 2006; Underwood et al., 1990), it is indeed difficult to predict quantified measures of reading comprehension (Martínez-Gómez & Aizawa, 2014). We have contributed to this research by investigating different methods for predicting reading comprehension (Chapter 5) and investigating how text difficulty affects eye gaze in such a way that reading comprehension prediction is improved (Chapter 6). Whilst the problem has not been solved, significant headway has been made. The lessons learnt from this investigation contribute to the production of effective eLearning materials. Most significantly, text difficulty should be considered from the students' perspective and that this differs for L1 and L2 readers (Chapter 7).

The results from this thesis are intended for application in eLearning environments. Consistent with past research these applications are intended to ultimately be incorporated in existing learning management systems (Barrios et al., 2004; De Bra et al., 2013). We will now discuss these uses in the next section.

9.2 Framework for dynamic text selection and presentation based on eye gaze

9.2.1 Framework Description

This thesis has investigated how eye gaze can be used, 1) to find optimal layouts of text and comprehension questions; 2) predict reading comprehension; and 3) predict implicit text difficulty. We now present how these conclusions are tied together by presenting their application in a framework for an eLearning system that dynamically presents text-based learning materials. The system utilises a commercial eye tracker. The framework for such a system is described in Figure 9.1.

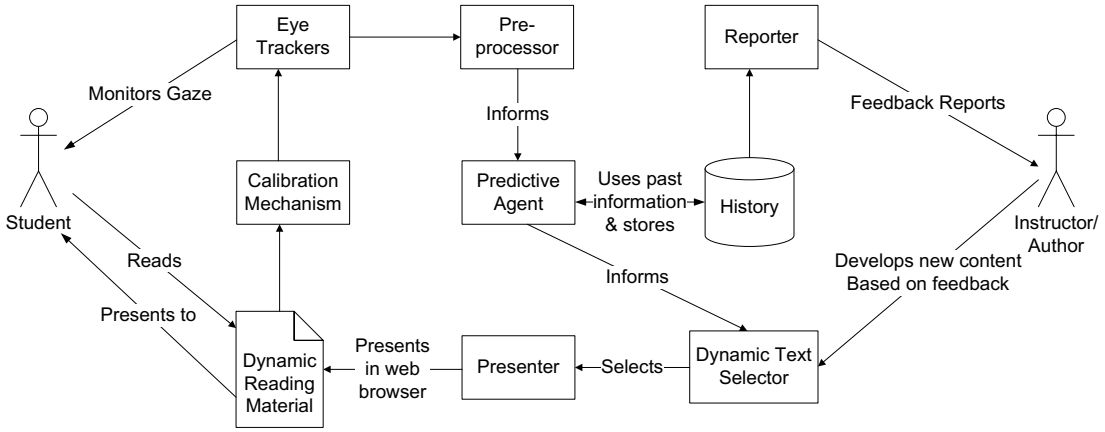


Figure 9.1. Framework for Dynamic presentation of reading material in an online learning environment (Copeland, Gedeon, & Caldwell, 2014).

Many of the components shown in Figure 9.1 are based on prior research, such as the calibration mechanism. These components will not be discussed in great detail. However, the components that showcase the use of the findings from this thesis will be discussed in more detail. These components are the Predictive Agent and the Reporter. The components of the framework are described in the rest of this subsection.

9.2.1.1 Calibration Mechanism

There is a need to account for error in recorded gaze location as it has been documented that eye trackers can lose precision during periods of use (Hyrskykari, 2006). A calibration mechanism will detect when the tracking data is out and prompt for a quick recalibration routine. It will do this by getting information from the content displayed. We propose using the calibration techniques described by (Hyrskykari, 2006); also see the auto-calibration we use described in Appendix C.

9.2.1.2 Pre-processor Mechanism

The output from eye trackers is x-y coordinate time series data, which is sent to the pre-processor mechanism to convert into eye movement measures. Pre-processing is necessary as it is the eye movements that can be used to make inferences about reading behaviour. The output from the eye tracker (eye gaze time series data) is sent to a pre-processor mechanism to turn the gaze points into fixation points and saccades using a fixation identification algorithm (Salvucci & Goldberg, 2000). A number of eye movement measures are calculated based on the content. Examples of these measures are answer-seeking behaviour as defined in this thesis in Chapter 4 and normalised number of fixations per paragraph. The output from the pre-processor mechanism is sent to the predictive agent.

9.2.1.3 Predictive Agent

This is one of the key parts of the system that highlights the use of the findings from this thesis. The eye movement measures used as the inputs for the predictive agent can be used to predict comprehension and implicit text difficulty. These predictions will be based on the presentation of the text. In Chapter 3 we identified that different presentation formats allowed for different deductions to be made

regarding comprehension and perceptions based on the sequence in which text and questions are shown to the student. Therefore, it is crucial to take this presentation method into consideration when drawing conclusions about a reader's current state. Additionally, the presentation method will determine which eye movement measures will be sent to the predictive agent. For example, the case where questions and text are shown together, data focussed on answer-seeking behaviour will be generated.

It is important to make it clear that the predictive agent has two functions, first to detect the reader's state in terms of comprehension and implicit text difficulty. Secondly, the predictive agent detects difficulty of the educational text and comprehension questions. Throughout the thesis we have highlighted two prospective uses of eye tracking in eLearning. The first is in regard to removing comprehension questions and implicitly predicting comprehension instead. The second is in regard to providing more information to the author of the eLearning materials so that materials can be optimised to facilitate learning. The results from the predictive agent from the latter function are passed to the Reporter, which is the second key part of the system, to achieve the latter goal.

Further, the student's previous learning behaviour is accessed to make an overall calculation of the student's current learning state. The current learning state is output to a content selector. Note from our analysis the predictive agent would be a combination of different machine learning techniques that are optimal for different situations. In the case where questions are shown with text, artificial neural networks using fuzzy output error (FOE-ANN) would be utilised as this provides optimal results (Chapter 5). However, when the questions are not shown with the text, then feature selection using genetic algorithms with a k-nearest neighbour classifier (GA-kNN) would be employed (Chapter 6). The output of the predictive agent is sent to the content selector.

9.2.1.3.1 Prediction of comprehension

The prediction of reading comprehension was covered in Chapters 4, 5 and 6. Chapter 4 highlights the use of eye movements to predict implicit comprehension when questions are presented with text. More specifically, when comprehension questions are presented with text, we showed in Chapter 3 that they are more likely to get the questions right. This is obvious; with the text there, the reader can search through the text and essentially do pattern matching with the words in the question. Do the questions then function to elicit true comprehension? This is not investigated here; instead we investigated answering behaviour, which can reveal the underlying state of the reader. More answer seeking indicates less confidence in answering a question. Whether this is due to not understanding, perception of not understanding, or simply that the reader did not read the text, this measure provides the system with key information that can be used with the reader's answers and their reading behaviour measures. The information in particular can be used as an implicit measure of how difficult a participant finds text and the corresponding questions. Of course, the use of the reading behaviour measures is crucial, as not reading the text just shows that the reader has not previously seen the text and therefore high amounts of answer seeking would be expected. In this case

the predictive agent can make recommendations to change the presentation format so that the questions are not visible with the text to ensure that reading occurs. Alternatively, no reading measured along with low or no answer seeking most likely indicates that the reader has prior familiarity with the content being assessed and so the predictive agent can recommend increasing the difficulty of the content or a change to the next subject matter.

Chapters 5 and 6 looked directly at the prediction of reading comprehension scores from eye movements. The results in Chapter 5 highlight that the prediction of reading comprehension scores when the questions are shown with the text is achievable with great accuracy using FOE-ANN. From Chapter 6 we found that the use of GA-kNN to predict reading comprehension was best when the questions are not shown with the text. Additionally, when the text is more difficult, the prediction results are better. The predictive agent can then inform the content selector on the level of understanding and the suitable next text can be shown. For example, if a student does not understand the content, simpler text can be shown. Or if a student has high understanding, then advanced level text can be shown, possibly skipping further basic and intermediate steps in the learning path.

Finally, in Chapter 7 we investigated the use of eye tracking to predict the implicit difficulty of the text. Whilst we mention that texts with differing degrees of difficulty should be shown to participants based on their measured comprehension, is it crucial to actually measure how difficult an individual student perceives the text. We have shown that this varies between students and is quite different between L1 and L2 readers. These results tie back to those found in Chapter 4 where we used answer-seeking behaviour higher as a measure of implicit comprehension and therefore difficulty in answering the comprehension questions. Therefore, the predictive agent can predict how difficult the student finds text / comprehension questions, depending of the sequence of presentation, which is crucial for successful selection of the next text to be shown to that student.

9.2.1.3.2 Prediction for feedback

The second function of the predictive agent is in predicting properties about the text and questions presented to students. More specifically, this involves analysing students' reading behaviours, answering behaviours, and understanding levels for each text. The predictive agent will predict how difficult text / comprehension questions are. Chapter 4 highlights the use of answer-seeking behaviour to measure question difficulty, which can be used as a feedback system to an instructor. This difficulty could be due to factors such as the technical nature of the material, and ambiguity in the material. Conversely, the instructor could see that the question is too easy and change it to be more challenging. This information could also be used to weight questions so that more difficult questions are weighted higher than those that are less difficult.

In Chapter 8 the clustering of eye movements revealed that the texts have different average reading behaviour that can be used to rate the texts' implicit difficulty. This provides the author of the text with a measure of the amount of reading that the text elicits and thus the implicit difficulty of the text, which may

indeed be different from the predefined text difficulty. This can be used to improve the quality of texts as well as to ensure that the appropriate level of text difficulty is shown to students.

9.2.1.4 Content Selector

The author of the learning material prefills the content selector with different texts. These texts will include different versions of the same content. The different versions will include different levels of text readability, concept difficulty as well as remedial and advanced level supplementary material. Based on the student's current state, as calculated by the predictive agent, a choice of version of the material is made by the content selector. This will also include generation of parameters that will change the rate at which the content is delivered to the student and change presentation format. The output of the selector is to the presenter.

9.2.1.5 Presenter

The presenter formats the selected content for the learning environment being used, such as Moodle. This is essentially the plug-in point to the existing learning environment.

9.2.1.6 Reporter

The reporting component is the second key component of the system is used by the author of the texts and questions to gain information about the difficulty of the questions and text, student performance and progression of learning data, in addition to reading behaviours. We have established that eye movements can be used in multiple ways to quantify the difficulty of text and questions. In Chapter 4 we showed how the average amount of answer-seeking behaviour that is observed could be used as an indicator for how difficulty on average students are finding particular questions to answer. This can then be used to check if particular students are performing above or below this average. The advantage of using this measure lies in the fact that it is a measure of the students' implicit behaviour. More specifically, just getting the average scores of students on questions will not give a true representation of the difficulty or ease of the questions. This was shown in Chapter 4 where for questions with similar average scores there were quite different ranges of answer seeking behaviour. We were therefore able to more accurately rank the questions on difficulty. The same argument applies for the rating of students' understanding. Below average answer seeking behaviour represents higher levels of understanding and high amounts of answer seeking behaviour indicates low levels of understanding, which may not be as accurately shown through the comprehension tests alone. The degree of answer seeking reflects how much they are learning now, while the comprehension score reflects the sum of prior and just learnt knowledge.

This line of inquiry was extended in Chapters 6, 7 and 8 where we investigated the effect of text difficulty on participants' eye movements and perceptions. Not all students have the same conception of difficulty so the predefined difficulty may not be how difficult the student finds the text. Calculating the implicit reading difficulty of texts would be performed at both the student and the cohort level. We have already discussed the purpose of measuring the text difficult. However on the

student level, as with the answer-seeking behaviour, the implicit difficulty for each student can be calculated to differentiate the students' abilities. The text complexity can be dynamically measured for the student rather than as a static measure. The long-term trends of these measured data can be used to assess how a student is progressing or if they are consistently underperforming and more assistance should be given to them. In an adaptive environment this provides a wealth of information and true power in giving learning material of the most appropriate level of difficulty to the student.

9.2.1.7 Student learning history

The student learning information is stored so that this information can be use in subsequent tutorials, and to track the learning progress. The information includes the basics such as what the student has learnt so far and their grades, but also includes their reading behaviour, how difficult they tend to find texts and questions, and the optimal way of presenting materials to them. This also allows the system to track how their perceptions change over time. This in itself is a measure for how the student is learning, as more accurate perceptions of text difficulty indicate increases in overall learning and comprehension, in addition to measuring levels of anxiety regarding the learning materials.

9.2.2 Replacement of Question and Answer Assessment

Since the predictive agent is designed to predict comprehension, the concept of removing question and answer-based assessment is plausible. In this case, a student's reading and eye movement behaviour could be used to assess the student's comprehension level. Prediction of reading comprehension from eye movement would allow for the removal of formative assessment of comprehension which could reduce learning time, workload, and potentially stress or anxiety of the students. Following on from this, predicting students' comprehension using eye tracking would allow the learning environment to 1) adapt the questions asked of students about the content and 2) alter the learning path to reflect the students' current understanding levels. This is similar to the traditional and summative interviews where answers to previous questions lead to easier or harder questions being asked.

The latter point is the main advantage of predicting comprehension from eye gaze. In the case where a student has read some learning materials and does not understand it, the student is then asked the same comprehension questions as all other students. Not understanding the questions makes it difficult and possibly increases the student's anxiety about the learning material. Two solutions arise from this, first is that the questions themselves are modified to be easier, perhaps covering more superficial understanding of the content, or text with more explanation could be provided to the student. Previous studies have shown that simplifying text can improve reading performance (Dingli & Cachia, 2014). Text with more explanation of the content that was not understood could then be given to the student, after which the student is assessed on the original comprehension questions. Secondly, instead of asking comprehension questions at all, the text with more explanation could be provided.

If we now consider the converse case where a student has an extremely high level of understanding, as is the case when the student has prior knowledge on a certain topic, this student may become frustrated or bored by being presented with easy content and unchallenging questions. Again, either the questions or the content could be altered to present these students with more difficult subject matter and questions that require much more thought and insight then the student with a lower level of understanding.

9.3 Summary

This chapter discussed the results from the chapters of this thesis. Each chapter addressed a sub-question to the question of whether eye tracking can be used to make learning more effective in eLearning environments. This overall question is approached in two ways. The first is that making eLearning environments better suited to the individual learner. The demonstration of these results is through the use of adaptive eLearning whereby the system adapts to the student's understanding levels and implicit difficulty. The second is the latent effects that eye tracking can have on making eLearning more effective. This is through the use of eye tracking to provide information to the author of the text and comprehension questions regarding their difficulty. This information can in turn be used to make better quality learning texts that are more accurately defined in difficulty. To show this we have tied the results from each chapter together in the presentation of a dynamic text selection method to make eLearning environments adaptive. The final chapter of this thesis is the conclusion and further work section.

Chapter 10

Conclusion

This thesis investigated the question of whether eye tracking can be used to make learning via reading more effective in the context of eLearning. Each chapter addressed a sub-question related to this research question. The investigation was approached from two directions: firstly, we investigated the use of eye tracking to adapt immediately to a student's understanding and implicit text difficulty. This involved investigating method for improving prediction of reading comprehension from eye gaze. The second approach was using eye tracking to analyse how a cohort of readers perceived, interacted with, and read text and comprehension questions within an eLearning environment. This information about reading behaviour of texts can, in turn, be used to make better quality materials that are more accurately defined in difficulty.

Throughout the investigation we explored the differences between L1 and L2 English readers, as this is an area of growing diversity in audiences of eLearning materials. This was accomplished by performing two large user studies in which readers' eye gaze was recorded. Chapters 3, 4 and 5 cover analysis of data collected from the first user study. In this study, different presentation sequences of text and comprehension questions were shown to readers as their eye gaze was recorded. In Chapter 3 we analysed how the sequence affected not only comprehension performance but also reading behaviour and student perceptions of performance. Chapter 4 covered analysis of reading and answering behaviour from a subset of the presentation sequences. From this analysis we proposed a new measure for reading

comprehension called answer-seeking behaviour, which can also be used to provide feedback to authors of the learning materials about the implicit difficulty of text and comprehension questions. Finally, in Chapter 5 investigated predicting reading comprehension from the eye gaze data collected. The purpose of this prediction is to make eLearning environments adaptive to students based on their implicit understanding. We found that good predictions can be made for a subset of the presentation sequences. However, there is still much improvement needed to predict reading comprehension scores when no questions were shown with text.

The second user study picked up from where the first left off. In Chapter 6 we investigated the effect of text readability and conceptual difficulty on eye movements and prediction outcomes of reading comprehension. This was in an attempt to improve classification results of reading comprehension prediction. Whilst we did not observe significant differences in eye movements and prediction accuracies between the levels of text difficulty that we expected, we were able to achieve higher prediction accuracies. This led us to investigation of predicting text difficulty from eye movements, which was explored in Chapter 7.

Chapter 7 also further examined the participants' perceptions of text difficulty, which indicated that text readability and conceptual difficulty interact to cause deviations from the predefined difficulty. Finally, in Chapter 8, cluster analysis of eye movements showed that average eye movements per text can be used to derive reading difficulty of the text. This can be used as feedback to the author of the text to assign derived text difficulty levels, as well as improve the quality of learning materials.

We conclude by tying these findings together in the discussion of how the research in this study can be applied to improve reading within eLearning environments. We propose an adaptive eLearning architecture that dynamically presents text to students and provides information to authors to improve the quality of texts and questions. However, much is left to investigate in this area so the following section of this chapter outlines some keys points that require further investigation as well as possible areas of interest for improving reading in eLearning environments.

10.1 Limitations

There are a number of limitations to the design of both studies and therefore the conclusions that can be drawn from them. In the first user study that is discussed in Chapters 3 through 5, a between-subjects design is used. However, this inherently introduces a lot of noise due to the differences in how people read. It would be beneficial to use a within-subjects design to reduce this noise and analyse in more depth the effects of layout on reading behaviour.

This leads to the next limitation, of both studies, which is that nature of the studies was highly artificial as they are conducted in a laboratory setting even though the tasks mimicked real life situations. The result of this highly artificial setting may, and probably did, influence their behaviour from a situation where

they are not being observed. Performing the studies in the laboratory setting likely altered the goals the readers purely because participants knew they were being observed. The most probable changes to goals and behaviour are 1) participants don't want appear lazy or unintelligent to the experimenter and so read the text more thoroughly than they would if unobserved, and 2) participants think they are being helpful by reading the text more thoroughly. In both cases, the end result is that the texts in the studies performed in this thesis are read more than they would be *in-the-wild*.

The results from *in-the-wild* studies would be highly beneficial in this area as they would capture more true to life behaviours. The focus of this thesis was to set ground work for the use of eye tracking to analyse reading and learning behaviour from text in eLearning environments. Additionally, the physical limitations of eye tracking hardware constrained our studies to be in a laboratory setting. For both reasons we limited the scope of studies performed so that they were in laboratory settings. However, recently small and portable eye tracking devices have become available. This does introduce the possibility of moving such studies into classrooms or *in-the-wild* settings and should most certainly be considered for future work.

Adding to these limitations are the highly restricted set of teaching materials that where considered in the studies. These limitations were briefly discussed in Chapter 3 but deserve more thorough discussion. As we concluded that studies should be extended to be *in-the-wild*, we so too conclude that the diversity and the nature of the content should be closer to a real world scenario. This means including different types of texts with different lengths, different topics, and especially those that are taken directly from course materials. Although the materials used in both studies were taken from a first year course at the university, they are a quite small subset of that course and they were chosen specifically because they were non-technical topics. This also resulted in a very particular subset of participants being used, mainly being selected from the course in which the materials were taken from. Whilst this meant we were testing on a realistic group of eLearning environments, and a realistic group who might access materials on these topics, it is still a subgroup of the population. Both of these factors are clear limitations on the studies and the results obtained.

Expanding the materials to a broader set of teaching materials, covering a much broader range of topics, and then testing these on a broader population would be necessary for future research. Firstly, because the current study only looked at a small subset and the results for different topics and different population subsets could differ from those reported here, and secondly, could increase the prediction results from the machine learning techniques.

The comprehension questions used in both studies were also limitations, as they represent significant subsets of the types of comprehension questions that could be asked. Additionally, the questions were designed in a way that made marking almost completely automated, as they were taken from a weekly tutorial quicz given to the students in the course. The questions where taken straight from the existing course materials, so they had been designed by the course convenor with

significant experience in teaching and developing the course materials. However, the questions themselves may be flawed in assessing comprehension for the purposes of the research being conducted. That is, the questions may not have spanned enough of the realms of comprehension assessment to fully and widely test the comprehension of students.

Extending the questions set to include short answer and essay questions would not only be valuable, but could also yield interesting results in more fully assessing comprehension. This limits the results from the second user study where the comprehension questions were specifically designed for each level of conceptual difficulty but not for readability. This was because different levels of concepts were being taught in the texts and so the questions had to reflect this. Although aligning comprehension questions to texts makes sense, there was no control for the quality control of questions. This means that some questions could have been inadvertently easier or harder than others, inherently skewing answering behaviour. This problem is prevalent in the first user study as well, where in Chapter 4 answer-seeking behaviour was used to assess question quality and consistency. In future research, there should be better quality control of comprehension questions and a larger set of different types of comprehension questions to more wholly assess comprehension.

In adapting the materials for the study there are many lessons learnt. This comes primarily in the length of texts given to students; in the two studies conducted as part of this thesis, the texts were likely too short. The reason for keeping the texts short was to keep down the experiment duration; however, the short nature of the texts most likely did not help in gathering more natural looking reading patterns. As the texts were short it was not difficult for the participants to read them and therefore not get bored and result in reduced reading behaviour. However, this is in itself an interesting research question to be investigated, does keeping text short increase reading?

The next lesson learnt is in the text construction. For both experiments, there were clear constraints on the readability level of the text and the conceptual difficulty. They were altered to keep them within limits so that the studies could test these effects on the reader. As we saw from Chapter 8, this is perhaps not the optimal or true way of finding the derived text difficulty. Previous studies have surveyed readers on the text difficulty and used that as the measure of text difficulty (Rayner et. al., 2006). Whilst this is more laborious than running readability formulae over the texts, it would provide more appropriate text difficulty results. Leading on from this, and in the spirit or testing *in-the-wild*, simply taking course material in its current state would be beneficial to do, rather than manipulating the materials to fit experimental conditions.

Perhaps the biggest limitation is that we did not formally test both prior knowledge and language skill. For both studies participants were only asked to rate their familiarity with the topics being examined and what language they first learnt to read in. This meant that they categorisation of participants into the L1 / L2 groups has limitations. This is observed in Chapter 8, where some L2 participants read like L1 participants and vice versa. There are clear benefits of accounting for factors like

prior-knowledge and language skill when constructing machine learning models of reading and learning behaviour, consequently, not formally accounting for either, limits the accuracy of the models and the potential of the models. Further work should be to take both into account.

This leads to the next point, which is, how would one deal with the situation where participants know the topic area sufficiently already that they answer the questions without reading the learning materials? Indeed, this is a limitation of the current studies given that prior-knowledge was not accurately assessed. It would be advantageous to detect that this situation, and would be an interesting future user study. A potential solution for this is to perform a pre-assessment similar to format D from Chapter 3, where participants were shown the questions before given the reading materials. Participants could be asked to complete the pre-assessment to the best of their ability and rate their knowledge on the subject matter and confidence in answering the questions. Then given the reading materials and observation of their eye gaze could take place. Such a scenario would set up testing for prior knowledge and therefore detection of eye gaze patterns of those who have (differing degrees of) prior-knowledge. This would allow for much more accurate personalisation of adaptive content. For example, detecting that a student has significant prior-knowledge of a subject allows the adaptive system to completely bypass the subject for that student. Moreover, if a student is detected to have partial prior-knowledge, then that student could be provided only with the materials that cover their knowledge gap. This also draws to light the potential benefits of combining pre-assessment with the use of eye tracking as complimentary drivers of adaptive and personalised eLearning.

Whilst the latter part of the analysis in the thesis considered the individual differences between readers, as we have discussed so far the clear limitations of not testing prior-knowledge and language and reading skills meant that individual models of readers were not considered more thoroughly. The machine learning prediction performance might be increased if there were more detailed models of each individual, particularly in terms of their actual reading skill, their prior-knowledge of the topic, and their demonstrated learning and comprehension performance.

10.2 Future work

There were several limitations of the first user study presented in the thesis. Some of these limitations were addressed in the second user study, such as variance of text complexity and analysis of perceptions of text complexity. Firstly, only two types of questions were investigated. Whilst these were chosen to test different parts of the comprehension spectrum, different types of questions and texts should be investigated. Further, given that inclusion of appropriate images and / or animations can enhance learning outcomes (Clark & Mayer, 2011; Harp & Mayer, 1998; Mayer et al., 2001; Sanchez & Wiley, 2006; Sung & Mayer, 2012) this should also be included in the different presentation sequence. The effects of inclusion of appropriate images and / or animations to text on L2 reading comprehension

performance as well as comparison of L1 and L2 perceptions of difficulty should also be investigated.

The second user study involved reading educational text about digital images. After reading the text, participants were asked to identify within digital images examples they had learnt from the text, such as resolution, manipulations, and bit depth. An example of this was that participants were asked to identify manipulated images and the manipulations (Caldwell et al., 2015). Accuracy of identifying the other factors related to what the students learnt should be investigated to assess the applied knowledge as well as the reading comprehension.

Additionally, eye gaze and pupil dilation data are the only biometric data used in this thesis. Inclusion of galvanic skin response (GSR), electrocardiogram (ECG), and electroencephalogram (EEG) data should be investigated. Such inclusion could lead to improve reading comprehension prediction results. These biometric measures have been used with great success to predict stress during reading (Sharma & Gedeon, 2012) as well as predicting differences in stress between males and females during reading (Sharma & Gedeon, 2011, 2013a, 2013b). These biometric data have also been used to predict the nature of document content in relation to national security (Chow & Gedeon, 2015).

The extension of these findings to mobile devices, such as smart phones and tablets should be investigated to see if the results are generalizable to these devices. The use of mobile devices, and hence mobile learning (mLearning), is becoming more widespread and therefore increasing the need for making learning materials effective on these devices. Indeed, there are differences in behaviours when using small screen devices compared to large screen devices, such as different search behaviour and that fact that users have trouble extracting information from search results on smaller screens (Kim et al., 2012; 2015). This implies that care should be taken when designing learning materials for different devices.

Whilst studies have shown that adaptive eLearning is beneficial in learning (Dingli & Cachia, 2014; Paramythis & Loidl-Reisinger, 2003) the effects of altering the difficulty of text shown to students based upon their understanding should be investigated further to see if, and to what extent, this provides learning benefits. Both short and longitudinal studies on these effects would be beneficial in determining any short and long terms benefits of such adaptions.

Throughout the thesis we have highlighted the use of eye gaze to predict reading comprehension. In Chapter 8 the idea of using eye tracking to calculate cognitive load was introduced, which is the strain being placed on the learner's working memory (Rosch & Vogel-Walcutt, 2013). The idea behind using cognitive load to adapt education materials is that there a limitations of working memory, where inducing too much load via an overly complex learning task is detrimental to learning, however so too is underload caused by a too simplistic task (Paas et al., 2004). Therefore if a learner were being too challenged according to their cognitive load then in an adaptive eLearning environment the material would be made simpler for the learner or more challenging in the opposite case. This is the same

preposition we use except that we have highlighted changes be made based on reading comprehension. However, the inclusion of cognitive load measures along with comprehension could be highly beneficial, especially in analysing the relationship between cognitive load and reading behaviour.

Attention guiding is another way in which learning environments can be made more optimal for learners. It can be used to both minimize distraction of the learner as well as draw the learner's attention to the important or relevant parts of the learning material. Attention guiding has been shown to improve problem solving by conveying task-relevant information (Groen & Noyes, 2010). Attention guiding can provide visual cues by using colours to emphasise relevant parts of animations (Boucheix & Lowe, 2010), or by zooming in on parts of animations (Amadieu et al., 2011), and signalling parts relevant parts of diagrams by adding temporary colour changes (Ozcelik et al., 2010). The addition of eye tracking data to the paradigms has been found to enhance their effectiveness (Boucheix & Lowe, 2010; Ozcelik et al., 2010). This leads to a similar concept which is the use of eye tracking to guide student learning using eye movement modelling examples (EMME) (Jarodzka et al., 2010). EMME is a technique where the eye movements of experts are superimposed onto a task to show how that expert performed a task. This is easily visualised when considering a visual task such as watching a video to learn how to classify fish locomotion (Jarodzka et al., 2013) or to diagnose seizures (Jarodzka et al., 2010). Importantly, this is achieved by blurring out areas where the expert was not looking at (Jarodzka et al., 2010) or using a dot or highlighting effect to focus on the parts that the expert was looking (Jarodzka et al., 2013). A similar approach has been used to in the context of reading and viewing an associated diagram (Mason et al., 2015). The use of EMME in reading could be explored.

Alternatively, investigation of methods of using rapid serial visual presentation (RSVP) to ensure reading of all text could be explored. RSVP of text has been shown to keep reading comprehension constant during speed-reading (Dingler et al., 2015). This poses the question of whether integrating the EMME and RSVP would be beneficial for eLearning. The idea being that using RSVP techniques to guide learning through text could possibly help reduce distractions, as it motivates the reader to keep up with the text, and secondly, could promote more thorough reading and comprehension of the text. In this way there would actually be no "rapid" presentation in the real sense of the technique, the goal would not be to promote speed-reading, or rapid reading, but rather to force reading (so the presentation of words or text would not be as fast). Examples of RSVP include the open source framework Squirt[19] where one word is presented to a reader at a time. Another example is dynamic underlining of text to mark (Dingler et al., 2015).

Both of these concepts underpin the idea of streamlining the reading process as well as mitigating distraction. Distractions affect reading comprehension and behaviour (Halin et al., 2014a; 2014b; Sörqvist et al., 2010). Digital environments present many distractions, often bombarding users with information that disrupts processing (Maglio & Campbell, 2003). Reduction of visual distractions is pertinent

[19] https://www.squirt.io/ Last Accessed: 25th August 2015

for avoid irrelevant objects increasing cognitive load (Sweller et al., 1998). When cognitive load induced by the primary task is made higher, then distractors are attended to far less (DeLeeuw et al., 2010). However, an interesting question for further research is whether eye tracking can be used to reduce the effect of distractions on reading. The idea of mitigating distractions during reading using eye-tracking technology plays on grabbing the attention back from students.

In Appendix D we present a preliminary study that looked at mitigating visual distractions during reading in a distracting environment. The results from the study first indicate that participants have high levels of distractions whilst studying, setting precedent for the need for distraction mitigation. The results show that the distraction mitigation signals helped L2 readers to restart reading and hence to read the hard text more effectively than in their absence. Additionally, the questionnaire data demonstrated that for both the L1 and L2 groups the mitigation signals helped to recover from a distraction by drawing participants' attention back to the text as well as indicating where to start reading from. While the study had limitations and was preliminary, the results indeed show that there is no potential for such technology. One of the main problems with the experiment that could have led to inconclusive results was inaccuracy of the eye tracker. Participants noted that the text effect did not always appear where they had last read and when it was not working at all. Instead the effect would appear sporadically around the page causing the process to be more distracting than the planned experimental distractions themselves. In the busy environments in which we now work, the concept of mitigating distraction is highly important especially when it is known that distractions affect reading outcomes (Halin et al., 2014a; 2014b; Sörqvist et al., 2010). This is an area of active research that needs further investigation.

The discussion of distraction mitigation leads to the integration of attention managers that use eye tracking to manage alerts to the user into eLearning environments. The idea behind attention managers is that information, namely alerts, is controlled by a managing service to minimise the effect of distractions by scheduling them during skimming rather than thorough reading. Interruptions not only have negative effects on users task performance and emotional state, but these effects are more intense if the user is under high mental load (Adamczyk & Bailey, 2004; Bailey et al., 2001).

References

Adamczyk, P. D., & Bailey, B. P. (2004). If not now, when?: the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 271-278). ACM.

Alsobhi, A. Y., Khan, N., & Rahanu, H. (2015). DAEL Framework: A New Adaptive E-learning Framework for Students with Dyslexia. *Procedia Computer Science, 51,* 1947-1956.

Amadieu, F., Mariné, C., & Laimay, C. (2011). The attention-guiding effect and cognitive load in the comprehension of animations. *Computers in human behavior, 27(1),* 36-40.

Amaratunga, D., Cabrera, J., & Lee, Y.-S. (2008). Enriched random forests. *Bioinformatics, 24(18),* 2010-2014.

Anderson-Inman, M. A. H., Lynne. (1999). Supported text in electronic reading environments. *Reading & Writing Quarterly, 15(2),* 127-168.

Anderson-Inman, L., & Horney, M. A. (2007). Supported eText: Assistive technology through text transformations. *Reading Research Quarterly, 42(1),* 153-160.

Atkins, M. S., Moise, A., & Rohling, R. (2006). An application of eyegaze tracking for designing radiologists' workstations: Insights for comparative visual search tasks. *ACM Transactions on Applied Perception, 3(2),* 136-151.

Bailey, B. P., Konstan, J. A., & Carlis, J. V. (2001). The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Proceedings of INTERACT* (Vol. 1, pp. 593-601).

Barrios, V. M. G., Gütl, C., Preis, A. M., Andrews, K., Pivec, M., Mödritscher, F., & Trummer, C. (2004). AdELE: A framework for adaptive e-learning through eye tracking. In *Proceedings of IKNOW* (pp. 609-616).

Barton, D. (2007). *Literacy: an introduction to the ecology of written language*. Wiley-Blackwell.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91(2),* 276.

Bernard, M., & Mills, M. (2000). So, what size and type of font should I use on my website. *Usability news, 2(2)*, 1-5.

Beymer, D., & Flickner, M. (2003). Eye gaze tracking using an active stereo head. In *Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. II-451). IEEE.

Beymer, D., Russell, D., & Orton, P. (2008). An eye tracking study of how font size and type influence online reading. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction* (pp. 15-18). British Computer Society.

Beymer, D., & Russell, D. M. (2005). WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze. In *CHI'05 extended abstracts on Human factors in computing systems* (pp. 1913-1916). ACM.

Biedert, R., Buscher, G., Lottermann, T., Schwarz, S., Möller, M., & Dengel, A. (2010). The Text 2.0 Framework: writing web-based gaze-controlled realtime applications quickly and easily. In *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction* (pp. 114-117). ACM.

Biedert, R., Buscher, G., Schwarz, S., Hees, J., & Dengel, A. (2010). Text 2.0. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems.* (pp. 4003-4008): ACM.

Bohn, R. E., & Short, J. E. (2009). *How Much Information?*: 2009 Report on American Consumers: University of California, San Diego, Global Information Industry Center.

Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J. M., Azevedo, R., & Bouchet, F. (2013). Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In *Proceedings of International Conference on Artificial Intelligence in Education.* (pp. 229-238). Springer Berlin Heidelberg.

Boucheix, J.-M., & Lowe, R. K. (2010). An eye tracking comparison of external pointing cues and internal continuous cues in learning with complex animations. *Learning and Instruction, 20(2),* 123-135.

Bowman, L. L., Levine, L. E., Waite, B. M., & Gendron, M. (2010). Can students really multitask? An experimental study of instant messaging while reading. *Computers & Education, 54(4),* 927-931.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. Cognitive *Psychology, 2(4),* 331-350.

Brasel, S. A., & Gips, J. (2011). Media multitasking behavior: Concurrent television and computer usage. *Cyberpsychology, Behavior, and Social Networking, 14(9),* 527-534.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24(2),* 123-140.

Breiman, L. (2001). Random forests. *Machine Learning, 45(1),* 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees:* CRC press.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment, 8(1),* 13-25.

Bunch, G. C., Walqui, A., & Pearson, P. D. (2014). Complex text and new common standards in the United States: Pedagogical implications for English learners. *Tesol Quarterly, 48(3),* 533-559.

Burton, L., Westen, D., & Kowalski, R. (2009). *Psychology 2nd Edition.:* Wiley.

Buscher, G., Dengel, A., Biedert, R., & Van Elst, L. (2012). Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond. *ACM Transactions on Interactive Intelligent Systems, 1(2),* Article 9.

Buscher, G., Dengel, A., & Elst, L. v. (2008). Eye movements as implicit relevance feedback. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems, Florence, Italy.* (pp. 2991-2996). ACM.

Buscher, G., Dumais, S. T., & Cutrell, E. (2010). The good, the bad, and the random: an eye-tracking study of ad quality in web search. If *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. (pp. 42-49): ACM

Caldwell, S., Gedeon, T., Jones, R., & Copeland, L. (2015). Imperfect Understandings: A Grounded Theory And Eye Gaze Investigation Of Human Perceptions Of Manipulated And Unmanipulated Digital Images. In *Proceedings of 3rd International Conference on Multimedia and Human-Computer Interaction (MHCI'15).*

Calvi, C., Porta, M., & Sacchi, D. (2008). e5Learning, an e-learning environment based on eye tracking. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies* (pp. 376-380). IEEE.

Campbell, C. S., & Maglio, P. P. (2001). A robust algorithm for reading detection. In *Proceedings of the 2001 workshop on Perceptive user interfaces* (pp. 1-7). ACM.

Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., & Enns, J. (2014). Highlighting interventions and user differences: informing adaptive information visualization support. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 1835-1844). ACM.

Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. Eye movements in reading: *Perceptual and language processes*, 275-307.

Chang, S. E. (2005). Computer anxiety and perception of task complexity in learning programming-related skills. *Computers in human behavior, 21(5),* 713-728.

Chen, C.-M. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers & Education, 51(2),* 787-814.

Chen, S.-C., She, H.-C., Chuang, M.-H., Wu, J.-Y., Tsai, J.-L., & Jung, T.-P. (2014). Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities. *Computers & Education, 74,* 61-72.

Chow, C., & Gedeon, T. (2015). Classifying Document Categories based on Physiological Measures of Analyst Responses. In *Proceedings of Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on* (pp. 421-425). IEEE.

Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning:* John Wiley & Sons.

Cohen, L. G., Celnik, P., Pascual-Leone, A., Corwell, B., Faiz, L., Dambrosia, J., Catala, M. D. (1997). Functional relevance of cross-modal plasticity in blind humans. *Nature, 389(6647),* 180-183.

Conati, C., Jaques, N., & Muir, M. (2013). Understanding Attention to Adaptive Hints in Educational Games: An Eye-Tracking Study. *International Journal of Artificial Intelligence in Education, 23(1)*, 136-161.

Conati, C., & Merten, C. (2007). Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. *Knowledge-Based Systems, 20(6),* 557-574.

Copeland, L. (2011). *Extraction of information from Eye Gaze Data* (Honours Thesis). Research School of Computer Science. Australian National University, ACT, Australia.

Copeland, L., & Gedeon, T. (2013a). The effect of subject familiarity on comprehension and eye movements during reading. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration.* (pp. 285-288): ACM.

Copeland, L., & Gedeon, T. (2013b). Measuring reading comprehension using eye movements. In *Proceedings of Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on.* (pp. 791-796): IEEE.

Copeland, L., & Gedeon, T. (2014a). Effect of presentation on reading behaviour. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*. (pp. 230-239). ACM.

Copeland, L., & Gedeon, T. (2014b). What are You Reading Most: Attention in eLearning. *Procedia Computer Science, 39,* 67-74.

Copeland, L., & Gedeon, T. (2015). Visual Distractions Effects on Reading in Digital Environments: A Comparison of First and Second English Language Readers. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. (pp. 506-516). ACM.

Copeland, L., Gedeon, T., & Caldwell, S. (2014). Framework for Dynamic Text Presentation in eLearning. *Procedia Computer Science, 39,* 150-153.

Copeland, L., Gedeon, T., & Caldwell, S. (2015). Effects of Text Difficulty and Readers on Predicting Reading Comprehension from Eye Movements. In *Proceedings of the IEEE 6th International Conference on Cognitive Infocommunications (CogInfoCom)* 2015, Győr, Hungary. (pp. 407-412). IEEE.

Copeland, L., Gedeon, T., & Mendis, S. (2014a). Fuzzy Output Error as the Performance Function for Training Artificial Neural Networks to Predict Reading Comprehension from Eye Gaze. In *Proceedings of The 21st International Conference on Neural Information Processing 2014.* (pp. 586-593). Springer International Publishing.

Copeland, L., Gedeon, T., & Mendis, S. (2014b). Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research, 3(3),* p35.

Copeland, L. D., & Gedeon, T. D. (2015). Tutorials in eLearning; How Presentation Affects Outcomes. *Emerging Topics in Computing, IEEE Transactions on,* PP(99), 1-1.

Coyne, J. T., Baldwin, C., Cole, A., Sibley, C., & Roberts, D. M. (2009). Applying real time physiological measures of cognitive load to improve training Foundations of augmented cognition. *Neuroergonomics and operational neuroscience* (pp. 469-478): Springer.

Crosby, M. E., Iding, M. K., & Chin, D. N. (2001). Visual search and background complexity: Does the forest hide the trees? *User Modeling 2001* (pp. 225-227): Springer.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly, 42(3)*, 475-493.

Crowder, R. G., & Wagner, R. K. (1992). *The Psychology of Reading: An Introduction (Second Edition ed.)*: Oxford University Press.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other

cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10(3)*, 277-299.

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies, 70(5),* 377-398.

De Bra, P., Smits, D., van der Sluijs, K., Cristea, A. I., Foss, J., Glahn, C., & Steiner, C. M. (2013). GRAPPLE: Learning management systems meet adaptive learning environments. In *Intelligent and Adaptive Educational-Learning Systems* (pp. 133-160): Springer.

De Jong, K. A. (1993). Genetic algorithms are NOT function optimizers. *Foundations of genetic algorithms, 2*, 5-17.

DeBoer, J., Stump, G. S., Seaton, D., & Breslow, L. (2013). Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002 x. In *Proceedings of the Sixth Learning International Networks Consortium Conference.* (Vol. 4).

Dednam, E., Brown, R., Dani, #235, Wium, l., & Blignaut, P. (2014). The Effects of Mother Tongue and Text Difficulty on Gaze Behaviour while Reading Afrikaans Text. In *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology, Centurion, South Africa.* (p. 334). ACM.

Dehaene, S. (2009). *Reading in the brain: the new science of how we read*: Penguin.

DeLeeuw, K. E., Mayer, R. E., & Giesbrecht, B. (2010). How does text affect the processing of diagrams in multimedia learning? *Diagrammatic Representation and Inference* (pp. 304-306): Springer.

DeStefano, D., & LeFevre, J.-A. (2007). Cognitive load in hypertext reading: A review. *Computers in human behavior, 23(3),* 1616-1641.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics, 35(10),* 1297-1326.

Dillon, A. (2004). *Designing usable electronic text: Ergonomic aspects of human information usage*: CRC Press.

Dillon, A., & Gabbard, R. (1998). Hypermedia as an educational technology: A review of the quantitative research literature on learner comprehension, control, and style. *Review of educational research, 68(3),* 322-349.

Dingler, T., Shirazi, A. S., Kunze, K., & Schmidt, A. (2015). Assessment of stimuli for supporting speed reading on electronic devices. In *Proceedings of the 6th Augmented Human International Conference, Singapore.* (pp. 117-124). ACM.

Dingli, A., & Cachia, C. (2014). Adaptive eBook. In *Proceedings of the Interactive Mobile Communication Technologies and Learning (IMCL), 2014 International Conference on.* (pp. 14-19). IEEE.

Dombi, J. (1990). Membership function as an evaluation. *Fuzzy sets and Systems, 35(1),* 1-21.

Dombi, J., & Gera, Z. (2005). The approximation of piecewise linear membership functions and Łukasiewicz operators. *Fuzzy sets and Systems, 154(2),* 275-286.

Dombi, J., & Gera, Z. (2008). Rule based fuzzy classification using squashing functions. *Journal of Intelligent and Fuzzy Systems, 19(1),* 3-8.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension A Brief History and How to Improve Its Accuracy. *Current Directions in Psychological Science, 16(4),* 228-232.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12(3),* 83-87.

Eagleman, D. (2011). *Incognito: The secret lives of the brain*. New York: Vintage Books.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes, 105(1),* 98-121.

Eklund, J., & Brusilovsky, P. (1999). Interbook: an adaptive tutoring system. *UniServe Science News, 12(3),* 8-13.

Engbert, R., & Kliegl, R. (2001). Mathematical models of eye movements in reading: a possible role for autonomous saccades. *Biological Cybernetics, 85,* 77-87.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research, 42(5),* 621-636.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). Swift: A dynamical model of saccade generation during reading. *Psychological Review, 112(4),* 777-813.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The quarterly journal of experimental psychology, 63(4),* 639-645.

Fahey, D. (2009). *A Preliminary Investigation into using eye-tracking to analyse a person's reading behaviour* (Honours Thesis). Research School of Computer Science. Australian National University.

Fletcher, J. M. (2006). Measuring Reading Comprehension. *Scientific Studies of Reading, 10(3),* 323-330.

Fox, A. B., Rosen, J., & Crawford, M. (2009). Distractions, distractions: does instant messaging affect college students' performance on a concurrent reading comprehension task? *CyberPsychology & Behavior, 12(1),* 51-53.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10(3),* 301-322.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14(2),* 178-210.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence, 14(771-780),* 1612.

Fritz, T., Begel, A., Müller, S. C., Yigit-Elliott, S., & Züger, M. (2014). Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering.* (pp. 402-413). ACM.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The Validity of Informal Reading Comprehension Measures. *Remedial and Special Education, 9(2),* 20-28.

Garrett, D., Peterson, D. A., Anderson, C. W., & Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 11(2),* 141-144.

Gedeon, T., Copeland, L., & Mendis, B. S. (2012). Fuzzy Output Error. *Australian Journal of Intelligent Information Processing Systems, 13(2),* 37-43.

Gedeon, T. D., Zhu, D., & Mendis, B. S. U. (2008). Eye gaze assistance for a game-like interactive task. *International Journal of Computer Games Technology*, 3.

Gedeon, T., Zhu, X., Copeland, L., & Sharma, N. (2015). Feature selection and interpretation of GSR and ECG sensor data in Biofeedback Stress Monitoring. In *Proceedings of the Ninth International Conference on Sensor Technologies and Applications (SENSORCOMM 2015)*, Venice, Italy.

Gehring, W. J. (2005). New Perspectives on Eye Development and the Evolution of Eyes and Photoreceptors. *Journal of Heredity, 96(3),* 171-184.

Gera, Z., & Dombi, J. (2005). Genetic Algorithm with Gradient Based Tuning for Constructing Fuzzy Rules. *Publications of International Symopsium of Hungarian Researchers of Computational Intelligence,* 86-95.

Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: *A practitioner's guide.* 493-516.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40(5),* 223-234.

Groen, M., & Noyes, J. (2010). Solving problems: How can guidance concerning task-relevancy be provided? *Computers in human behavior, 26(6),* 1318-1326.

Gustavsson, C. J. (2010). *Real Time Classification of Reading in Gaze Data* (Masters Thesis). School of Computer Science and Engineering. Royal Institute of Technology. Stockholm, Sweden.

Gütl, C., Pivec, M., Trummer, C., García-Barrios, V. M., Mödritscher, F., Pripfl, J., & Umgeher, M. (2005). Adele (adaptive e-learning with eye-tracking): Theoretical background, system architecture and application scenarios. *European Journal of open, Distance and E-learning (EURODL),* 2.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3,* 1157-1182.

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on, 5(6),* 989-993.

Halin, N., Marsh, J. E., Haga, A., Holmgren, M., & Sörqvist, P. (2014a). Effects of speech on proofreading: can task-engagement manipulations shield against distraction? *Journal of Experimental Psychology: Applied, 20(1),* 69.

Halin, N., Marsh, J. E., Hellman, A., Hellström, I., & Sörqvist, P. (2014b). A shield against distraction. *Journal of Applied Research in Memory and Cognition, 3(1),* 31-36.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence, 12(10),* 993-1001.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90(3),* 414.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences, 7(11),* 498 - 504.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143(3611),* 1190-1192.

Hornof, A. J., & Halverson, T. (2002). Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers, 34(4),* 592-604.

Houts, P. S., Doak, C. C., Doak, L. G., & Loscalzo, M. J. (2006). The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence. *Patient education and counseling, 61(2),* 173-190.

Howland, J. L., & Moore, J. L. (2002). Student Perceptions as Distance Learners in Internet-Based Courses. *Distance Education, 23(2),* 183-195.

Huey, E. B. (1968). *The Psychology & Pedagogy of Reading*. Cambridge: MIT Press.

Hyona, J., Lorch Jr, R. F., & Rinck, M. (2003). Chapter 16 - Eye Movement Measures to Study Global Text Processing. In *J. Hyona, R. Radach, R. R. H. DeubelA2 - J. Hyona & H. Deubel (Eds.), The Mind's Eye* (pp. 313-334). Amsterdam: North-Holland.

Hyrskykari, A. (2006). Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading. *Computers in human behavior, 22(4),* 657-671.

Hyrskykari, A., Majaranta, P., Aaltonen, A., & Räihä, K.-J. (2000). Design issues of iDICT: a gaze-assisted translation aid. In *Proceedings of the 2000 symposium on Eye tracking research & applications.* (pp. 9-14). ACM.

Initiative, C. C. S. S. (2010). *Appendix A: Research supporting key elements of the standards*. Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects.

Initiative, C. C. S. S. (2012). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*: Common Core Standards Initiative.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 311-320). ACM.

Iqbal, S. T., & Bailey, B. P. (2004, October 6–9). Using Eye Gaze Patterns to Identify User Tasks. In *Grace Hopper Celebration of Women in Computing 2004*, (pp. 5-10).

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. In *CHI '04 extended abstracts on Human factors in computing systems* (pp 1477-1480). ACM.

Isokoski, P., Joos, M., Spakov, O., & Martin, B. (2009). Gaze controlled games. *Universal Access in the Information Society, 8(4),* 323-337.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49(10),* 1295 - 1306.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind, 2(3),* 4.

Jacobsen, W. C., & Forste, R. (2011). The wired generation: Academic and social outcomes of electronic media use among university students. *Cyberpsychology, Behavior, and Social Networking, 14(5),* 275-280.

Jain, A. K., Mao, J., & Mohiuddin, K. (1996). Artificial neural networks: A tutorial. *Computer, 29(3),* 31-44.

Jaques, N., Conati, C., Harley, J. M., & Azevedo, R. (2014). Predicting affect from gaze data during interaction with an intelligent tutoring system. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. (pp. 29-38). Springer International Publishing.

Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2010). Learning perceptual aspects of diagnosis in medicine via eye movement modeling examples on patient video cases. In *S. Ohlsson & R. Catrambone (Eds.), Cognition in flux: Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1703–1708). Austin: Cognitive Science Society.

Jarodzka, H., van Gog, T., Dorr, M., Scheiter, K., & Gerjets, P. (2013). Learning to see: Guiding students' attention via a model's eye movements fosters learning. *Learning and Instruction, 25*, 62-70.

Kahneman, D. (1973). *Attention and Effort*: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science, 154(3756),* 1583-1585.

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology, 79(1, pt. 1),* 164.

Kang, H. (2014). Understanding online reading through the eyes of first and second language readers: An exploratory study. *Computers & Education, 73,* 1-8.

Kardan, S., & Conati, C. (2012). Exploring gaze data for determining user learning with an interactive simulation. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization* (pp. 126-138). Springer Berlin Heidelberg.

Kardan, S., & Conati, C. (2013). Comparing and combining eye gaze and interface actions for determining user learning with an interactive simulation. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization* (pp. 215-227). Springer Berlin Heidelberg.

Kardan, S., & Conati, C. (2015). Providing Adaptive Support in an Interactive Simulation for Learning: An Experimental Evaluation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* (pp. 3671-3680). ACM.

Kareal, F., & Klema, J. (2006). Adaptivity in e-learning. In *A. Méndez-Vilas, A. Solano, J. Mesa and JA Mesa: Current Developments in Technology-Assisted Education, 1,* 260-264.

Katidioti, I., Borst, J. P., & Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of Experimental Psychology: Applied, 20(4),* 380.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12(3),* 281-300.

Kim, J., Thomas, P., Sankaranarayana, R., & Gedeon, T. (2012). Comparing scanning behaviour in web search on small and large screens. In *Proceedings of the Seventeenth Australasian Document Computing Symposium* (pp. 25-30). ACM.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H. J. (2015). Eye-tracking analysis of user behavior and performance in web search on large and small screens. *Journal of the Association for Information Science and Technology, 66(3),* 526-544.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*: DTIC Document.

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In *M. J. Snowling & C. Hulme (Eds.), The Science of Reading: A Handbook*: Blackwell Publishing.

Kirschner, P. A., & Karpinski, A. C. (2010). Facebook® and academic performance. *Computers in human behavior, 26(6),* 1237-1245.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium onEye tracking research & applications.* (pp. 69-72). ACM.

Kozek, K. K. (1997). *Classification of eye tracking data using hidden markov models* (Honours thesis). University of New South Wales, NSW, Australia.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology, 77(6),* 1121.

Lach, P. (2013). Intelligent Tutoring Systems Measuring Student's Effort During Assessment. In *Canadian Conference on Artificial Intelligence* (pp. 346-351): Springer.

Lallé, S., Conati, C., & Carenini, G. (2016). Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data. In *Proceedings of the 25th International Joint Conference in Artificial Intelligence.*

Lallé, S., Toker, D., Conati, C., & Carenini, G. (2015). Prediction of Users' Learning Curves for Adaptation while Using an Information Visualization. In *Proceedings of the 20th International Conference on Intelligent User Interfaces.* (pp. 357-368). ACM.

Lankford, C. (2000). Effective eye-gaze input into windows. In *Proceedings of the* 2000 *symposium on Eye tracking research & applications.* (pp. 23-27). ACM.

Little, J. L., & Bjork, E. L. (2012). Pretesting with multiple-choice questions facilitates learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society.* (pp. 23-27). ACM.

Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation, 61(6),* 700-712.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in cognitive sciences, 4(1),* 6-14.

Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: implications for theory and practice. *Studies in Higher Education, 27(1),* 27-52.

Loboda, T. D., Brusilovsky, P., & Brunstein, J. (2011). Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th international conference on Intelligent user interfaces.* (pp. 175-184). ACM.

Longmore, M. A., Dunn, D., & Jarboe, G. R. (1996). Learning by doing: Group projects in research methods classes. *Teaching Sociology*, 84-91.

Maglio, P. P., & Campbell, C. S. (2003). Attentive agents. *Communications of the ACM, 46(3)*, 47-51.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58(0),* 61-68.

Mansfield, J. S., Legge, G. E., & Bane, M. C. (1996). Psychophysics of reading. XV: Font effects in normal and low vision. *Investigative Ophthalmology & Visual Science, 37(8),* 1492-1501.

Marshall, C. C., & Bly, S. (2005). Turning the page on navigation. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, (JCDL'05).* (pp. 225-234). IEEE.

Martinez-Conde, S. (2006). Fixational eye movements in normal and pathological vision. *Progress in brain research, 154*, 151-176.

Martínez-Gómez, P., & Aizawa, A. (2014). Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th international conference on Intelligent User Interfaces.* (pp. 95-104). ACM.

Mason, C., & Kandel, E. R. (1991). Central visual pathways. *Principles of neural science, 3*, 420-439.

Mason, L., Pluchino, P., & Tornatora, M. C. (2015). Eye-movement modeling of integrative reading of an illustrated text: Effects on processing and learning. *Contemporary Educational Psychology, 41*, 172-187.

Mayer, R. E. (1999). Research-based principles for the design of instructional messages: The case of multimedia explanations. *Document design, 1(1)*, 7-19.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93(1)*, 187.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial fixations on words. *Vision Research, 28(10)*, 1107-1118.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1989). Eye movement control during reading: II. Frequency of refixating a word. *Perception & Psychophysics, 46(3)*, 245-253.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Attention, Perception, & Psychophysics, 17(6)*, 578-586.

McKay, D. (2011). A jump to the left (and then a step to the right): reading practices within academic ebooks. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference, Canberra, Australia.* (pp. 202-210). ACM.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*: Cambridge University Press.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). *Coh-Metrix version 3.0*. Retrieved 30th July 2015, from http://cohmetrix.com

Mehigan, T. (2014). Chapter Four Assessing Eye-Tracking Technology For Learning-Style Detection. In *Adaptive Game-Based Learning Tracey Mehigan And Ian Pitt. Game-Based Learning: Challenges and Opportunities, 77*.

Mehigan, T., & Pitt, I. (2013). Intelligent mobile learning systems for learners with style. In *Tools for Mobile Multimedia Programming and Development (May 2013), D. Tjondronegoro, Ed., IGI-Global,* 131-149.

Mehigan, T. J. (2013). *Automatic detection of learner-style for adaptive eLearning*.

Mehigan, T. J., Barry, M., Kehoe, A., & Pitt, I. (2011). Using eye tracking technology to identify visual and verbal learners. In *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo (ICME).* (pp. 1-6). IEEE.

Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional blindness. *Consciousness and Cognition, 15(3)*, 620-627.

Mendis, B. S. U., & Gedeon, T. D. (2008, 1-6 June 2008). A comparison: Fuzzy signatures and Choquet Integral. In *Proceedings of the IEEE International Conference on Fuzzy Systems, 2008 (FUZZ-IEEE 2008).* (pp. 1464-1471). IEEE.

Merten, C., & Conati, C. (2006). Eye-tracking to model and adapt to user meta-cognition in intelligent learning environments. In *Proceedings of the 11th international conference on Intelligent user interfaces.* (pp. 39-46). ACM.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63(2),* 81.

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks, 6(4),* 525-533.

Moresi, S., Adam, J. J., Rijcken, J., Van Gerven, P. W. M., Kuipers, H., & Jolles, J. (2008). Pupil dilation in response preparation. *International Journal of Psychophysiology, 67(2),* 124-130.

Morimoto, C. H., & Mimica, M. R. M. (2005). Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding, 98(1),* 4-24.

Murata, A. (2006). Eye-gaze input versus mouse: Cursor control as a function of age. *International Journal of Human-Computer Interaction, 21(1),* 1-14.

Nugrahaningsih, N., Porta, M., & Ricotti, S. (2013). Gaze behavior analysis in multiple-answer tests: An Eye tracking investigation. In *Proceedings of the International Conference on Information Technology Based Higher Education and Training (ITHET), 2013.* (pp. 1-6). IEEE.

O'Hara, K., & Sellen, A. (1997). A comparison of reading paper and on-line documents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 335-342). ACM.

O'Regan, J. K. (1981). The convenient viewing position hypothesis Eye movements: *Cognition and visual perception* (pp. 289-298): Erlbaum.

O'Regan, J. K. (1984). How the Eye Scans Isolated Words. In *A. G. Gale & F. Johnson (Eds.), Theoretical and Applied Aspects of Eye Movement Research Selected/Edited Proceedings of The Second European Conference on Eye Movements* (Vol. 22, pp. 159 - 168): North-Holland.

Okoso, A., Toyama, T., Kunze, K., Folz, J., Liwicki, M., & Kise, K. (2015). Towards Extraction of Subjective Reading Incomprehension: Analysis of Eye Gaze Features. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, Republic of Korea.* (pp. 1325-1330). ACM.

Oluleye, B., Armstrong, L., Leng, J., & Diepeveen, D. (2014). A genetic algorithm-based feature selection. *British Journal of Mathematics & Computer Science, 4(21),* pp. 889-905.

Ozcelik, E., Arslan-Ari, I., & Cagiltay, K. (2010). Why does signaling enhance multimedia learning? Evidence from eye movements. *Computers in human behavior, 26(1),* 110-117.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science, 32(1),* 1-8.

Paramythis, A., & Loidl-Reisinger, S. (2003). Adaptive learning environments and e-learning standards. In *Proceedings of the Second European Conference on e-Learning.* (Vol. 1, pp. 369-379).

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia, 4(2),* 1883.

Pollatsek, A., & Rayner, K. (2009). Reading. In *L. R. Squire (Ed.), Encyclopedia of Neuroscience.* Oxford: Academic Press.

Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI.* (pp. 542-546).

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*. Pennsylvania: Idea Group, Inc.

Porta, M. (2008). Implementing eye-based user-aware e-learning. In *Proceedings of the CHI'08 Extended Abstracts on Human Factors in Computing Systems*. (pp. 3087-3092). ACM.

Prusty, B. G., & Russell, C. (2011, 21 - 26 August 2011). Engaging students in learning threshold concepts in engineering mechanics: adaptive eLearning tutorials. In *Proceedings of the International Conference on Engineering Education (ICEE2011)*, University of Ulster, Belfast, Northern Ireland, UK.

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001). Types of eye movements and their functions. *Neuroscience. 2nd edition*: Sunderland (MA): Sinauer Associates.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1(1),* 81-106.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning: Elsevier.*

Ramakrisnan, P., Jaafar, A., Razak, F. H. A., & Ramba, D. A. (2012). Evaluation of user Interface Design for Leaning Management System (LMS): Investigating Student's Eye Tracking Pattern and Experiences. *Procedia-Social and Behavioral Sciences, 67,* 527-537.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 372-422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology, 62(8),* 1457-1506.

Rayner, K., & Bertera, J. H. (1979). Reading without a fovea. *Science, 206(4417),* 468-469.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10(3),* 241-255.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research, 16(8),* 829 - 837.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105(1),* 125-157.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research, 7(1),* 4-22.

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the EZ Reader model. *Vision Research, 39(26),* 4403-4411.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences, 26(4),* 445-476.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2012). Eye movements in reading versus nonreading tasks: Using EZ Reader to understand the role of word/stimulus familiarity. *Visual cognition, 20(4-5),* 360-390.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16(1),* 1-21.

Rho, Y. J., & Gedeon, T. D. (2000). Academic articles on the web: reading patterns and formats. *International Journal of Human-Computer Interaction, 12(2),* 219-240.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL-International Review of Applied Linguistics in Language Teaching, 45(3),* 193-213.

Rockinson- Szapkiw, A. J., Courduff, J., Carter, K., & Bennett, D. (2013). Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education, 63(0),* 259-266.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33(1),* 1-39.

Rokach, L., & Maimon, O. (2005). *Clustering methods Data mining and knowledge discovery handbook* (pp. 321-352): Springer.

Romer, D. (1993). Do Students Go to Class? Should They? *The Journal of Economic Perspectives, 7(3),* 167-174.

Rosch, J. L., & Vogel-Walcutt, J. J. (2013). A review of eye-tracking applications as tools for training. *Cognition, technology & work, 15(3),* 313-327.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20,* 53-65.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., & Edwards, D. D. (2003). *Artificial intelligence: a modern approach* (Vol. 2): Prentice hall Upper Saddle River.

Salojarvi, J., Puolamaki, K., Simola, J., Kovanen, L., Kojo, I., & Kaski, S. (2005). Inferring relevance from eye movements: Feature extraction: In *Workshop at NIPS 2005, in Whistler, BC, Canada, on December 10, 2005*.(p. 45).

Salvucci, D. D. (1999). Inferring intent in eye-based interfaces: tracing eye movements with process models. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit* (pp. 254-261). ACM.

Salvucci, D. D., & Anderson, J. R. (1998). Tracing eye movement protocols with cognitive process models. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp 923-928).

Salvucci, D. D., & Anderson, J. R. (2001). Automated Eye-Movement Protocol Analysis. *Human–Computer Interaction, 16(1),* 39-86.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications* (pp. 71-78). ACM.

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & cognition, 34(2),* 344-355.

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., & Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and

reward during decision making: Evidence from pupillometry and event-related fMRI. *Neuroimage, 37(3),* 1017-1031.

Scherr, K. C., Agauas, S. J., & Ashby, J. (2015). The Text Matters: Eye Movements Reflect the Cognitive Processing of Interrogation Rights. *Applied Cognitive Psychology, 30*, 234–41.

Schroder, M., Bogdan, M., Hinterberger, T., & Birbaumer, N. (2003). Automated EEG feature selection for brain computer interfaces. In *Proceedings of the First International IEEE EMBS Conference on Neural Engineering, 2003.* (pp. 626-629). IEEE.

Schwarz, U., & Schmückle, T. (2002). Cognitive Eyes. *Schweizer Archiv Für Neurologie Und Psychiatrie, 153(4),* 175-179.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 40(1),* 185-197.

Sharma, N., & Gedeon, T. (2011). Stress classification for gender bias in reading. In *Proceedings of Neural Information Processing*. (pp. 348-355). Springer Berlin Heidelberg.

Sharma, N., & Gedeon, T. (2012). Artificial neural network classification models for stress in reading. In *Proceedings of Neural Information Processing*. (pp. 388-395). Springer Berlin Heidelberg.

Sharma, N., & Gedeon, T. (2013a). Computational Models of Stress in Reading Using Physiological and Physical Sensor Data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 111-122). Springer.

Sharma, N., & Gedeon, T. (2013b). Hybrid genetic algorithms for stress recognition in reading. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. (pp. 117-128). Springer Berlin Heidelberg.

Sharmin, S., Spakov, O., & Raiha, K.-J. (2012). The effect of different text presentation formats on eye movement metrics in reading. *Journal of Eye Movement Research, 5(3),* 1-9.

Shibata, H., Takano, K., Omura, K., & Tano, S. i. (2015). Page Navigation on Paper Books and Electronic Media in Reading to Answer Questions. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, Parkville, VIC, Australia.* (pp. 526-534). ACM.

Sibert, J. L., Gokturk, M., & Lavine, R. A. (2000). The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th annual ACM symposium on User interface software and technology, San Diego, California, United States.* (pp. 101-107). ACM.

Sibley, C., Coyne, J., & Baldwin, C. (2011). Pupil Dilation as an Index of Learning. In *Proceedings of the human factors and ergonomics society 55th Annual meeting*, (Vol. 55, No. 1, pp. 237-241). SAGE Publications.

Siedlecki, W., & Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters, 10(5),* 335-347.

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.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28(9),* 1059-1074.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5(4),* 644-649.

Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*: Rand Corporation.

Sörqvist, P., Halin, N., & Hygge, S. (2010). Individual differences in susceptibility to the effects of speech on reading comprehension. *Applied Cognitive Psychology, 24(1),* 67-76.

Spada, D., Sánchez-Montañés, M., Paredes, P., & Carro, R. M. (2008). Towards inferring sequential-global dimension of learning styles from mouse movement patterns. In *Proceedings of the Adaptive Hypermedia and Adaptive Web-Based Systems.* (pp. 337-340). Springer Berlin Heidelberg.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science, 333(6043),* 776-778.

Specht, M., Kravcik, M., Klemke, R., Pesin, L., & Hüttenhain, R. (2006). Adaptive learning environment for teaching and learning in WINDS. In *Proceedings of the Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 572-575). Springer Berlin Heidelberg.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. *The Oxford handbook of psycholinguistics, 327*, 342.

Steichen, B., Conati, C., & Carenini, G. (2014). Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Transactions on Interactive Intelligent Systems, 4(2),* 1-29.

Sung, E., & Mayer, R. E. (2012). When graphics improve liking but not learning from online lessons. *Computers in human behavior, 28(5),* 1618-1625.

Surjono, H. D. (2011). The design of adaptive e-Learning system based on student's learning styles. *International Journal of Computer Science and Information Technology (IJCSIT), 2(5),* 2350-2353.

Surjono, H. D. (2014). The Evaluation of a Moodle Based Adaptive e-Learning System. *International Journal of Information and Education Technology, 4(1)*: 89.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review, 10(3)*, 251-296.

Toker, D., & Conati, C. (2014). Eye Tracking to Understand User Differences in Visualization Processing with Highlighting Interventions. In *V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings* (pp. 219-230). Cham: Springer International Publishing.

Traphagan, T., Kucsera, J., & Kishi, K. (2010). Impact of class lecture webcasting on attendance and learning. *Educational Technology Research and Development, 58(1),* 19-37.

Tsai, M.-J., Hou, H.-T., Lai, M.-L., Liu, W.-Y., & Yang, F.-Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education, 58(1),* 375-385.

Underwood, G., & Batt, V. (1996). *Reading and Understanding*. Massachusetts, USA: Blackwell Publishers.

Underwood, G., Hubbard, A., & Wilkinson, H. (1990). Eye fixations predict reading comprehension: The relationships between reading skill, reading speed, and visual inspection. *Language and speech, 33(1),* 69-81.

Victor, T. W., Harbluk, J. L., & Engström, J. A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour, 8(2),* 167-190.

Vo, T., Mendis, B. S. U., & Gedeon, T. D. (2010). Gaze Patterns and Reading Comprehension. In *International Conference on Neural Information Processing* (pp. 124-131). Springer Berlin Heidelberg.

Waniek, J., & Ewald, K. (2008). Cognitive costs of navigation aids in hypermedia learning. *Journal of Educational Computing Research, 39(2),* 185-204.

Welsh, E. T., Wanberg, C. R., Brown, K. G., & Simmering, M. J. (2003). E-learning: emerging uses, empirical results and future directions. *International Journal of Training and Development, 7(4),* 245-258.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing, 4(2),* 65-85.

Whitley, D. (2001). An overview of evolutionary algorithms: practical issues and common pitfalls. *Information and software technology, 43(14),* 817-831.

Woodfield, R., Jessop, D., & McMillan, L. (2006). Gender differences in undergraduate attendance rates. *Studies in Higher Education, 31(1),* 1-22.

Xu, R., & Wunsch, D. (2008). *Clustering* (Vol. 10): John Wiley & Sons.

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection* (pp. 117-136): Springer.

Yarbus, A. (1967). *Eye Movements and Vision*. New York: Plenum Press.

Yatabe, K., Pickering, M. J., & McDonald, S. A. (2009). Lexical processing during saccades in text comprehension. *Psychonomic Bulletin and Review, 16,* 62-66.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: neural correlates of pupil size as a measure of cognitive listening load. *Neuroimage, 101*, 76-86.

Zhang, L., Liu, Z., & Ni, J. (2013). Feature-Based Assessment of Text Readability. In *2013 Seventh International Conference on Internet Computing for Engineering and Science* (pp. 51-54). IEEE.

Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class adaboost. *Statistics and its Interface, 2(3),* 349-360.

Appendix A

Materials for Experiment 1 - Eye Gaze in eLearning Environments

This appendix includes the supporting documentation and resources that were used for the first experiment explained in this thesis – Eye Gaze in eLearning Environments. The resources included the participant information sheet, the consent form, the texts and questions used for the experiment, and the pre- and post-experiment questionnaires.

Ethics approval was sought from the Australian National University Research Ethics Committee before the experiment was conducted. The experiment was conducted under Human Ethics Protocol 2012/006. The participant information sheet and consent form was designed according to the requirements of the ethics approval.

A.1 Participant Information Sheet

Participant Information

Project Title: Analysis and Computational Modelling of Reading

Investigators:

- **Ms. Leana Copeland** (0422521154, leana.copeland@anu.edu.au)

Introduction
This document concerns a user study to be conducted at the Research School of Computer Science at the Australian National University. The study requires that participants read through a set of paragraphs on a subject and then answer a set of questions based upon what was just read. During this time user behaviour will be recorded.

What would be involved?
The time needed to complete this user study will be about 60 minutes. This time will include an introduction to the tasks, setup, and completion of the tasks mentioned above.

Data Collection and Contact Details
The main purpose of the user study is to collect some data to enable useful information to be gained on the interface, the interaction techniques, and tasks. We will give you a pre- and post-task questionnaire that may contain some questions of an identifying nature. You do not need to complete these or any of the other questions if you have any objections to them. The data from the experiment will be made unidentifiable to retain privacy of each participant. Until that time, if you give your permission, your contact details will be retained for follow-up testing.

Data Use
The data collected will be used to draw conclusions about certain interaction techniques and the nature of the tasks. Any data collected, either raw or processed, may be used in a thesis and other research and publications. The data will be made unidentifiable so that no participant will be able to be identified from any data collected.

Risks
As the study is conducted in a carefully designed lab environment, all care will be taken to make participants as comfortable as possible, given the nature of the interaction tasks. Some physical discomfort such as eye and muscle strain may occur with some people ***including, in rare cases, motion sickness. Participants are free to request that your participation in the user study cease at any stage without explanation.***

Your rights
You may ask for a copy of any data collected or research publications written. You may also end the test session or ask for a break at any time and request that any or all data collected from you be destroyed. You have the right to completely withdraw from the experiment at any point. You can ask that your name be deleted from our contact list for future testing at any time.

This usability study has nothing to do with assessment and participation or non-participation will not directly affect your assessment in any course at ANU and it is completely voluntary. If you have any concerns with the ethics of this study please contact the ANU ethics committee by emailing Human.Ethics.Officer@anu.edu.au or calling 6125 2900.

A.2 Participant Consent Form

Consent Form

1. I have had the project explained to me, and I have read the Participant Information sheet.
2. I agree to participate in the user study as described.
3. I acknowledge that:
    a) I understand that my participation is voluntary and that I am free to withdraw from the user study at any time and to withdraw any unprocessed data previously supplied (unless follow-up is needed for safety).
    b) The user study is for the purpose of research. It may not be of direct benefit to me.
    c) The privacy of the personal information I provide will be safeguarded and only disclosed where I have consented to the disclosure or as required by law.
    d) The security of the research data will be protected during and after completion of the study. The data collected during the user study may be published in this and subsequent research. Any information that may identify me will not be used.

Please list any Special Considerations (e.g. any medical conditions you have which you would like to bring to the attention of the user study supervisor)

_____
_____
_____

Participant's Name: _____

Signature: _____ Date: _____

A.3 Experiment texts

Original lecture notes written by: Nandita Sharma

Tutorial texts written by: Leana Copeland

A.3.1 The World Wide Web

The World Wide Web (WWW), or colloquially the Web, is a widely used information system that enables locating and viewing of a variety of multimedia based files including text documents, audio, visual and graphic files.

Sir Tim Berners-Lee wrote a proposal in 1989 based on earlier concepts of hypertext systems for what eventually became the Web. It was Berners-Lee that built the first web browser, web server and web pages, which are the main components of the Web, and he is now the Director of the Web Consortium (W3C), which is the main international standards organization for the Web.

The Web is essentially a big graph made up of billions of web pages and hyperlinks. A Web page is a document or information that can be viewed using a web browser. Web pages can contain content such as text, images, videos, audio, as well as hyperlinks, which enable navigation to other Web pages. Web pages are generally formatted in HyperText Markup Language (HTML). HTML provides the ability to embed images, create interactive forms, and a means of structuring documents into headings, paragraphs, lists, links, and so on. Although some formatting and presentation of information can be handled by HTML, it is generally the Cascading Style Sheets (CSS) that are used to define the appearance and layout of the web pages.

Scripts can be embedded into HTML that affect the behaviour of a Web page. This allows the content of Web pages to be dynamically generated. These are termed dynamic Web pages and refer to Web content that is based on user input. Examples of these types of Web pages are on websites for flight status or stock exchange rates. Usually dynamic Web pages are assembled at the time of a request from a browser and typically their URL has a "?" character in it. Scripts to create dynamic Web pages can be written in languages such as Javascript and Ajax.

Web pages are requested and served from Web servers using the Hypertext Transfer Protocol (HTTP). For example, when you enter a Uniform Resource Locator (URL) in your browser, this actually sends an HTTP request command to a Web server directing it to fetch and transmit the requested Web page. HTTP is an application layer protocol designed within the framework of the Internet protocol suite. This means that it presumes there is an underlying transport layer protocol such as the Transmission Control Protocol (TCP).

A.3.2 The Importance of Search Engines

The Web is popular. Every day the number of Web pages on the Web increases. The Indexed Web contained at least 15.2 billion pages as of Wednesday, 13 February, 2013. Similarly, the number of Internet hosts connected to the Internet increases, with close to 1 billion hosts as of July, 2012. The content on the Web is rapidly changing and expanding and there are users of the Web all over the world. This also means that the Web is full of information in different languages. There is no central coordination over content, presentation or location of Web pages. Most web pages are titled by their author and are located on servers with cryptic names. With the vast number of resources that are scattered in an ad hoc way, located in different locations, and in no order, how does anything get found? This is where search engines come in. Web search engines are designed to search the information on the Web based upon keywords that the user enters into their interface and return a list of results referred to as search engine results pages (SERP's). Search engines essentially make the content of the Web accessible and they make the web seem organised to the user.

There are numerous search engines and they are often specialised to perform certain searches. The major search engines are Google, Yahoo!, Bing and Ask. There are also different types of search,

including text search, audio search, location-based search, image recognition search, barcode search, and many more. Of the uses of the Web, searching the Web is very popular. Other uses of the web include social networking, accessing news, sending and receiving email, online shopping, and many more. However, the fact remains that people often use search engines to first find their way to one or more of these other uses. Furthermore, there is no incentive in creating content on the Web unless it can be easily found.

Other methods of finding Web pages exist, such as web directories, taxonomies and bookmarks, but have not kept up the pace of search engines to perform large and very fast searches of the Web. There are also answer engines that are a type of search engine that answers natural language queries directly by computing an answer from structured data as opposed to returning a list of the most suited web pages for queries.

A.3.3 Brief non-technical History of Search Engines

During the early development of the Web a manual list was kept of the Web servers but as the number of Web servers grew, the central list could not keep up. The first search engine on the Internet was called Archie, which was created in 1990. The name stands for "archive" without the "v" and was a database of file names that could be searched manually rather than be indexed.

The Gopher protocol was created in 1991 by Mark McCahill, which led to two new search programs, Veronica and Jughead. These programs searched the file names and titles stored in Gopher index systems. The Gopher protocol is a TCP/IP application layer protocol designed for distributing, searching, and retrieving documents over the Internet. Gopher was eventually superseded by HTTP.

The W3Catalog was the first primitive search engine for the Web, which was released in 1993. One of the first publicly available crawler-based search engines called WebCrawler was introduced to the Web In 1994. The difference between this search engine and its predecessors was that it allowed users to search for any word in any Web page, which has become the standard for all major search engines today.

There was a rapid emergence of search engines in the 1990's with search engines such as, Yahoo!, Lycos, Magellan, Excite, Infoseek, Inktomi, Northern Light, and AltaVista. By the end of this period, search engines had begun to adopt the use of paid placement rankings and the selling of search terms. This move made search engine companies one of the most profitable businesses on the Internet at the time.

The Google search engine rose to prominence around 2000 and has remained the most popular search engine. Up until Google's search engine, the conventional method of ranking search results was by counting the number of times a search term appeared on a web page. However, Google's search engine employed the use of the PageRank algorithm to rank its search results. PageRank is a ranking system where the number of pages and the importance of those pages that linked back to the original site determine a website's relevance.

In 2012, Google released Open Drive, which is a file search engine that enables files stored in cloud storage that are publically available. Open Drive will return search results from cloud storage content services including Google Drive, Dropbox, SkyDrive, Evernote and Box.

A.3.4 Web Search Basics

A Web search engine is a program that is designed to search for information on the Web for a user query and return a set of results to the user. In short, a Web search engine performs the following tasks: Web crawling, indexing, calculating relevancy and rankings, and serving results back to the user. Web search engines need to store information about a lot of Web pages for effective and efficient search. They get this information by Web crawlers that record information from the HTML of a Web page and follow every link from the Web page. The data collected by the Web crawlers about Web pages they visit get stored in an index database so that it can be used when users make queries.

To serve its users, Web search engines must take user input in the form of a query, which are usually keywords that they wish to find information or Web sites on. The search engine examines its index and then provides a listing of the most suitable Web pages given the search criteria the user has given as input. The results of a text-based search are the document's title as well as a short excerpt of text from the page or document, or an image in the case of image search, or a location in the case of location based search.

Of course, just because the keywords that a user has queried appear on a Web page does not mean that it is an appropriate result to return to the user because some other pages may be more relevant or reliable than others. Search engines rank results using different ranking methods before they are returned to the user so that the most relevant results are returned towards the beginning of the list or ranked higher in the search results.

Advertising revenue to some extent supports most commercial search engines and is what made a lot of search engine companies quite profitable. Search engines allow advertisers to pay to have their listings ranked higher in search results. Also, search engines feature related ads next to the search engine results for a query. Every time a user clicks on one of these ads the search engine is paid. This way search engines can maintain their credentials with their users and the advertisers – users get their search results and advertisers have their ads placed towards the best search results.

A.3.5 Web Crawling

Web crawling is the first step that a search engine takes to return results of a search query to a user. This step is invisible and most people do not know that it exists. This is the step in which a search engine identifies that a file or document exists. Simple automated programs or scripts, colloquially called Web spiders and crawlers, perform Web crawling whereby a list of words and notes about where they were found is generated. These Web spiders build lists of words found on Web pages by methodically scanning through web pages and creating an index from the information they scanned. Web crawlers are not only used by search engines, but are used by linguists and market researchers or anyone trying to find information from the Internet in an organised manner.

Web crawlers usually start at popular sites and servers where they index the words on the pages and follow every link within the site. The crawler eventually builds an index based on its own system of weighting. For example, words in titles or headings may be deemed more important. This data is encoded to save space and stored for users to access through search queries.

There are limits to how much a web crawler can download at any one time and given that there is a large amount of rapidly changing data web crawlers have access to, the behaviour of a web crawler can determine how efficient and how up-to-date the information that is collected and stored. There are several policies that contribute to the behaviour of a web crawler. The selection policy of a web crawler determines which pages to download and the re-visit policy determines when the web crawler checks for changes to the pages. The politeness policy determines how the Web crawlers avoid overloading Web sites, and the parallelization policy coordinates distributed Web crawlers.

Different search engines employ Web crawlers that record different types of words on Web pages. Different approaches are usually an attempt to make the spider operate faster, allow users to search more efficiently, or both. For example, some Web crawlers will keep track of the words in the title, sub-headings and links while others will keep track of the 100 most frequently used words on the page. The early Google search engine was built with only a few crawlers that could keep around 300 connections to Web pages open at any one time.

A.3.6 Building an Index

A Web search engine must store the information that is constantly being collected by web crawlers so that it is accessible to users when they make queries. It would be neither computationally efficient nor fast for a search engine to scan every page in its collection of crawled pages. Instead a search engines

index is a compact storage of Web information that is designed to optimize the speed and performance in finding results for a search query.

There are many challenges in search engine indexing which centre on the fact that the Web has an enormous amount of data that is constantly changing. So indexes must be designed to maintain efficient indexing, fast retrieval and compact storage. This motivates the index policy to consider which pages should be indexed and the extent to which these pages are indexed. Often information such as how many times the word appears on the page, whether the word was just used in a trivial way, and whether there are links from that page to other pages containing the word are stored in the index along with the words and URLs. This additional information is used to assign a weight to each entry in the index that is later used in ranking of the search results. Search engines have different methods of assigning weight to entries but an example is that higher weights are assigned when the word appears in the title, sub-headings or in links of the document. Some search engines store all or part of the source page as well as information about the page whilst others store every single word on the pages they crawl.

There are many factors that affect the design of a search engine's index such as how information is entered into the index and how and if information is compressed or filtering to reduce the storage size. Lookup speed of finding an entry in the index, as well as update and removal speeds are another factor that affect the design of search engine indexes. Furthermore, maintenance and fault tolerance are also considered in designing the index. The method of index storage also plays an important role in how search engines perform indexing and although there are many types of data structures that a search engine could be built from, a common web search engine index structure is the inverted index i.e. a hash table.

A.3.7 User Queries

Once the search engine has built an index of the parts of the web that its web crawlers have explored, users can submit queries to find information from that part of the web. The query submitted to the search engine actually queries the index that was built by the web crawlers. Queries can be quite simple, such as one word, or quite complex to make a query more specific. Boolean operators such as AND, OR and NOT can be used to make a query more specific. The AND operator allows the user to specify that they want all the words joined by the AND to be present in the results. The OR operator allows the user to specify that they want at least one of the words joined by the OR to appear in the results. The NOT operator allows the user to define terms that they do not want to appear in the results. Searches of this kind are termed literal searches because the search engine looks for the words or phrases exactly as they are queried.

There are additional advanced queries that can be made such as searching for an exact word or phrase, which in Google search is denoted by "search query", or finding words similar to a query term, again denoted in Google search as ~query term. There are also wildcard characters that are used as placeholders for unknown terms, i.e. fill in the blank, and are denoted as * in Google search. Furthermore, searching directly within a site or domain directly, which in Google search is denoted query term site: site or domain.

There are searches that are concept-based which involve using statistical analysis on the pages that contain the words or phrases that were queried in order to find the pages that the user would be more interested in. The problem with this approach is that it requires more information to be stored about the crawled pages and the processing time for a search of this type would be longer.

Furthermore, there are natural language based searches that are based on the premise that the user types a question as a query. The question is structured the same as if you were asking another human and hence a natural approach. An example of this is Wolfram Alpha, which takes questions of many different forms, such as mathematical equations, and returns a computed answer rather than a list of results.

A.3.8 Calculating Relevancy and Rankings

When the user inputs a query to a search engine they expect the most relevant web pages to be returned in the list of search results. However, relevance means more than simply finding a page that contains some keywords, as some pages may be more appropriate, popular, or authoritative than others. Often how useful a search engine is considered is dependent upon the relevance of the search result set it gives back. Search engines risk losing users, to other providers and to offline methods if they cannot provide relevant results, this is why search engines rank search results.

The search results with the highest rankings are deemed most relevant and are presented at the top of the page to the user. There are many factors that affect how the search engine calculates relevance and hence ranking of pages. Some of these factors include: page content, frequency and location of keywords within a page, age of the page, number of pages linking to the page, discovery of additional sites, updates made to indices, changes to the search algorithm, and many more. The rankings each search engines uses are different which is why submitting the same query into several different search engines will return different results. However most search engines have a few things in common such as the more popular a site is the more important it must be, as well as, the location and frequency of keywords on the web pages. So the more popular a site, page or document is, the more valuable the information must be.

Furthermore, adverts are included in the search results, which also must be appropriately matched to what was searched by the user. This means that if a user searches for cars they will be presented with adverts that are about products related to cars and not products relating to boats.

It has long been established that users generally tend to look at the first page of search results and generally gravitate to the top of the first page of search results. A sample of over 8 million clicks showed that over 94% of users clicked on a first page result with the first spot being the clicked the most. This fact motivates search engines to order search results with the "better" pages at the top. However, it also motivates designers of web pages to optimise how search engines can find their web pages.

A.3.9 Search Engine Optimisation

Search engines allow companies or individuals to pay to have their Web pages placed at the top of search results for certain queries. This is not the only way to ensure a web page tops the search results for a given query. The alternative is to use search engine optimisation (SEO) which is the process of tuning a website or web page to rank higher in the search results for certain queries and hence increase visibility and visitors to the site.

SEO considers how search engines work, what people search for, and which search terms are used. The first step of SEO is to get indexed by a leading search engine. These search engines use web crawlers to find pages and offer either free or paid submission of pages. Web crawlers look at a number of different factors when crawling a site so the search engines index not every page. However, web crawlers intentionally avoid some content because the owner has specified for it not to be indexed. This is done through the robots.txt file in the root directory of the domain. Typically pages such as shopping carts and user-specific content are prevented from being indexed because search engines such as Google consider those pages as search spam. Finally, there are a number of ways to increase the visibility of a webpage within the search results, such as by cross-linking web pages on a website to provide more links to most important pages. Other methods include writing content that includes frequently searched keyword phrase because that will make the page or site relevant to a wider variety of search queries. Also, updating content to give additional weight to a site because web crawlers will have to visit the site or page more frequently.

There are two categories of SEO techniques, which are term white hat and black hat SEO. The white hat SEO techniques are ones that are approved and recommended by a search engines guidelines, and involves no deception. Using white hat techniques tends to produce results that will last longer. Black hat SEO techniques are techniques of which search engines do not approve and involve deception. An

example of black hat SEO is hidden text, which is either text coloured similar to the background or positioned off screen. Once the search engine realises that a site is using black hat techniques it may be banned either temporarily or permanently.

A.4 Web Search Tutorial Quiz

The quiz for the web search tutorial to test understanding consists of 18 questions. There are 2 questions for each tutorial heading/slide, one of which is a multiple-choice question and the second, which is a cloze (fill-in-the-blanks) question.

Questions for "The World Wide Web" Slide:

1. What are Web pages formatted in and what protocol are they transmitted in?
    a) Cascading Style Sheet (CSS) and Hypertext Transfer Protocol (HTTP), respectively.
    b) HyperText Markup Language (HTML) and Hypertext Transfer Protocol (HTTP), respectively.
    c) JavaScript and Transmission Control Protocol (TCP), respectively.
    d) HyperText Markup Language (HTML) and Transmission Control Protocol (TCP), respectively.
2. Web pages that are generated based upon user input are called _____(dynamic/interactive) web pages. These types of web pages have _____(scripts/forms) embedded into the HTML.

Questions for "The Importance of Search Engines" Slide:

3. Why are Web search engines important?
    a) They are the only way to find web pages on the Web.
    b) They control the location and presentation of the content on the Web
    c) They control the information of the Web
    d) They make the content of the Web accessible and seem organised to the user
4. Web search engines are designed to search the information on the Web based upon _____(keywords/queries/search terms) that the user enters into their interface and return a list of results.

Questions for "Brief non-technical history of Search Engines" Slide:

5. Open Drive is a file search engine that enables files stored in _____(cloud) storage that are publically available to be searched.
6. The conventional method of ranking results was by counting the number of times a search term appeared on a web page. What changed this?
    a) PageRank
    b) Web crawlers
    c) Archie
    d) Gopher Protocol

Questions for "Web Search Basics" Slide:

7. What is the correct order that Web search engines perform their four main tasks:
    a) Web crawling; Indexing; serving results; calculating relevancy and rankings
    b) Indexing; Web crawling; Serving results; calculating relevancy and rankings
    c) Web crawling; Indexing; calculating relevancy and rankings; serving results
    d) Indexing; Web crawling; Calculating relevancy and rankings; serving results

8. Web search engines gain _____(revenue/money/profit) by allowing advertisers to pay to rank their websites higher in search results and by running related ads _____(next) to the search results.

Questions for "Web Crawling" Slide:

9. Web crawlers are automated programs responsible for methodically scanning through _____(Web pages/Web sites) and creating an _____(index) of information so that users can make queries on it later.
10. Which is false:
    a) Web crawlers are also called spiders.
    b) Web crawlers start at the home page of the search engine.
    c) Web crawlers build lists of words found on a web page.
    d) Web crawler behaviour is dictated by a set of policies.

Questions for "Building an Index" Slide:

11. Which of these is not a factor that affects how a search engine's index is designed?
    a) Relevancy ranking within the index
    b) How data is entered into the index
    c) The data structure used to build the index
    d) Fault tolerance of the index
12. There are many ways to build an index but the main purpose for search engines is to provide _____(compact/efficient) storage and _____(quick/fast/efficient/rapid/optimised) retrieval of the information.

Questions for "User Queries" Slide:

13. When a search engine looks for the words or phrases exactly as they are queried this is an example of:
    a) a concept based search
    b) a literal search
    c) a natural language based search
14. Complex queries to search engines can include _____(Boolean) operators to make the search more specific.

Questions for "Calculating Relevancy and Rankings" Slide:

15. The search results with the highest rankings are judged most _____(relevant) to the search query and are presented at the _____(top) of the page to the user.
16. Why do web search engines rank search results?
    a) Search engines rank results based on how much they are paid by website owners, so they rank the highest paying sites highest.
    b) Users want results with a high frequency of query keywords within the web page.
    c) Users tend to look more frequently at the top of the search results list.
    d) Users don't care how the search results list is ordered, so search engines do not rank results.

Questions for "Search Engine Optimisation" Slide:

17. Search engine optimisation (SEO) is:
    a) the process of making search engines faster.
    b) the process of making a web site or page more highly ranked.

c) the process of making search engines return more relevant results.
d) The process of making search engines more profitable.

18. There are two types of SEO, white hat SEO are techniques that are _____(approved and recommended) by search engines and black hat SEO are techniques that involve _____(deception).

A.5 Questionnaires

A.5.1 Pre-experiment questionnaire

The pre-study questions that participants were asked:

1. Gender: (Female / Male)
2. Age:
3. Is English your native language? (Yes / No)
4. Do you have any form of dyslexia or difficulties reading? (Yes / No)
5. Do you have normal or corrected to normal vision? (Yes / No)
6. Highest education level: _____
7. What area did major in or are currently majoring in?
8. Are you studying COMP1710? (Yes / No)

A.5.2 Post-experiment questionnaire

The post-study questions that participants were asked to complete:

1. Were you already familiar with the content you have just read and been quizzed on? (Very Familiar / Familiar / Somewhat Familiar / Not Familiar)
2. Would you find it useful/helpful to be given feedback about the parts of the text that you should re-read before attempting the quiz? (Yes / No / I don't know)
3. Would you find it useful/helpful to have electronic documents annotated with the areas of the text that you: *(tick any relevant)* Skimmed / Did not read properly (e.g. mindless reading) / Seemed to have trouble reading (may have not understood the concepts or language used / Read thoroughly / Read normally /Annotate nothing; I don't think this would be useful/helpful feedback.
4. Do you have any other comments?

Appendix B

Materials for Experiment 2 - Adaptive eLearning and Digital Images

This appendix includes of the supporting documentation and resources that were used for the Second experiment explained in this thesis – Adaptive eLearning and Digital Images Experiment. The resources included the participant information sheet, the consent form, the run sheet for the experiment, the texts used in the experiment, and the pre-experiment questionnaires.

Ethics approval was sought from the Australian National University Research Ethics Committee before the experiment was conducted. The experiment was conducted under Human Ethics Protocol 2012/006. The participant information sheet and consent form was designed according to the requirements of the ethics approval.

B.1 Participant Information Sheet

Participant Information

Project Title: Adaptive Learning Environments and Digital Images

Investigators:

- **Dr. Sabrina Caldwell** (0261259663, sabrina.caldwell@anu.edu.au)
- **Ms. Leana Copeland** (0422521154, leana.copeland@anu.edu.au)

Introduction
This document concerns a user study to be conducted at the Research School of Computer Science at the Australian National University. The study consists of two parts; the first requires that participants read through a set of paragraphs on a subject and then answer a set of questions based upon what was just read. In the second part, participants are required to view a set of images and answer questions about them. During this time user eye gaze and responses will be recorded.

What would be involved?
The time needed to complete this user study will be about 60 minutes. This time will include an introduction to the tasks, setup, and completion of the tasks mentioned above.

Data Collection and Contact Details
The main purpose of the user study is to collect data to enable useful information to be gained on the interface, the interaction techniques, and tasks. We will give you a pre- and post-task questionnaire that may contain some questions of an identifying nature. You do not need to complete these or any of the other questions if you have any objections to them. The data from the experiment will be made unidentifiable to retain privacy of each participant. Until that time, if you give your permission, your contact details will be retained for follow-up testing.

Data Use
The data collected will be used to draw conclusions about certain interaction techniques and the nature of the tasks. Any data collected, either raw or processed, may be used in a thesis and other research and publications. The data will be made unidentifiable so that no participant will be able to be identified from any data collected.

Risks
As the study is conducted in a carefully designed lab environment, all care will be taken to make participants as comfortable as possible, given the nature of the interaction tasks. Some physical discomfort such as eye and muscle strain may occur with some people ***including, in rare cases, motion sickness. Participants are free to request that your participation in the user study cease at any stage without explanation.***

Your rights
You may ask for a copy of any data collected or research publications written. You may also end the test session or ask for a break at any time and request that any or all data collected from you be destroyed. You have the right to completely withdraw from the experiment at any point. You can ask that your name be deleted from our contact list for future testing at any time.

This usability study is completely voluntary. If you have any concerns with the ethics of this study please contact the ANU ethics committee by emailing Human.Ethics.Officer@anu.edu.au or calling 6125 2900.

B.2 Participant Consent Form

Consent Form

1. I have had the project explained to me, and I have read the Participant Information sheet.
2. I agree to participate in the user study as described.
3. I acknowledge that:
    a) I understand that my participation is voluntary and that I am free to withdraw from the user study at any time and to withdraw any unprocessed data previously supplied (unless follow-up is needed for safety).
    b) The user study is for the purpose of research. It may not be of direct benefit to me.
    c) The privacy of the personal information I provide will be safeguarded and only disclosed where I have consented to the disclosure or as required by law.
    d) The security of the research data will be protected during and after completion of the study. The data collected during the user study may be published in this and subsequent research. Any information that may identify me will not be used.

Please list any Special Considerations (e.g. any medical conditions you have which you would like to bring to the attention of the user study supervisor)

_____

_____

_____

Participant's Name: _____

Signature: _____ Date: _____

B.3 Run sheet for user study

The following document outlines the run sheet for the user study so that the study is consistent no matter who is there.

1. Participant ID's are assigned as follows, concatenate the following:
    a. The first 3 letters of the day, e.g. mon, tue, wed, etc.
    b. The first 2 digits of the sign up time (in 12 hour time), e.g. 9am is 09, 12pm is 12, 3pm is 03
    c. Whether it is am or pm
    d. The date of the experiment in the form: DDMM

    For example, an experiment run on Monday 14th April at 9am is mon09am1404

2. Open FaceLab and use the stereo-head "jointExpSH2"

3. Open EyeWorks Record and set the following:
    a. The script file: C:\Users\faceLAB\Documents\My EyeWorks Projects\Leana\Joint Experiment\HCIExperiment.egs
    b. The Output file:
        i. location of where to store the data: C:\Users\faceLAB\Documents\My EyeWorks Projects\Leana\Joint Experiment\Data
        ii. filename: the above participant ID

4. Log into Wattle and got to the course: *Tom Gedeon's Sandpit*; go to the Administration panel and select *Users>Enrolled Users*. Select *Enrol User*. Get the participants Uni Id and search for it in the Not Enrolled Users section and enrol the student.

5. Go back to the course's main page, Turn editing on, and make visible one quiz under each of the sections *Part 1*, *Part 2* and *Part 3* in the Topic area *Leana and Sabrina's eLearning User Study*. Make visible the quizzes based on the combinations outlined in \Dropbox\JointExperiment\Administration_of_Experiment\Combinations.xlsx and update the spreadsheet to include the participant's ID in the attendance column.

6. Ask participant to read the Participant Information sheet (\Dropbox\JointExperiment\Administration_of_Experiment\Consent Form.docx) and sign the consent form (\Dropbox\JointExperiment\Administration_of_Experiment\Consent Form.docx)

7. Explain the following to the student:
    a. "Have you ever been in front of a Gaze Tracker before?" If No, show the participant what the tracker is and the video feed.
    b. "The gaze tracker has a narrow field of view so throughout the study you must remain relatively still as your face must remain inside these boxes at all times so that you eye gaze can be constantly monitored throughout the task. Please not get into a position that you will be comfortable to remain in for the next hour and make sure you can reach and keyboard and mouse. Try not to have a really straight back as people always slump their shoulders when they start to relax." Change the height of the desk to ensure that their face is within the video box.
    c. "When we start the experiment the gaze tracker will initialise by locating your pupils, once this has occurred a calibration sequence will begin. This involves a series of 9 red dots appearing on the screen; please look at each dot as it appears.

Once the calibration sequence has completed the experiment script will load. This can take a few moments.

Once the script loads you will be asked a few pre-experiment questions such as your age and gender. Once you have completed the pre-experiment questionnaire the experiment will begin. The experiment is divided into three sections in which you will complete a quiz and then view a set of images. Please keep in mind that you cannot click on the 'back' button in the browser to view earlier pages due to experimental constraints.

When you view the images a series of questions will be asked verbally. There are no right or wrong answers to these questions and they are not graded. You will begin by completing the first quiz and then you will view the first set of images. Then you will move onto the second quiz and then view the second set of images. Finally you will complete the last quiz and view the final set of images. You will not be shown the grade that you get for each quiz that you complete, as we do not want to affect your confidence. *Note that you cannot go back within the quiz so make sure that you feel confident that you understand the content before pressing the next button.* Do you have any questions before we begin?" If yes, answer the questions.

8. Ask the participant to look straight ahead and then press the Start Button in EyeWorks Record.

9. Once the calibration sequence is complete press Enter on the screen with 5 dots on it and then wait for the script to load. Since an IE window will open press the EyeWorks Presenter Icon in the system tray to bring up the start of the script. Hand over control to the participant from this point.

10. Allow the participant participant to complete the questionnaire and then the first quiz. Once they have completed the quiz make sure you click on the EyeWorks Presenter icon in the system tray to bring them back to the script so that they can view the images. Ask the following questions, record answers with iPhone or recording pen and take notes of answers:
    a. Two images: vector graphic frogs and raster graphic chameleon
        - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
        - "Of these two images, which is a vector graphic and which is a raster graphic?"
        - "Do you believe that the raster graphic has been manipulated?"

    b. Two images: one 8-bit b&w and one 24-bit colour image of boy
        - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
        - "Of these two images, which do you think has the highest bit depth?"
        - "Do you believe that either of these images has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"

    c. Two images: one low-res cat and one high-res cat
        - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
        - "Of these two images, which do you think has the highest resolution?"
        - "Do you believe that either of these images has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"

    d. Two images: one low-res cat eyes and one high-res cat eyes.

- "This is just a back up image to assist in identifying higher vs lower resolution. Which of these do you think has the highest resolution?

e. One image, coins
   - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
   - "Do you believe that this image has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"

11. Once they have finished with the images the participant will return to Wattle and complete the next quiz. Again once they are finished ensure that you click the EyeWorks Presenter icon in the system tray to bring them back to script so that they can view the images. Ask the following questions:
    a. One image, watermarked keypad
       - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
       - If they mention the text / water mark, ask "What do you think the text / watermark is for?

    b. One image, James Blundt photo
       - Say "this is a screen capture from Facebook. Looking at it, how do you think Facebook as a company benefits from uploads and discussions like this?"

    c. One image, Creative Commons logo
       - Question "this is the logo of an organisation, can you tell me who it is and what they do?"

    d. One image, John Howard and image of Queen in media scrum
       - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
       - "Do you believe that this image has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"

    e. One image, anemone in pond
       - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
       - "Do you believe that this image has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"

12. Once they have finished with the images the participant will return to Wattle and complete the next quiz. Again once they are finished ensure that you click the EyeWorks Presenter icon in the system tray to bring them back to script so that they can view the images. Ask the following questions:
    a. Two images, 'Fading Away' and zebra in clothes shop
       - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
       - "Do you believe that either of these images has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"

    b. One image, group of girls
       - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"

- "Do you believe that this image has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?" If participant answers no, say, "Actually, this image has been manipulated; one girl has been spliced in. Can you guess who it is?"

c. One image, missiles
   - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
   - "Do you believe that this image has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"
     If the participant answers no, say "Actually, this image is manipulated; one missile has been added. Can you tell which one?"

d. One image, car cow
   - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
   - "Do you believe that this image has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"
     If the participant answers no, move on.

e. One image, people on pier with 'jumping from pier' sign
   - Interest question "what are the interesting bits in these images" and/or "what in these images draws your eye?"
   - "Do you believe that this image has been manipulated?" If participant answers yes, ask "How do you think it was manipulated?"
     If the participant answers no, move on."

13. The participant is now complete. Stop the eye tracker and ask if they have any questions.

14. Request the student not discuss the experiment with other students since they may participate in the experiment and should not have knowledge of its contents beforehand.

15. Once the participant has left un-enrol the student from the course and disable all three quizzes.

B.4 Pre-experiment Questionnaire

1. What is your age?
2. Are you enrolled in COMP1710?
3. What are you currently studying and how many years have you been studying in this degree?
4. What is you gender? *(Male / Female / I would prefer not to say)*
5. What Language did you first learn to read in?
6. Do you have any form of dyslexia or difficulties reading? *(Yes / No)*
7. Do you have normal or corrected to normal vision? *(Yes / No)*
8. How fast do you think you read in comparison to others? (Very slow / slow / average / fast / very fast)

B.5 Experimental Content

The following text was written by Dr Sabrina Caldwell.

B.5.1 Topic 1: Working with Digital Images

Conceptual Difficulty	Readability		
	Easy	**Medium**	**Difficult**
Basic	Digital images come in many forms: photographs, icons, clipart, graphs, diagrams and sketches to name a few. They have many sources including scanning, photography, 'born digital' art and video stills. Digital images can be either vector or raster graphics. Vector graphics are created using mathematic descriptions such as lines and curves. The vector graphics we know best are fonts, but they are also used for clipart and icons. Raster graphics are better known as bitmaps. Bitmaps include the digital photographs we know as jpgs, tiffs and pngs. Digital cameras arrived in Australia in 1998, and rapidly overtook conventional photography. Today digital photographs are the most prevalent type of digital image. Over the years cameras have been included in many devices including mobile phones and tablets. Millions of digital photos find their way onto websites every day as media content, where they provide communication, information and entertainment. Digital cameras work by registering the light that falls on the camera sensor when the shutter button is pressed. Camera sensors are normally CCDs (Charge Coupled Devices) or CMOSs (Complementary Metal-Oxide Semiconductors). Together with other hardware and software within the camera, sensors record a series of bits known as pixels. Pixels (short for picture elements) store information about the light that fell on the sensor. The camera or your computer then assembles the pixels into an image that you can see. ***Words: 227*** ***FK Reading ease: 39.6, FK Grade level 11.2***	Digital images derive from numeric representations of two-dimensional areas of two types: vector and raster. They include photographs, clipart, video still captures and digital art. Vector graphics use scalable mathematical expressions to store and represent images. Geometrical primitives (lines, curves, polygons) are mapped onto the x,y axes of a plane and the resulting outline graphic is 'painted' with textures, shading and colours. Font types are simple vector graphics we use every day, and in addition, their high quality and economical space requirements make vector graphics useful for icons and clipart, and they are popular in 'born digital' graphical art. Raster graphics are dot matrices using rectangular squares known as pixels (contraction of the term picture elements) painted onscreen or printed one horizontal line at a time. We use these rectangular grids of pixels every day in digital photography as bmps, jpgs, tiffs, and pngs. Since their introduction in Australia in 1998, digital cameras have become ubiquitous in computers, mobile devices and tablets, however all such cameras work similarly, in that they use sensors (normally Charge Coupled Devices [CCDs] or Complementary Metal-Oxide Semiconductors [CMOSs]) to evaluate light and convert it into stored bits for later assemblage into viewable images. These images are uploaded to and shared across the Internet at a rate of millions per day. ***Words: 216*** ***FK Reading ease: 25.0, FK Grade level 15.7***	Digital images are electronic renderings using either vector or raster graphics and depicting representations of existing real world scenes, scanned texts and art. Vector graphics is a mathematical language modelling geometrical primitives on a working plane in two dimensions. Unlike resolution-dependent raster images, vector graphics render independent of device resolution, and scale seamlessly according to the screen device used for display or printing, although ultimately the graphic is displayed/printed as a raster image due to the constraints of current hardware and software infrastructure. Raster graphics such as GIF, PNG and BMPs store a dot matrix of individual pixels to compose an image. While not as compact as and more difficult to process and analyse than vector graphics, and unable to scale as well as vector graphics to arbitrary resolutions and sizes, raster graphics are more common, particularly in their use in digital photography. Raster graphic images, or digital photographs, obtain from the light incidence upon the multi-element camera sensors, which is predominately either a Charge Coupled Devices [CCD] sensor that exposes to the light all elements simultaneously, or a Complementary Metal-Oxide Semiconductors [CMOSs]) which has a rapidly rolling shutter that sequentially exposes the elements of the sensor. In either model the associated software computes the values of the light incidence for each element and converts it into stored pixel (picture element) information. Since 1998 when digital cameras were introduced in Australia, digital photo upload to the web has risen to millions of images daily. ***Words: 245*** ***FK Reading ease: 13.1, FK Grade level 19.0***
Intermediate	Digital photographs are an assemblage of information your camera stores about the pixels that make up the image. The quality and quantity of the information is	Digital photographs are constructed from pixel information recorded and stored by your camera. The colour information and size of your image file are based	Bit depth and resolution determine the extent of the palette of colours available to digital images (bits) and the level of detail defining their resolved vs pixelated

	dependent upon the bit depth and resolution respectively, which construct the pixels that form your image. Pixel information for digital images is contained in bits. The higher the bit depth, the more colours can be used for the pixel. Many standard colour image formats use 24 bits, comprising 8 bits for each red, green and blue channel of a pixel; combined, they define the overall colour of the pixel. This means that in its uncompressed state, each pixel requires a total of 3 bytes to record. This large amount of information provides a palette of 16 million colours and is known as 'true color.' Bit depth can be used in special ways for managing pixel colour. Using 32 bit colour offers transparency for formats such as png, and 32 or even 48 bit colour provides extra-large colour palette. The more pixels a camera can record, the better the resolution of the photo. A digital photo has high resolution when you can see a lot of detail in the picture, even when you zoom in. Low resolution photos quickly 'pixelate,' that is they degrade into blocks of pixels when you zoom in. This is because of the amount of information that the camera records when it takes the photo. The higher the megapixels, the more information. So for example a camera that takes photos using 3 megapixels records only a quarter of the information about a scene that a 12 megapixel camera captures. ***Words: 268*** ***FK Reading ease: 52.5, FK Grade level 10.0***	on the bit depth of your image and the resolution of your camera sensor. Digital image data is stored in variable numbers of bits (from 1 to 48) that dictate the pixel colour value; as more bits are employed to store data about the pixel colour, the range of colour choices available for pixels increases. The majority of image formats commonly used today use 'true color.' True color requires each pixel to be defined utilising 24 bits, comprising 8 bits each of red green and blue, or 24 bits (3 bytes) per pixel, which are blended to provide any one of 16 million distinct colours. Bit depth strategies can be employed to achieve specific colour management outcomes, for instance additional bits enable transparency in formats such as png, and 32-48 bits are used for ultra high colour resolution. Quality images result from a smooth gradation of pixels at sizes too small for the human eye to resolve. The lower the resolution, the more likely the eye will see the individual pixels, a phenomenon called 'pixelation' in which the image appears jagged and choppy, distracting from the image viewing experience. High resolution images result from cameras with 8-12 megapixel sensors that record high volumes of data about the image, meaning that the image will not pixelate under reasonable zoomed conditions. Conversely, low resolution images pixelate quickly and are best used as small images to avoid apparent pixelation. ***Words: 260*** ***FK Reading ease: 35.4, FK Grade level 14.1***	range (resolution). The array of colours available for the pixel value bit determinants increases exponentially as the bit depth increases linearly with minimum bit depth being 1 bit (enabling only 2 colours) and ranging upwards to as high as 48 bits (enabling 281 trillion colours). Bits can be deployed in alternative models such as 32 bit png files that facilitate fine-grained colour image resolution, calibrate germane colour indexes, and/or instill transparent regions, however the prevailing standard for bit depth is 24 bit 'true color.' This paradigm employs an 8 bit per colour arrangement in which component colours red, green and blue are sequentially represented by 8 bits per channel resulting in 3 bytes of information per pixel. This configuration enables pixel colour determinations derived from a colour palette encompassing 16 million hues. Image resolution is dictated by the number of sensing elements in the camera sensor behind the lens and iris of the camera. Low end cameras record little more than 1 megapixel of information and yield images that pixelate (become jagged and sharp-edged with visible squares of individual pixels) while high end cameras can record 8-12 megapixels or more of image data, and images of these megapixel magnitudes can be printed in large format or zoomed in without pixelation artifacts visible in the print or screen display. ***Words: 246*** ***FK Reading ease: 15.3, FK Grade level 19.8***
Advanced	More megapixels and high bit depths mean larger file sizes, but there are bit depth strategies for reducing files sizes, and many image formats have evolved that compress these large files into sizes that are easier to use. In their 'raw' state, digital images are quite large. With a 24-bit true color image and an 8 megapixel camera, each raw image file would require 24 megabytes to store. To manage this problem, especially when using digital images online, a number of image formats are available, often in camera. One of the most common formats is the jpg. Jpgs work by discarding information about colours our eyes are less sensitive to and encoding similarly	Memory costs are incurred when recording high resolution 'true color' images, however image formats have been developed to address this problem by compressing the information into manageable sizes with varying results. To calculate the uncompressed (raw file) size of a digital image, you multiply the number of horizontal pixels times the number of vertical pixels times the bit depth. Consequently, when taking a photo with a 12 megapixel camera set at the common setting of 24 bits, each photo will require 288 million or 36 megabytes. A common format to reduce these image sizes is the jpg.	As the quantity of pixel information recorded increases at higher bit depths and resolutions, so do computer storage requirements; in their raw form images of 48 bits and 12 megapixels would necessitate 78 megabytes storage space. To address this prohibitively space-expensive exigency, several industry-standard image formats have been developed that use compression/expansion algorithms to reduce file sizes on compression and restore images to full or partial original resolution upon viewing time. Jpgs are amongst the most common compression formats; using downsampling of chrominance channels,

	weighted blocks of pixels. This is a 'lossy' method, which means that once compressed, jpgs will not expand back to the same quality. An example of lossless compression is the png file. Pngs deflate the data into smaller pieces. The process is completely reversible. You can use bit depth settings to reduce file size as well. Grayscale images can be effectively rendered using only 8 bits (256 colours), which allows for small image sizes. Bits can also be assigned an index of colours from your image, thus reducing the file size without losing colour quality. Other techniques for reducing image size include resizing and cropping. ***Words: 220*** ***FK Reading ease: 54.1, FK Grade level 9.62***	Jpgs are a lossy compression method with a three pass methodology of a) reducing colour information in the less visible spectra, b) assigning relevance to blocks in the image, and c) identifying repeating patterns to reduce encoding. The png format is more recent and is a lossless compression method. As a raw image is reduced to a png, the png 'sliding-window' algorithm identifies data that can be transformed into reconstructable mathematical expression and stores it. In addition to other techniques such as image resizing and cropping, the use of bits can be varied so that they also act to reduce image size. They can be set to hold only a subset of colours, thereby creating an indexed colour image. Also, grayscale images require fewer shades and render well in as few as 8 bits. ***Words: 231*** ***FK Reading ease: 39.0, FK Grade level 13***.	quantization and entropy coding heuristics to reduce image sizes. Jpg is a lossy format, and information discarded through jpg compression cannot be recovered, leading to often substandard quality reconstructions when upsampling including block artifacts, jagged edges and pixelation. Pngs, a more recent compression format, use a deflate algorithm based on a sliding-window concept that capitalizes on the fact that data in images contains identical pixel repetitions, fragment repetitions, and gradients. As a raw image is reduced to a png, the compression engine algorithmically identifies data that can be transformed into and stored as one of a suite of mathematical expressions in a reversible framework. Resizing and cropping are additional image size reduction options as is alternative bit depth strategies: bit depth can be deployed in a selective colour model which calibrates germane colour indexes thereby removing unused 'placeholder' values, and grayscale images that render well at 256 shades of gray can be set to 8 bits for certain applications, greatly reducing images size. ***Words: 249*** ***FK Reading ease: 11.1, FK Grade level 18.7***

B.5.2 Topic 2: Copyright and Intellectual Property

Conceptual Difficulty	**Readability**		
	Easy	**Medium**	**Difficult**
Basic	To protect people who create original work, copyright law provides a way of deciding who pays and who gets paid for the use of original work. Copyright is one of a range of intellectual property rights, which also includes patents and trademarks. A photographer has copyright in his or her photos from the moment they are created, which lasts for 70 years after the photographer's lifetime. However, it is easy for others to acquire and use your work without permission. This is called copyright infringement and it is something others should not do to	Copyright is an automatic right afforded to creators of original works giving these creators exclusive economic rights to control copying, adaptation, issuance of copies to the public, performance and broadcasting of the work that they create. In return for licensing their materials the creators are entitled to receive royalties. Infringing copyright by using images without permission is ethically and legally wrong. Copyright, together with patents, trademarks, database rights, design rights, and performers' rights form part of	Photographs fall under the auspices of Intellectual Property and are afforded protection through the concept of copyright, a protection that persists until 70 years after your lifetime and gives photographers exclusive rights to license their image to others in respect of copying, performing, broadcasting and publishing. Should others use your images without permission, they are committing copyright infringement and may be liable to remunerate and/or make reparations for such infringement should you decide to take civil legal action

	you and you should not do to others. Your photos and your image art belong to you, but if you don't protect your copyright, who will? Digital images should be associated through metadata with the name of their creator. However this is often not the case, and a photograph can easily become 'orphaned' (an image without an author); as an orphaned image your photograph is more susceptible to infringement. Once you've created your work and associated your name to it through metadata, how can you empower your image to go out into the world and work for you? Licensing. There are many approaches to licensing your work. Images can be licensed to others via a stock image organization like Shutterstock or Getty Images. They can carry a bespoke license you create. A popular approach is to offer your image through a Creative Commons license. Creative Commons is an international non-profit organization that offers six standard licenses that brand your image as available for uses ranging from simple attribution to fully commercial, modifiable, and able to be on-licensed. ***Words: 271*** ***FK Reading ease: 49.7, FK Grade level 10.5***	the family of Intellectual Property Rights, which is the name of the broad range of rights that protect the fruits of human innovation, creation and invention. To ensure these rights are attributable to the correct creator, adequate author identification is required to ensure the work does not become 'orphaned,' or disassociated from the author; orphaned works are difficult to police and can easily be reused without recompense. This can be accomplished with metadata outlining authorship and licensing requirements. When you want to license and/or commercialise your photographs, there are a range of options from licensing your images through a commercial service like Shutterstock or Getty Images through to offering your images to the public under a Creative Commons license. By attaching a Creative Commons license to your image you specify who is able to use your photo and in what manner. ***Words: 223*** ***FK Reading ease: 31.1, FK Grade level 14.8***	against them. A particular risk with digital photographs is that an image can quickly become disassociated with its author by virtue of being transmitted and retransmitted without any attendant information identifying the copyright owner. When this occurs, it is said that the photograph has become an 'orphaned' work. An important precaution against this eventuality is ensuring adequate author identification and permissions identification in the metadata fields of the image, which can significantly reduce this risk. Photographers do not usually wish to sequester their photographs from the world; quite the opposite. To deal with copyrighted works legally and ethically requires licensing arrangements to be executed. Licenses can be adhered to a work by proffering it through a commercial stock photo company such as Getty Images or Shutterstock, licensing using a bespoke license or utilizing an open source copyright such as Creative Commons. Creative offers internationally recognised licensing that is embedded in or attached to copyright protected material; the open source non-profit organisation offers a range of standard licenses that specify what licensees are allowed to do with your work, from attribution only through to modification, distribution, commercialization and licensing derivative works to others. ***Words: 271*** ***FK Reading ease: 16.3, FK Grade level 17.0***
Intermediate	Online environments are complicated things. In addition to protecting your intellectual property, you need to consider how you use other people's information and intellectual property, and how other people are affected by the content you upload. For example, commonsense should tell you that unflattering images of friends and family should not be uploaded lest it create a future problem for them. Less obviously, you may need to consider how to content you receive from others. Web site owners soon find themselves on the receiving end of a range of information about individuals. Over the years two key	Posting images to websites entails responsibility that extends beyond copyright infringement and encompasses ethics, privacy and security. The photos you choose to upload may have consequences for you or people you know if they are ill-advised; images of friends may linger on the Internet and ultimately influence perspectives of others, including prospective employers, in future. Furthermore, in considering how to safeguard your intellectual property, it is wise to consider where you distribute your photos: does the site to which you intend to upload your work have privacy and security policies	There are ethical considerations in distributing/publishing your images. Sharing too much information within the public domain, particularly in image form, can be problematic; first it can reveal things to future employers, colleagues and social contacts about the person or people involved that may be detrimental, and second it may create personal safety issues. Understanding the privacy policies (methodologies and procedures by which your information is retained and shared with third parties) and security policies (safeguards in place to protect your information from unauthorized access) of the websites to which you

	documents have evolved to manage user expectations: privacy policies and security policies. A privacy policy describes what information the website keeps and how it is shared. Security policies describe how that information is secured against accidental or fraudulent access. If you decide to try to make money from your site, exploiting user information is an obvious but controversial choice. Many high profile sites have chosen this path with mixed results. Facebook for example does not charge fees. The CEO, Mark Zuckerburg and the Facebook shareholders benefit financially from users uploading images, personal information and other information such as 'likes' which enable them to target users with relevant ads. And they safeguard that information, which becomes an asset of the company. The data you upload may follow you on the Internet for years. ***Words: 229*** ***FK Reading ease: 38.7, FK Grade level 11.9***	(a privacy policy describes what information is retained by site owners and how it is shared; a security policy describes how it is protected)? Keep in mind that privacy and security are also the responsibility of anyone managing a website of their own. This becomes particularly important and potentially problematic if you decided to monetize your site. User data is in fact an asset prone to exploitation. Facebook is a good example of a high profile site that allows unlimited free uploads of images and other data. However, Facebook is not a beneficial society, it is a commercial, publicly-traded company with shareholders seeking profits; while Facebook does not charge users money the CEO, Mark Zuckerburg and the company derive $3 billion dollars per year in advertising revenues. However as users continue to find their user experience tailored ever more tightly to their preferences and content and ads customized around them in accordance with their needs, they may become increasingly wary of how the company uses their information. ***Words: 257*** ***FK Reading ease: 28.7, FK Grade level 14.5***	upload your content is important. While issues relating to inappropriate content sharing may be generally understood, many website owners do not realize the complexity of receiving and managing user information themselves. Website owners should develop privacy and security policies appropriate to the specific arena the website occupies (for example a counselling site may have higher privacy and security obligations than a movie review site), and special care needs to be taken when seeking to obtain revenue from the activities of and information inherent in the website. Facebook is a good example of how user information can be monetized in a manner that has implications for users' data. Mark Zuckerburg's business model is not predicated on a user pays approach but rather a 'data mining' model in which personal information gleaned from 'likes' and personal profiles is aggregated as marketing demographic data or used for targeted advertising known as 'relevance ads.' Depending on perceptions of this tactic, it may be viewed as beneficial or invasive by users. ***Words: 257*** ***FK Reading ease: 16.5, FK Grade level 18.0***
Advanced	Once intellectual property has been created it is tempting to believe that it will remain in existence. However, with digital images, this is not always the case. With conventional photography the output was almost always prints, physical copies of the image. These prints were shared and stored, where, depending on storage conditions, they could be expected to remain viable for many decades while aging gracefully. Digital images however are vulnerable to 'the digital cliff,' a phenomenon in which these photographs can become completely non-existent overnight due to technology obsolescence or failure of storage media. To protect your digital images it is important to upgrade them to current technologies and ensure backups are taken. Once these basic precautions have been taken, other forms of intellectual property protection can be considered. To preclude others from using your images	The ephemerality of electronic constructs implies additional vulnerability to destruction in comparison with their analog counterparts. The 'digital cliff' effect is a symptom of this phenomenon; rather than degrading over time and space as does an analog signal, digital signals are normally either received in their entirety or else not received at all. Archivists have compared conventional photographic prints to an analog signal (the print fades in storage over time) and digital images to a digital signal (the file persists until outdated or destroyed). Protecting digital assets from the digital cliff is the first and most basic step in intellectual property protection. But there are other electronic rights management strategies to apply on top of this step. One such strategy is using watermarking, is the application of faint or even invisible images and patterns	In contrast to conventional photography in which one or more examples of physical prints could be counted upon to exist, the intellectual property of digital photography must be safeguarded by electronic means. Without regular backups and technology upgrades including updating storage media, digital images may become obsolete or unavailable and fall over the 'digital cliff' into non-existence. Assuming that the existence of the digital image is protected, additional electronic rights management features can be employed to protect your intellectual property. One prevalent strategy is the use of watermarking, which is the application either visibly or invisibly of an auxiliary image or pattern to the surface of the image that identifies ownership of the image. A further useful modification intellectual property holders can make to their images is to identify their

	without permission, it may be desirable to use a watermark. Watermarks are a visible form of electronic right management information embedded in your image that make your photograph an undesirable target for infringement. Lastly, it is important to understand what metadata is and how it can work to protect your copyright. Some metadata schemes are EXIF, IPTC and XMP. Of these, XMP metadata is the most flexible for photographers. EXIF data is supplied by the camera when writing the image file to record camera settings. IPTC metadata is a particular system developed for news services who need verification details. But XMP, or Extensible Metadata Platform allows users to define and edit metadata tags to provide information about your image. You can use this for copyright information to safeguard your claims to your own original work. ***Words: 271*** ***FK Reading ease: 35.0, FK Grade level 12.5***	demarcating copyright. Most readily apparent of all is metadata tagging. This is information stored within the image file in one of a range of metadata systems. For example, IPTC metadata is an important tool for news investigation and broadcasting to clarify reporter and photographer identities and specific details about the evidence presented in the image. EXIF metadata is an automatically written set of metatags denoting camera settings. And XMP (Extensible Metadata Platform) is a flexible user customizable metadata system that photographers can use to identify their work and attendant information about that work, potentially warding against copyright infringement. ***Words: 232*** ***FK Reading ease: 18.7, FK Grade level 15.4***	ownership using image metadata, in particular XMP fields. While EXIF metatags are written automatically in camera and contain details of camera setting such as shutter speed and white balance, and IPTC metadata is used by news reporters and agencies in industry-specific ways, XMP can be customized by users. This means that of EXIF, IPTC and XMP metadata, XMP is most useful for photographers. In XMP, or Extensible Metadata Platform, users can define and edit metadata tags to provide information about images and ownership to safeguard against infringements. ***Words: 217*** ***FK Reading ease: 10.1, FK Grade level 18.5***

B.5.3 Topic 3: Photo Credibility

Conceptual Difficulty	**Readability**		
	Easy	**Medium**	**Difficult**
Basic	Altering photos is as old as photography itself. Back in the 1850's, only about ten years after the invention of photography by Henry Fox Talbot and Louis Jacques Mande Daguerre, photographers were using many negatives to create ghostly apparitions and photoart. The difference between then and now is that with conventional photography altering photos was an expensive and time consuming process only skilled photographers could do. Today anyone can modify a digital image using Photoshop or Instagram. They can then share it with the world in moments. People usually make these changes for fun or art. However, sometimes people change photographs to create false images to criticize others or to make money. While there is no process at present to authenticate a	Image tampering has been around since the advent of the photographic process. In the 150 years commencing with the invention of conventional photography by Henry Fox Talbot and Louis Jacques Mande Daguerre, and before digital photography was introduced, photographers were staging images and/or creating seemingly real but actually false photographic prints or photoart pieces crafted from disparate negatives in photographic darkrooms. Digital photography and the long arm of the Internet increased the problem of photo manipulation. Image manipulation software has become inculcated into photographer's postprocessing of photographs, offering easy access to an extensive palette of image tampering tools. Such manipulated images are now common;	In the1840s, photography was simultaneously invented by William Henry Fox Talbot and Louis Jacques Mande Daguerre. Initially a completely scientific discipline, within a decade photographers were superimposing negatives to manufacture manipulated photographs and offering specimens as photoart, or more problematically, factual representations of reality for publication. Once a vanishingly small cohort of practitioners, such image tampering is now ubiquitous, done for reasons of fun or art but sometimes, more insidiously, for profit or libel. "Owing to such sophisticated digital image/video editing software tools, the establishment of the authenticity of an image has become a challenging task, encompassing a variety of issues." This is of particularly relevant social concern given the pervasiveness of manipulated

	digital photograph from the moment it was taken, there are a number of tests that can indicate if an already existing image has been altered. Together, these tests are called digital image forensics. Digital image forensics analysts compare elements within an image to identify changes. It is an important field because digital images are increasingly being used in areas such as intelligence gathering, court proceedings, news, medical imagery and sports. This can have a direct impact on people such as defendants, insurance claimants and ordinary citizens, as well as industries such as news publishing, betting, and medicine. For example, defendants can claim that digital photographs are unreliable, and insurance claimants can falsify photographs of damage. ***Words: 237*** ***FK Reading ease: 35.1, FK Grade level 12.8***	although usually manipulated for fun or art some photos are manipulated for political or commercial ends. Presently, photos are often illusive electronic constructs, globally distributed at the speed of light with little context or explanation. Despite attempts by some camera manufacturers, no authentication process has yet been successfully implemented. However, a range of digital image forgery detection techniques have been developed in recent years. Collectively these techniques are known as digital image forensics, a field that analyses images to determine image veracity through identifying image manipulation artifacts. Interest in and development of digital image forensics techniques is increasing due to the potential of image manipulation to impact on medicine, justice, news reporting and the legal and accounting professions. Recently, defendants have been successful in rejecting photographic evidence based on the fact that they cannot be authenticated. ***Words: 242*** ***FK Reading ease: 12.9, FK Grade level 16.4***	photographs throughout most disciplines and social platforms, which has logarithmically exacerbated the problem. In fact, positively authenticating an image is not currently possible due to the lack of an accepted proactive authentication software solution. However there are several technological approaches extant, which together comprise the fledgling discipline of digital image forensics. "Digital image forensics is a field that analyses images of a particular scenario to establish (or otherwise) credibility and authenticity through a variety of means. It is fast becoming a popular field because of its potential applications in many domains, such as intelligence, sports, legal services, news reporting, medical imaging and insurance claim investigation." In courts, defendants are beginning to challenge digital photographic evidence on the grounds that their veracity cannot be guaranteed. ***Words: 238*** ***FK Reading ease: 1.3, FK Grade level 18.9***
Intermediate	Let's consider the three most popular kinds of photo manipulation: copy/move, splicing, and retouching. Copy/move is an approach in which an area is copied from one place in the image and moved to another place in the same or similar image. These types of manipulations can be found using a technique called approximate block matching. In this technique, a range of overlapping blocks are separated out and each is compared to its neighbour to identify similarities and differences. Image splicing has more potential to create fictional images. In image splicing a false image is created by combining more than one image. Detecting spliced images mainly occurs by identifying adjoining regions and edges. For example sharp edges or abrupt changes between different regions suggest that an image has been created by splicing. Image retouching is the third main type of image alteration. It includes airbrushing (which we are familiar with from photos of models in magazines) and using	There are three main forms of image tampering (copy/move, splicing, and retouching), each with their own suite of forensics detection techniques. Copy/move forgery is one of the most popular forms of tampering, in which a target region is copied from a particular location in an image and thereafter pasted at one or more locations within the same image or a different image of preferably the same scene." These types of forgeries are detected using approximate block matching strategy. "A typical approximate block matching strategy splits the image into overlapping blocks and applies a suitable technique to extract features on the basis of which the blocks are compared to determine similarity." Image splicing techniques are used to compose one image from multiple images. "Splicing detection is a challenging problem whereby the joining regions are investigated by a variety of methods. The presence of sharp edges (or changes) between different regions and their surroundings constitute valuable clues to splicing in	Copy/move, splicing, and retouching are three popular image tampering paradigms, for each of which forensics detection techniques have been developed. Image forgery employing 'copy/move,' a technique in which regional image components are cloned and applied intra-image or more rarely inter-image, is amongst the most common tampering strategies. In this instance, approximate block matching detection is utilized to sequester areas of the image exhibiting repetitive pixels between overlapping blocks. Spliced images involve multiple photographic sources from which salient features are extracted and combined to create new images; this technique affords greater potential for forgery and falsification of photographs. Detecting spliced images is a difficult problem wherein forensic investigators seek indicative artifacts such as sharp edges and changes suggesting combinatorial regions. Retouching is historically the most prevalent form of image tampering (particularly as used in airbrushing

	filters to soften or sharpen or adjust colour. Retouching enhances or diminishes individual features in an image or applies a global change to the whole image. Individual changes are usually made using a number of small copy/moves such as cloning skin pixels to cover a blemish. Detecting the use of this technique involves finding one or more or enhancements, blurring, illumination and colour changes. If the source photo is available this may be easy. Otherwise, the task may be very difficult. Individual manipulations are investigated using copy/move forensics. To detect contrast and colour enhancements (if visible) investigators usually for global modifications. ***Words: 262*** ***FK Reading ease: 41.6, FK Grade level 11.2***	the image under investigation." "Image retouching is another class of forensic methods that pertains to a slight change in the image for various aesthetic and commercial purposes, not necessarily conforming to the standards of morality. The retouching is mostly used to enhance or reduce the image features." "Forgery detection, in case of image retouching, involves finding the enhancements, blurring, illumination and colour changing." Enhancements may be local (usually copy/move modifications) or global (contrast enhancements affecting the entire image) and forensic investigation requires the application of an extensive range of techniques. "Forgery detection may be an easy task, if the original version is available. Otherwise, with blind detection, the task may be very challenging." ***Words: 274*** ***FK Reading ease: 29.5, FK Grade level: 14.3***	photos of models). Retouching involves either global filtering of images for softening/sharpening effects or chromatic enhancement, or local changes, especially move/copy such as cloning pixels to erase imperfections. Where the source image is available detection of retouching is more facile, however in the absence of reference images the image under investigation must be 'blind detected,' and the task becomes significantly more difficult. "For the manipulation of retouched images, two types of modifications are applied, namely local and global modifications Local modifications are usually used in the copy/move forgery or in the case of splicing. For the detection of contrast enhancements that perceptually impact the image, global modifications are usually investigated." ***Words: 249*** ***FK Reading ease: 11.6, FK Grade level 17.2***
Advanced	We communicate with each other in images far more frequently than once was the case. For example, we may take a photograph of our coffee and cake in a café and forward it on our iPhone to a friend instead of chatting on the phone and describing our trip to the café. What has also changed is how easy it is to change our photographs. It was once the case that few photographers could tinker with their photos, now almost anyone can. While the effects of airbrushed models and product image enhancement is commonly understood, there is little understanding of the effects of day to day manipulation of photographs in social media, family photos, and public images. Why is all this image manipulation a problem? There are many reasons, but to take one significant issue, image manipulation is a problem because we are manipulating our personal stories one image at a time. According to Dr Ira Hyman, Professor of Psychology at Western Washington University, "our photographs can actually change and modify our memories over time." He asks, "How many of your childhood memories resemble the pictures that your parents took? Is it your memory or their picture?" When we alter our photos, we also alter	Increasingly, we encounter information about the world in visual form. At the same time, human capability to manipulate images is greater than at any previous point in history; it is an intuitive and rapid process within the reach of anyone with an iPhone and Instagram. Copious research has been undertaken on the use of manipulated images in advertising and marketing, but there is inadequate understanding of the effects of casual photo manipulation such as is prevalent in news, social media and family photos. However, some issues are coming to the fore. One such issue pertains to the relationship between human memory and photographs. Photographs are memory aids that represent events, people and places in our personal experiences. That they are effective in this role can be seen in the fact that photographs can act upon our memories to emphasise some aspects and minimize others; in effect, they influence our memories (Ira Hyman, Professor of Psychology, Western Washington University). Altering our photographs equates to altering our memories of the locations, participants and events within which the photograph transpired. When considering the construction of history through	An effect of the information technology revolution is that society is increasing its consumption of information in visual form (witness Facebook, Pinterest, Instagram). Simultaneously, the ability of humans to tamper with visual information through Photoshop and 'on-the-fly' image altering has dramatically increased. Extensive research exists regarding advertising and marketing uses of image manipulation, however the effects of ubiquitous use of photo manipulation in news, social media and family photographs is as yet largely unquantified. One initial issue gaining attention is the impact of image manipulation on the photograph-memory link. Photographs are mnemonics whose importance increases with chronological imperatives; memories fade over time and we rely upon photographs as reportage to prompt our memories of the events contemporaneous with the image acquisition. Tampered images create a flow-on effect of tampered memories: our memory acuity is influenced by the photographic representations appertaining to them (Ira Hyman, Prof. Psych, WWU). As unverifiable images continue to provide evidentiary reportage of the real world our understanding of those

	our memories. Further, in the big picture, we as a society are amassing a large, mobile body of images that cannot be relied upon as records of actual people, places and events as was the case in the era of conventional photography. ***Words: 247*** ***FK Reading ease: 47.7, FK Grade level 11.6***	personal narratives and the body of evidence including photographs, these alterations may affect our societal cognition of history, subverting our evidence of events, places and people over time. ***Words: 212*** ***FK Reading ease: 28.8, FK Grade level 14.0***	events becomes questionable; evidence of societal meta-narratives become polluted with unquantifiable and unqualifiable falsifications; future generations will be forced to consider whether the photographs extant relevant to their interests and investigations are reliable as reportage of real events, places and people. ***Words: 204*** ***FK Reading ease: .3, FK Grade level 18.2***

Appendix C

Dealing with imperfect eye gaze data

Gaze data often needs to be cleaned up and inference must be made about where fixations actually occurred subsequent to data collection. Noise in eye gaze data can be due to inaccuracy of the equipment, and characteristics of a participant's eye that make it hard to track them (Hyrskykari, 2006). Stereo camera eye tracking allows for head movement (Beymer & Flickner, 2003) and so are used, as the eye tracker in these experiments uses stereo cameras. The stereo cameras are first calibrated and for each participant the eye tracker begins with a 9-point calibration sequence. Even so, it is often noted that calibration must be done throughout experiments to ensure that it is correct throughout the experiment (Hornof & Halverson, 2002; Hyrskykari, 2006). There are methods for adjusting and recalibrating the eye tracker during use such as the use of implicit required fixation locations (RFLs) (Hornof & Halverson, 2002). Implicit RFLs are locations on a screen that a participant must look at as part of a task and therefore provide a location where the eye gaze data can be recalibrated from if deviation has been encountered. Other algorithms such as presented by Hyrskykari (2006) are highly related to reading tasks and involve using lines of text as the locations where fixations are reference points for mapping of the gaze data. This algorithm is used in real time as part of a reading aid called iDict and allows for manual corrections to be made if the fixations are not be mapped to the right words (Hyrskykari, 2006). This algorithm focuses highly on the vertical disposition of gaze points rather than the horizontal disposition.

For post-collection recalibration of data, inference about where the fixations should occur can use the same logic about the above examples of recalibration of eye gaze trackers during experimentation. To deal with the distortion of the data, one solution is to apply transformations to the data to move points to where they reasonably should be. Of course this begs the questions of how one defines where the data points should reasonably be. The next problem is applying such a transformation. The data points are distorted in different ways between and within participants. That is, even for the same participant, from text presentation to the next text presentation the distortion may be different. The simplest option is to manually apply the transformation for each set of fixations for each participant. The next step is to automate the process. Examples of misaligned fixations are shown in Figure C.1.

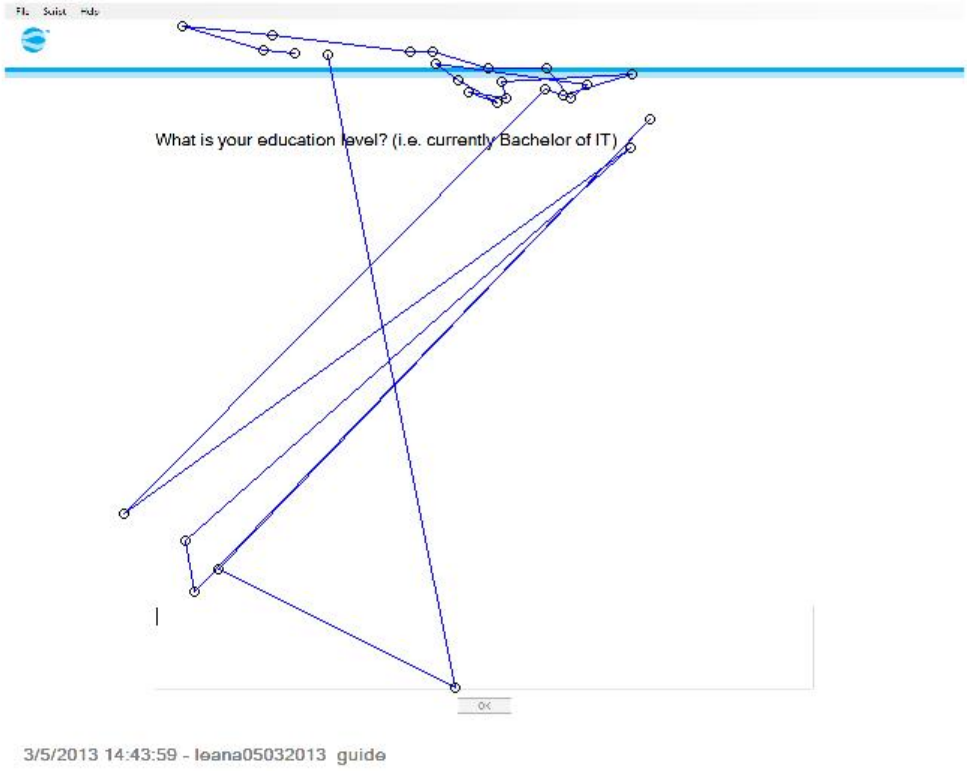


Figure C.1. Example of misaligned fixation data.

The code used to shift the fixations is:

```
vert_box_ratio=y/(max(y)-min(y));

hor_box_ratio=x/(max(x)-min(x));

new_y=y+(vertical_shift*((1-vert_box_ratio)+(1-hor_box_ratio)));

new_x=(x*hor_spread_factor)-horizontal_shift;
```

Note that x indicates the x coordinate of the fixation and y indicates the y coordinate of the fixation and the experimenter defines the values for variables.

The outcome of this shift is:

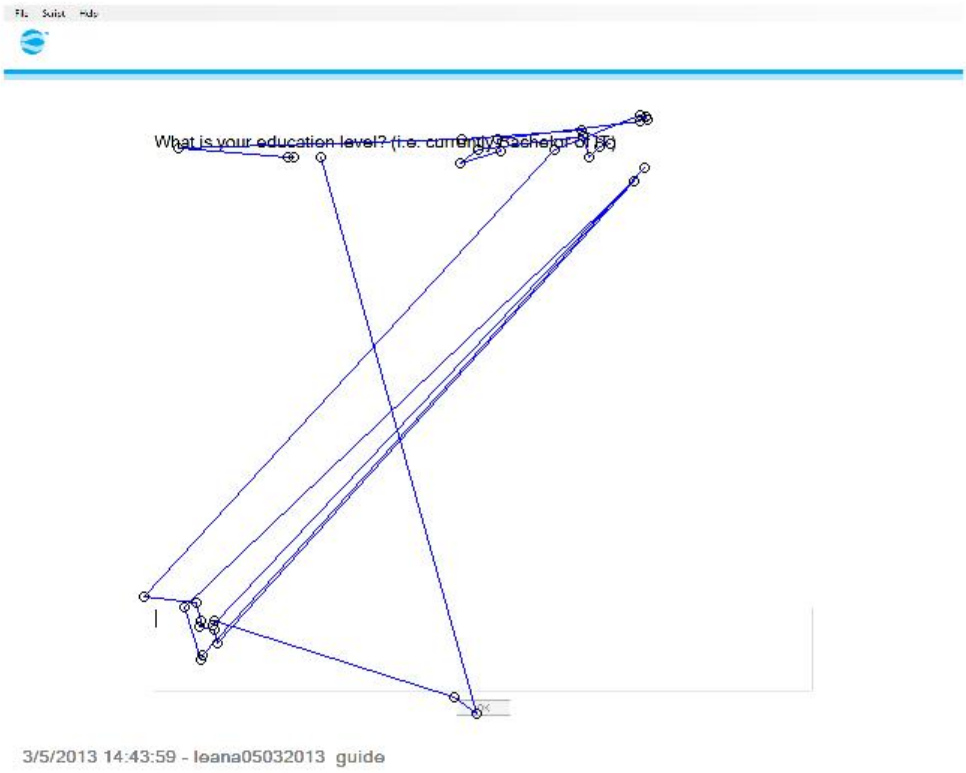


Figure C.2. Example of re-aligned fixation data.

The experimenter would manually go through the fixation data and re-align the fixations using this shift. The process was not automatic.

Appendix D

Reading in distracting environments

"Any distraction tends to get in the way of being an effective gangster."

— Terence Winter, *creator of Boardwalk Empire*

Reading in digital environments can be very distracting. In this appendix we present a preliminary user study in which participants' eye gaze was recorded as they read text in a visually distracting environment. We explore two distraction mitigation signals using real-time eye gaze data to investigate whether the effects help reduce distraction rate as well as aid recovery from distractions. These signals involved adding a signal to the last word read before a distraction occurred to show the reader where they were up to. We compared these experimental conditions on both first (L1) and second (L2) English language readers and for easy and hard to read texts. The results demonstrate that the mitigation signals helped recovery from a distraction by drawing participants' attention back to the text as well as indicating from where to recommence reading. We conclude with recommendations on implementing distraction mitigation signals in text and limitations of this study. This appendix is based on work presented at OzCHI 2015 (Copeland & Gedeon, 2015).

D.1 Introduction

Digital environments make vast amounts of information readily available. However, these environments are dynamic, distracting the user with alerts, advertising, social media, and other distractions. It has been shown that auditory distractions, such as background noise, impair reading comprehension (Sörqvist, Halin, & Hygge, 2010) and that visual distractions lead to disruptions in cognition (Atkins, Moise, & Rohling, 2006). In the case of educational material, irrelevant and attention grabbing images or animations alongside text material have negative effects on learning (Clark & Mayer, 2011; Harp & Mayer, 1998; Mayer et al., 2001; Sung & Mayer, 2012). However, distractions can be avoided by using attention

guiding to ensure that important information is seen (Rosch & Vogel-Walcutt, 2013). Our hypotheses therefore are that visual distractions have a negative impact on reading behaviour and comprehension, but can be mitigated using attention guiding, to help reduce the disruption of visual distractions during reading.

We explore these hypotheses by also investigating the effects of text readability on the extent to which the visual distractions impact comprehension and distraction rate. We know that auditory distractions impair proofreading performance and prose recall, but the impairments only occur when the reading task is easy (Halin, Marsh, Haga, Holmgren, & Sörqvist, 2014; Halin, Marsh, Hellman, Hellström, & Sörqvist, 2014). In digital environments many visual distractions are possible, such as the reader having dual screens open with Facebook showing on one screen, advertising on webpages, or simply the pop-up alerts used by many applications such as email.

The objective of this study is to investigate firstly, the effects of text readability on the rate at which participants are distracted and secondly, whether attention guiding can be used to mitigate distractions for test with different readabilities. Readability is determined by readability formulas such as the Flesch-Kincaid Grade level. We investigate the effects of text readability and distractions on first language English (L1) and second language English (L2) readers. Distractions are induced using images that change at constant rates in a side bar. An eye tracker was used to record and monitor eye gaze of participants. Using this live data, we implemented a signal to trigger on the last word read before a distraction.

We hypothesize that the easy readability text and the L1 readers will be associated with higher distractions rates, and that the mitigation signals will reduce distraction rates and will help the reader recover after distraction.

This appendix is organized into the following sections: background information; user study method; results and discussion; and further work.

D.2 Background

Much of the background has been covered in the literature review (Chapter 2) of the thesis. As follows, only the literature that has not been covered in that review will be addressed in this section.

D.2.1 Images and text

It is generally accepted that including images along with text is beneficial to the learning process, the basis of which lies in dual coding theory (Mayer, 1999). Put simply, the activation of two cognitive subsystems results in more effective learning. In this way Mayer (1999) proposed five design principles for multimedia education, amongst which using words and images is primary. Images have a large effect in real word scenarios, such as educating patients in health care. Images improve understanding of health care instructions and change adherence such instructions (Houts et al., 2006).

However, it has been shown extensively that the images or animations must be relevant to the learning materials (Clark & Mayer, 2011; Harp & Mayer, 1998; Mayer et al., 2001; Sanchez & Wiley, 2006; Sung & Mayer, 2012). Use of *seductive* images, those that attract attention but are irrelevant to the learning materials have been shown to have a negative effect on learning because the images draw the reader's attention away (Sanchez & Wiley, 2006; Sung & Mayer, 2012). The effects of seductive images explored using eye tracking suggest that readers with low working memory capacity are affected more as they spend longer looking at the seductive images than those with high working memory capacity (Sanchez & Wiley, 2006). Another image type that is used in learning materials is decorative images, which are irrelevant to the learning material but not attention grabbing. Whilst it has been shown that decorative images do not negatively impact learning, they do not improve learning (Sung & Mayer, 2012).

D.2.1.1 Distractions during reading

Irrelevant and attention grabbing images can be considered distractions from the text rather than helpful resources. Simplification and reduction of distractions is best when aiming to avoid unnecessary cognitive load (Sweller et al., 1998). Visual distractions from unnecessary elements have been shown to lead to disruptions to cognition (Atkins et al., 2006). Additionally, auditory distractions such as background noise have been found to impair reading comprehension (Sörqvist et al., 2010). The extent of the impact of these distractions is aligned with the complexity of the task, where impairments on prose recall and proofreading performance only occurred when the reading task was easy (Halin et al., 2014a; 2014b).

Distractions, such as television, provide both visual and auditory disturbance. Computer use in front of a television has shown that people switch between the two medias frequently and that they underestimate the extent of how frequently they are switching (Brasel & Gips, 2011). Whilst not directly related to reading, these results emphasise the importance of investigating how distractions affect readers in a digital environment.

As stated, digital environments provide many distractions within themselves. One such distraction is computer mediated communication technologies such as instant messaging (IM). Whilst using IM during reading does not appear to negatively impact reading comprehension, extensive used of IM is associated with lower reading comprehension scores as well as lower GPA scores (Fox et al., 2009). Whilst "IMing" during a reading task does not negatively impact reading comprehension scores (Bowman et al., 2010; Fox et al., 2009; Jacobsen & Forste, 2011), it negatively impacts the time taken to complete the reading task.

IM is not the only distraction ever-present in digital environments. Recently the use of social media has proliferated in use, especially amongst the young generations. These are the generations now studying so the effects of such technology on learning are especially important. It has been found that students who use Facebook spend less time studying and have lower GPAs (Kirschner & Karpinski, 2010).

D.2.1.2 Mitigating distractions

Attention guiding can be used to minimise distractions by providing visual cues using colours to emphasise relevant parts of animations (Boucheix & Lowe, 2010), or by zooming in on parts of animations (Amadieu et al., 2011), and signalling parts relevant parts of diagrams by adding temporary colour changes (Ozcelik et al., 2010). The addition of eye tracking data to the paradigms has been found to enhance their effectiveness of attention guiding (Boucheix & Lowe, 2010; Ozcelik et al., 2010).

D.3 Method

D.3.1 Participants

Data was collected from 66 (28 female) participants with an average age of 21.7 years (standard deviation of 3.9). All participants had normal or corrected to normal vision and were primarily (n=54) recruited from a first year Computer Science course on Web Development and Design offered at the Australian National University (ANU). The remaining participants were all students from ANU. Participants were divided into two groups; those that first learnt to read in English, denoted L1, and those that first learnt to read in another language, denoted L2. There were 42 L1 participants and 24 L2 participants.

D.3.2 Design

The study used a between-subjects design with 3 independent factors: 1) text difficulty; 2) distraction mitigation signal; and 3) whether English was their first reading language. There were two levels of text difficulty, three distraction mitigation signal conditions, and two language groups. All participants were exposed to the same distracting environment.

We experimented using two distraction mitigation signals and had a control condition, these conditions are denoted and described as:

Condition A: Cue is yellow highlighting and bolding the last word the reader fixated on.

Condition B: Cue is the last word the reader fixated on coloured grey and italicized.

Condition C: No cue applied to text

The aims of these conditions are to explore the effects of bringing the readers' eyes back to the text, in particular the point they were up to in the text. Secondly, rather than actively drawing their attention back to the text, just give the reader a signifier of where they are up to in the text to help when they do focus their attention back on the text. In both cases the reader will feel the presence of the system monitoring them. The question is whether the cues reduce the effect of distraction? The remainder of this section discusses the design of the texts used, the distracting environment and finally the mitigation techniques.

D.3.2.1 Text Properties

The experiment involved two parts; firstly, the participant was asked to read a piece of text with either easy or hard readability. The readability was calculated using several readability formulae and the average of the tests was used. The readability formulae used were, Flesch-Kincaid Grade Level, Gunning-Fog Score, Coleman-Liau Index, SMOG Index, Automated Readability Index. The easy-to-read text has an average score of 10.6 (Table D.1); this equates to only a high school level of education needed to comfortably read this text. Given that participants are university students the text should be comfortable to read by participants. However, the hard-to-read text has an average score of 18.0 (Table D.1) indicates that a much higher level of education is needed to comfortably read the text. Participants should therefore find it difficult to read.

Table D.1. Readability scores for each text type.

Readability Formula	Easy Text	Hard Text
Flesch-Kincaid Grade Level	9.5	17.8
Gunning-Fog Score	12.2	21.3
Coleman-Liau Index	12.7	15.8
SMOG Index	9	15.2
Automated Readability Index	9.5	19.7
Average Grade Level	*10.6*	*18*

The statistics of each text type are shown in Table D.2. Whilst the number of words is different by more than 100 words the number of characters is kept roughly the same, which in turn equates to the lengths of the text being approximately the same. We can see that the hard text has significantly longer words as well as longer sentences compared to the easy text.

Table D.2. Text statistics for each text type

Text Statistics	Easy Text	Hard Text
Character Count	3,693	3,746
Syllable Count	1,215	1,246
Word Count	764	698
Sentence Count	47	22
Characters per Word	4.8	5.4
Syllables per Word	1.6	1.8
Words per Sentence	16.3	31.7

The experiment used a *between-subjects* design so each participant was shown either an easy or a hard text to read. After the text was read, participants' comprehension was tested using 10 comprehension questions that were the same for both texts.

D.3.2.2 Making the environment distracting

Participants were required to read text in a distracting environment. This involved creating an environment with a controlled level of distraction so that each participant would be exposed to distraction to the same degree. To accomplish this, a sidebar on the right of the screen was added. In the sidebar a picture at the top is changed every 20 seconds. The pictures in this box are different animals, for example a meerkat. Below this in a rectangular box, names are changed at random every 5 seconds. Both are shown in Figure D.1. The right sidebar is designed to stay constantly in focus whilst the participant scrolls through the text. This mimics some properties of Facebook pages, while being consistent for each subject.

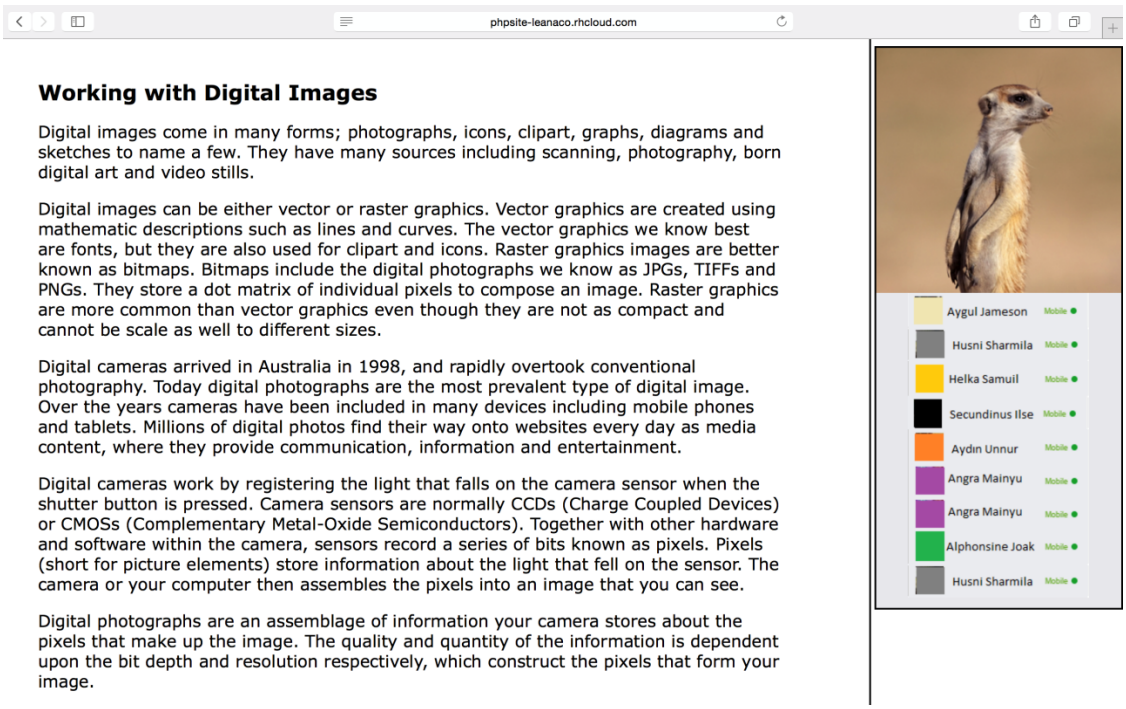


Figure D.1. Example of distracting environment

Distraction mitigating signals are added to the text to show where the reader was up to in the text before they were distracted. This was to investigate whether adding text signals helps the reader recover after reading, and if the participants consider it helpful. Two signals were used in the study, the first is an overt signal and the second is a subtler signal. In both cases the signal is only applied to the last word the reader fixated on according to the eye tracker, before a distraction drew the reader's eyes away from the text. Both signals were designed so that as soon as the reader looks at the affected word the signal would disappear.

Signal A: Highlighting (yellow) and bolding the last word read before a distraction, shown in Figure D.2.

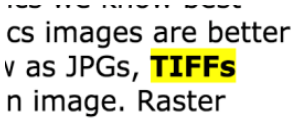


Figure D.2. Example of signal A; highlighting and bolding of the last word read before a distraction.

Signal B: Italicizing and making the last word read before a distraction grey, shown in Figure D.3.

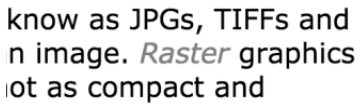


Figure D.3. Example of signal B; greying out and italicizing the last word read before a distraction.

The aim of the two text signals is to explore the effects of bringing the readers' eyes back to the text, in particular, where they were up to in the text. Secondly, the signals are designed to only give readers a signifier of where they are up to in the text to help when they do focus their attention back on the text.

D.3.3 Materials and Procedure

The experiment duration was approximately 30 minutes. First, the experiment was explained to participants. Then participants were asked to read and sign a consent form. Participants were given a pre-experiment questionnaire. The questions were designed so that we could gauge participants' use of potentially distracting technologies. The questions asked of the participants are:

1. Do you use social media? (If yes, how regularly?)
2. Do you use email? (If yes, how regularly?)
3. Do you use instant messaging? (If yes, how regularly?)
4. Do you often use social media, email and/or instant message while you are reading course materials or work materials? (If yes, how regularly?)
5. Do you find that you are distracted by these technologies during study or work time? (If yes, how regularly?)

Note that how regularly was restricted to the following options: Never; Once a month; Once a week; Once a day; 2-5 times per day; 5-10 times per day; and 10+ times per day.

Calibration of the EyeTribe eye tracker was performed until 'perfect' calibration was obtained according to the tracker. A 9-point calibration protocol was used, shown in Figure D.4. According to the EyeTribe software, perfect calibration is the optimal calibration result and equates to accuracy being < 0.5°. The eye tracker recorded eye gaze at 30Hz.

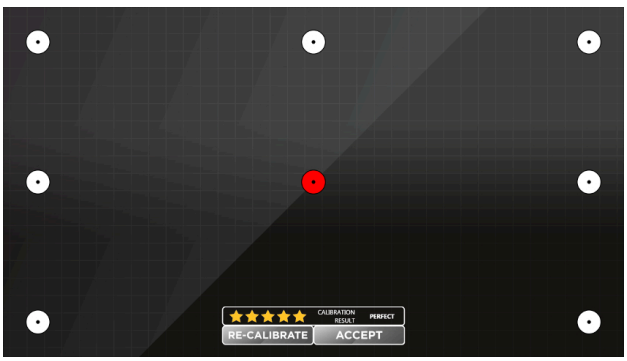


Figure D.4. Example of the 9-point calibration screen used in the experiment showing that perfect calibration was accomplished.

The experiment was run on a Macbook Pro 13" and participants were free to move their heads, however, they were asked to stay relatively still while the tracker was on. The setup of the experiment is shown in Figure D.5.

Figure D.5. Experiment setup

After the calibration routine, participants read the text whilst their eye gaze was being monitored and recorded. Finally, a post-experiment questionnaire was given to the participants. In the post questionnaire participants were asked if they were: 1) distracted whilst reading; and 2) whether they thought this had an impact on their understanding. In the conditions where a text signal was used, participants were also asked if they thought the text effect 1) reduced their distraction rate; and 2) helped them to start reading the text again.

D.3.4 Data pre-processing

The raw eye gaze data collected from the eye tracker consists of x,y-coordinates recorded at equal time samples. Fixation and saccade identification was performed on the eye gaze data. To detect fixations, the dispersion threshold identification algorithm (Salvucci & Goldberg, 2000) was used. The duration threshold was set to 150ms and the dispersion threshold was set to 30 pixels.

Once the fixations have been identified, eye movement measures were derived to characterise the reading behaviour. The measures used in this analysis are:

Number of fixations: From the fixation identification algorithm the number of fixations observed for the page is calculated. We report the total number of fixations.

Total fixation duration: The sum of the durations of all recorded fixations is calculated as well as the sum of fixation durations.

Number of distractions: The number of times a participant moves their eyes to the distractions from the text.

Percentage of fixations on distractions: Number of fixations recorded on the distractions divided by the total number of fixations. This provides information about the extent to which a participant was distracted rather than a raw count of distractions.

D.4 Results

There are a number of results from the study. First we look into the pre-questionnaire data to investigate their use of communications technologies, and then we investigate the effect each condition had on participants' eye movements and reading comprehension. Finally, we look at the post-experiment questionnaire data to explore their perceptions of the distracting environment and the text signals.

D.4.1 Pre-experiment questionnaire data

Participants completed a pre-experiment questionnaire to reveal their use of communication technologies. 50 stated that they use social media, however all participants (n=66) stated that they use email and instant messaging technology. Additionally, 65 of the 66 participants stated that they use social media and / or emails and / or instant messaging while they are reading learning materials for university.

We can also analyse the self-rated frequency with which participants are distracted by social media, email, or instant messaging while studying (or working). Participants were asked to rate their use and distraction levels on a Likert scale as described in

Figure D.6. When asked how regularly they use these technologies whilst reading learning materials, 46% of these participants stated that they use these technologies more than 10 times per day. 56 stated that these technologies distract them while they are studying. As Brasel and Gips (2011) people underestimate the amount they are distracted so this level could in fact be a lot higher. This establishes that participants have quite a high level of usage of communicative technologies and on average are quite distracted by them while they are studying.

Participants who stated that they use communication technologies the most (more than 10 times a day) are also those who were distracted most (see Figure D.6). However, the figure also shows that there are certainly discrepancies in participants' ratings of distractions versus their ratings of use of the technologies. This is seen most prominently in the case where the individual who stated they never get use communication technologies and yet is highly distracted by them. We can observe that for almost every bracket of the frequency of use

.

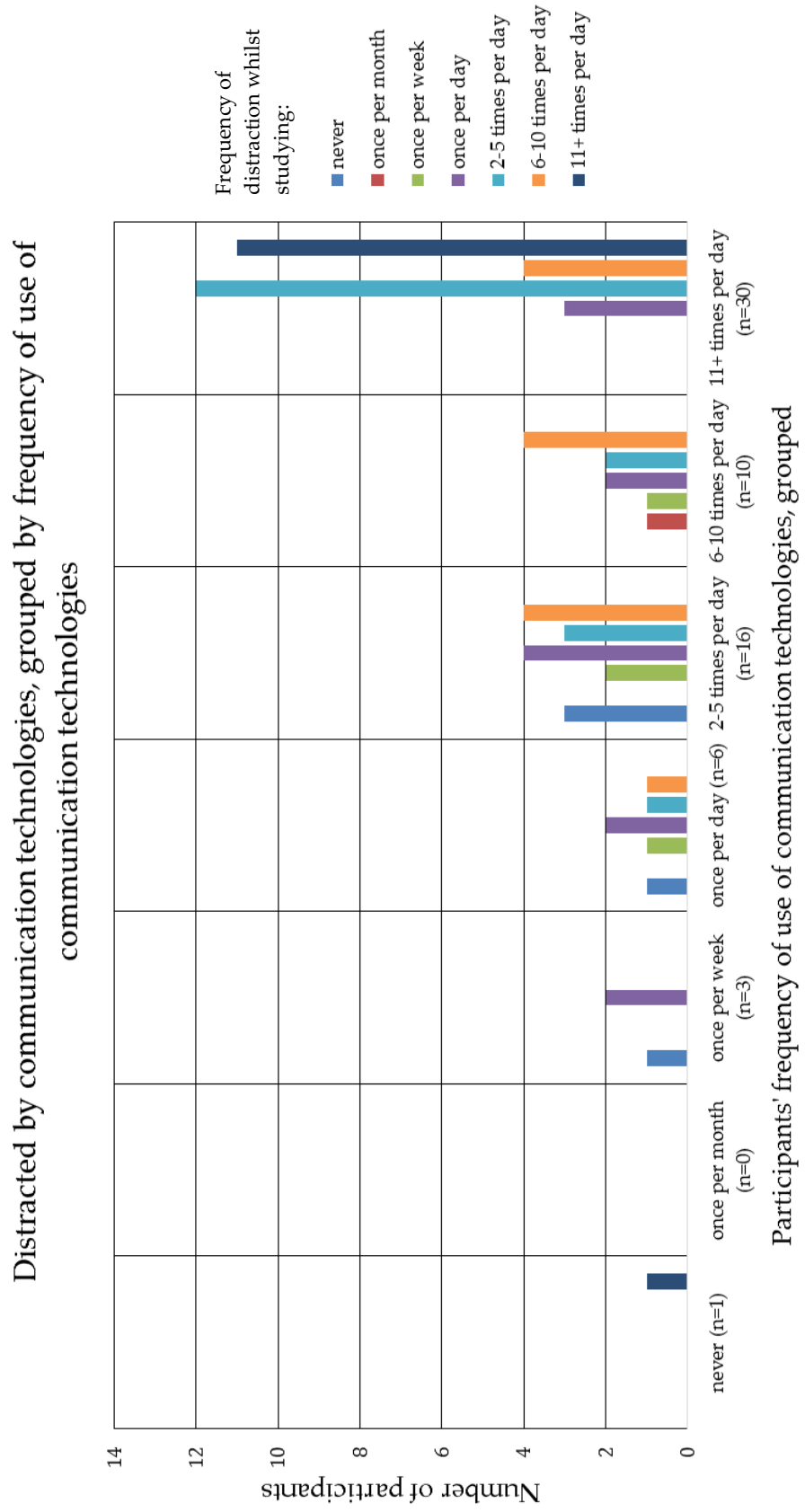


Figure D.6. Pre-experiment questionnaire data on self-rated distraction levels from communication technologies, grouped by frequency use of technologies

D.4.2 Performance outcomes

The times taken to complete the reading task for each condition are shown Figure D.7 and participants measured comprehension levels for each condition are shown in Figure D.8. L2 readers take longer to complete the reading task, for all conditions. Contrary to our predictions, there is no visible increase in time taken to read the hard text compared to reading the easy text. The distraction mitigation signals appear to only affect the L2 readers, however in the opposite way to what we expected – reading time increases for the signal conditions.

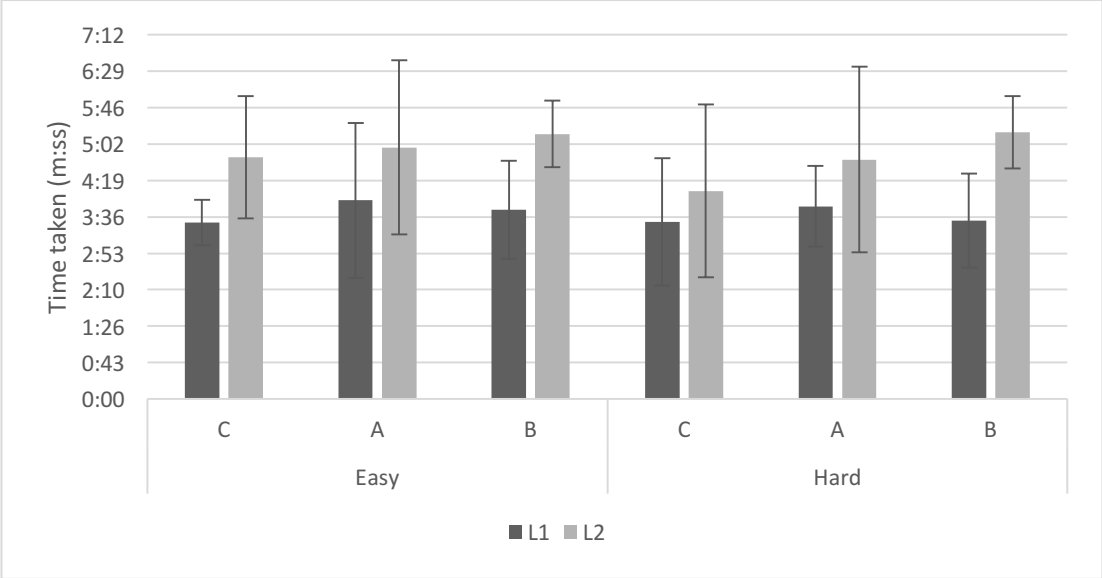


Figure D.7. Time taken to complete for each condition

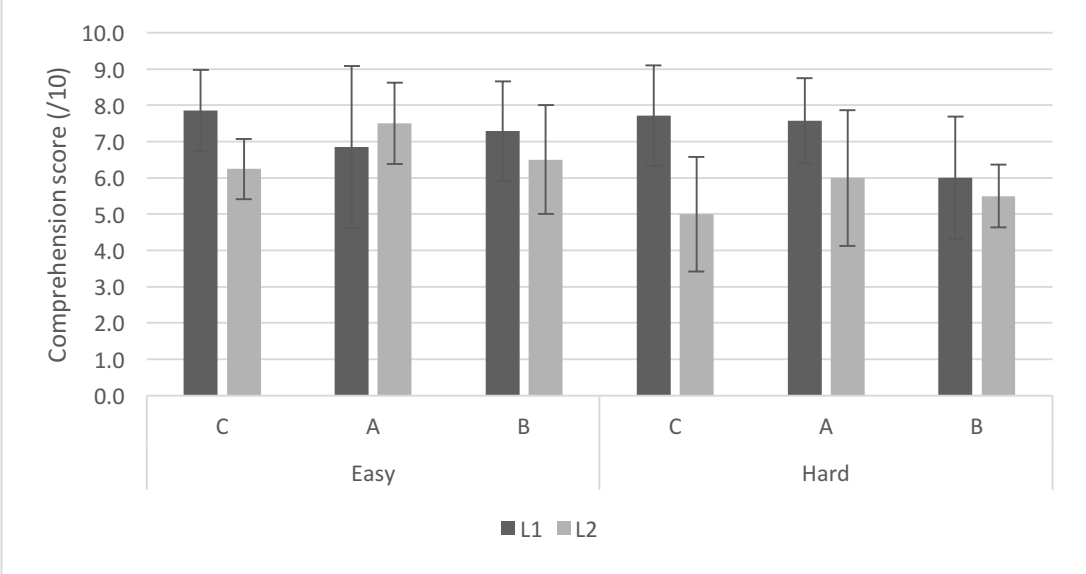


Figure D.8. Comprehension for each condition

Additionally, we can observe from Figure D.8 that in most cases L1 readers score higher on the comprehension tests compared to the L2 readers. The distraction mitigation signals do not appear to help the L1 readers, if anything there is an

observable decrease in reading comprehension when the signals are used. The opposite is seen for the L2 readers where an increase in comprehension score is seen when the signals are used.

To address the above hypotheses a MANOVA is used to determine if there are any statistical differences between the conditions. The correlations between the dependent variables are within the acceptable limits for MANOVA outcomes, i.e. the correlations lie between r=-0.4 and r=0.9. To test for normality in the dependent variables the Shapiro-Wilk Test is used, as it is more appropriate for small sample sizes. The quiz scores are normally distributed for all formats. The times taken are normally distributed for both of the L1 and L2 data sets. The comprehension scores are normally distributed for the L2 data set and not for the L1 data set. The Levene's test for equality of variances shows that there is homogeneity for all dependent variables (significance>0.05). Finally, the homogeneity of variance-variance-covariance matrices is satisfied as the Box's M value of 36.92 (p=0.653).

The MANOVA shows there is a statistically significant difference between L1 and L2 participants, $F(2,53)=10.94$, $p<0.0005$; Wilk's $\lambda=0.708$, partial $\eta^2=0.292$. However, there is no significant difference based on text difficulty, $F(2,53)=1.82$, $p<0.172$; Wilk's $\lambda=0.936$, partial $\eta^2=0.064$, or text signal condition, $F(4,106)=0.818$, $p<0.516$; Wilk's $\lambda=0.945$, partial $\eta^2=0.030$. Additionally, there is no significant effect of interaction between the format and reader type. Since statistically significant results have been found between-subjects ANOVAs are performed. L1 readers have significantly lower reading times ($F(1,54)=13.25$; $p=0.001$, partial $\eta^2=0.197$) and higher comprehension scores compared to L2 readers ($F(1,54)=6.36$; $p=0.015$, partial $\eta^2=0.105$). The difference in reading duration is not only consistent with similar research (Kang, 2014) but also with results from this thesis. The differences in comprehension score is consistent with the results from Chapter 7 of this thesis that showed that there is a difference between L1 and L2 readers when the difficulty of the text is increased. Whilst there is no statistically significant difference in comprehension scores based on text difficulty we can see that the difference between the L1 and L2 readers in comprehension scores largely comes from the hard text conditions.

D.4.3 Eye gaze and distractions

The comparison of percentages of fixations on the distractions and the distraction rates for each of the conditions are shown in Table D.3. MANOVA analysis of the eye gaze measures cannot be performed as the data violates the preconditions of the test. However, we can make observations about the recorded data. In all cases there are low distraction rates, as shown in Table D.3. On average L1 participants only look away from the text about 5 times and L2 participants only look away from the text about 4 times. Even when participants did get distracted they spent relatively no time looking at the distraction. For the L1 participants, only about 2.4% of the fixations were recorded on the distraction area and only 1.9% for the L2 participants. Our expectation was that there would be a higher level of distraction. However, two key points can be made from these results; firstly, the L2 participants tend to be distracted less than the L1 participants, and secondly, there is

considerable variation in the distraction of participants, as seen in the large standard deviations. The latter point suggests that the amount to which an individual is distracted is largely based on the characteristics of that individual.

Table D.3. Distraction rates for each experimental condition

Text Readability	Mitigation Condition	Reader Group	Eye Gaze Measures			
			Total number of fixations	Total fix. dur. (m:ss)	% fixations on distractions	Number of distractions
Easy	A	L1	532 ± 270	2:12 ±1:29	2.4 ±3.0	5.9 ± 3.2
		L2	464 ± 566	1:37 ± 2:07	1.8 ±2.5	3.8 ± 5.0
	B	L1	519 ± 231	1:53 ± 0:59	2.3 ± 1.4	5.0 ± 2.2
		L2	550 ± 208	2:12 ± 1:07	1.3 ± 3.7	2.3 ± 1.0
	C	L1	452 ± 129	1:33 ± 0:36	1.9 ± 8.3	3.6 ± 2.3
		L2	560 ± 63	2:23 ± 0:25	3.9 ± 23.0	5.8 ± 3.3
Hard	A	L1	613 ± 135	2:22 ± 0:47	1.8 ± 11.1	4.4 ± 5.3
		L2	692 ± 298	2:35 ± 1:11	1.5 ± 3.1	5.3 ± 2.5
	B	L1	477 ± 128	1:44 ± 0:35	3.8 ± 12.1	6.3 ± 3.8
		L2	512 ± 258	1:59 ± 1:08	1.4 ± 3.0	2.5 ± 2.4
	C	L1	564 ± 245	2:08 ± 1:08	2.1 ± 4.7	5.4 ± 4.9
		L2	400 ± 310	1:35 ± 1:13	1.6 ± 1.1	3.5 ± 3.1

In most cases the L2 readers have higher numbers of fixations and longer fixation durations. Notably, this is not the case for the hard text condition C, where the L2 group has a considerably lower average number of fixations and fixation duration. This is an interesting point given the results from Chapter 6, 7 and 8 of this thesis that highlight that L2 readers tend to stop reading thoroughly when a text becomes too difficult for them. However, for the other two conditions, A and B, in the hard text condition, the number of fixations and fixation durations are certainly higher than for the L1 readers. In these cases, the distraction signals may not have worked in mitigating distractions but perhaps have helped the L2 readers to keep reading the text more thoroughly than if the signals were not there. Our conclusion therefore is that the distraction mitigation signals do not seem to reduce the amount of distractions but they may provide encouragement to read the text more thoroughly, especially for the L2 readers when reading difficult text.

Finally, the difference in eye gaze measures between the easy and hard texts appears to be minimal, contrary to what we would expect. Further analysis using more participants is required to investigate this further.

D.4.4 Participants' perceptions

After the reading and comprehension tasks participants were asked if they were: 1) distracted whilst reading; and 2) whether they thought this had an impact on their

understanding. Of participants, 82% stated that they were distracted whilst reading and 61% stated that it did affect their comprehension.

There is no difference found in the perceptions between L1 and L2 readers using Chi-square test for independence ($\chi^2(1)=1.99$, $p=0.16$) but there is a relationship between the language group and whether the participants thought the distractions affected their understanding ($\chi^2(1)=4.99$, $p=0.03$). Of the L1 participants, 50% thought the distractions affected their understanding, whereas 79% L2 participants thought the distractions affected their understanding.

Again using the Chi-square test for independence, the distraction mitigation signal conditions were found to have no relationship to whether participants thought they were distracted ($\chi^2(2)=0.26$, $p=0.88$) nor whether they thought the distractions affected their comprehension ($\chi^2(2)=0.72$, $p=0.69$). Finally, text difficulty was found to have no relationship to whether participants thought they were distracted using Chi-square test for independence ($\chi^2(1)=0.88$, $p=0.35$) nor whether they thought the distractions affected their comprehension ($\chi^2(1)=0.74$, $p=0.39$).

D.4.4.1 Perceptions of the distraction mitigation signals

For the conditions where the distraction mitigation signal are applied to the text participants were also asked, 1) did you find that the text effect reduced your distraction? And, 2) did you find that the text effect helped you to start reading the text again? The results from this in general point to three main findings; firstly, that the majority of participants did not even notice the distraction mitigation signal in condition B. Only 9% of participants thought that the signal in condition B helped reduce their distractions however, 14 of these participants did not even see that there was a signal. Unsurprisingly, only 14% of participants in the B condition stated that the signal helped them recover after reading.

The second point that can be made is that whilst the signal was meant to be applied with the last word read, this was seldom the case. That is, limitations in the eye tracking accuracy impacted the effectiveness of the signal. This was not picked up in condition B since a large majority of participants did not even notice the effect. However, in condition A the signal was more noticeable and hence the limitation was detected. For condition A, 32% of participants found that the signal reduced their distraction rate. Whilst this is a low percentage, for those that it worked for it did do the job it was supposed to do with participants noting: "*Yes it showed me I was distracted*" and "*Yes as it went a bright colour and reminded me I should be reading*". But for the rest of the participants the signal was not working correctly with participants noting "*It actually distracted me more than the pictures did because it went to something that I either hadn't read yet or already read*", "*No, it was the reason why I distracted.*" and "*Nope. Very random.*"

Remarkably, even with the effect not working correctly, 55% of participants actually stated that they thought it helped to start reading again. So even for participants who stated that the effect was not working correctly, they still found it helped, mainly because it drew their attention back and made them re-read text. Participants stated: "*It did bring me back to the text a bit.*", "*Some help, it always drag my*

attention to the start point to read again." "*Yes - but it was a bit behind so I re-read the sentence I had previously read*", "*Well it made me reread things*", and "*Yes, I kind of forgot what I was reading after I saw the text effect, and then I just read from the highlighted text again*".

This brings us the third and final point that perhaps it is useful to consider the distraction mitigation signal not being on the word that was read before distraction occurred but to being slightly behind that point, therefore inducing re-reading of the text.

D.5 Discussion

In this study we investigated two methods for mitigating distractions during reading. The insights gained from this study come from several directions, the first of which is in regards to the pre-experiment questionnaire about usage of distracting technologies during study periods. All participants stated that they use emails and IM but the shock comes from the fact that a large majority (98%) of participants use social media and / or email and / or instant messaging while they are reading materials for university. And 85% of participants admit that these technologies distract them while studying. Almost half (46%) of the participants are using these technologies more than 10 times a day and 85% of them are using these technologies at least 2 times per day. Perhaps more interesting is that 65% of participants admit that they are distracted by these technologies during study at least 2 times per day. This indicates that people are getting distracted whilst reading and studying and therefore there is a need to mitigate these distractions.

We hypothesised that the L1 readers would be associated with higher distractions rates and hence the eye gaze would be more affected in this case. The eye gaze analysis in the study is not conclusive enough to provide evidence for or against this hypothesis, but what they do show is that L2 readers tended to be slightly less distracted than L1 readers. The L2 readers were seen to take longer to read the texts and scored lower than the L1 readers. In general, the L2 readers have higher numbers of fixations for longer durations, as we would expect from past research (Kang, 2014).

The hypothesis that the easy-to-read text would be associated with higher distraction rates was based on past research that auditory distractions impair proofreading and prose recall task performance when the task is easy and not when it is hard (Halin, Marsh, Haga, et al., 2014; 2014). However, there are several differences to these studies, mainly being, the distraction type and the way in which the text is made difficult to read. In our study the visual distractions we used may not have been distracting enough. Participants on average fixated about 2% of the time in the distractions area which is a very small percentage and raises the question of whether the environment is actually "highly" distracting or not.

The visual distractions were an experimental condition and not entirely a realistic situation. However, the images rapidly change, which is common for advertising on webpages as well as the rapid changes that occur in social media site

such as Facebook. The choice was made to not use a real scenario, i.e. a webpage with changing adverts, because the objective of the experiment was to control the distraction rate to keep it constant for all participants. Changing the images at a random rate could perhaps increase the level of distraction.

Another explanation that is that whilst attention grabbing irrelevant images and animations alongside text material have negative effects on learning (Clark & Mayer, 2011; Harp & Mayer, 1998; Mayer et al., 2001; Sung & Mayer, 2012), decorative images have been found to have neither a negative nor positive effect on learning (Sung & Mayer, 2012). The images chosen have only a covert association with the topic in that primarily they are digital images and the topic of the text was on digital images. Given that the images have no overt association to the topic they are perhaps more similar to decorative images rather than seductive images. In either case it would be desirable to redesign the environment to be more overt in distracting participants.

The second difference from previous research on auditory distractions lies in the fact that task difficulty was altered using the readability of the text rather than by changing the font used. The reason for this is because we are interested in investigating reading behaviour and the effects of distractions on reading. This is different to previous studies where only the outcomes of reading, in terms of comprehension, recall, or time taken, and not the reading process itself. There is a large body of research on reading behaviour that we can compare against. For these reasons, we decided to change the readability instead of the font. In the study a sans serif font was used throughout the whole experiment, namely Verdana. However, the hard to read font used by Halin et al. (2014b) was the sans serif font Haettenschweiler and the easy to read font was serif font Times New Roman. In follow-up studies the use of Times New Roman as the font for text display could be tested to see if the font indeed has an effect.

Finally, we hypothesized that the signals would reduce the distraction rate and help the reader recover after being distracted. Neither signal used in the experiment was found to affect the distraction rate; however, the distraction rates themselves are quite low. Even though on average participants were distracted about 5 times during the reading task, the distractions were short with only about 2% of recorded fixations lying on the distractions. Therefore, it is not surprising that the mitigation signals had little overt effect. Additionally, the distraction rates are highly variable between participants indicating that some participants are much more easily distracted than others.

D.5.1 Implications for eLearning

The pre-experiment questionnaire shows that there is a problem with participants being distracted by communication technologies whilst studying. There is a need to mitigate these distractions and help students in their learning. Attention guiding could be used to both minimize distraction of the learner as well as draw the learner's attention to the important or relevant parts of the learning material.

Another use of adaptive eLearning is to overcome the effects of distractions. Detection of distractions of readers could be used to determine whether text should be reshown to students. Additionally, labelling parts of the text that the reader was highly distracted during reading could be used to either show the student where they were distracted or be used to control what content is re-shown to the student, where the parts of text that the student was highly distracted during reading could be re-shown.

D.6 Future work

The study showed interesting results about the presence of distractions during reading and the potential of distraction mitigation signals, especially for L2 readers. However, the results from the study are preliminary, primarily due to the fact that more participants are needed and that better eye tracking technology needs to be used to produce more accurate eye tracking and thus better implementation of the distraction mitigation signals. Follow-up experiments are suggested to address these limitations of the experiments.

Furthermore, given the relatively low distraction rate observed in this study, it is suggested that the environment be made more distracting and have more overt distractions. In this way we could see if an even more distracting environment causes more distractions and therefore has a more prominent effect on eye gaze and reading behaviour. The optimal setting for this would be the use of wearable eye trackers that monitor the student being distracted off the laptop screen as well. Thus, we can induce more distractions such as those that come from mobile phones or televisions, as well as the onscreen distractions that were proposed. Additionally, we observed that some participants are more easily distracted than others, trying a within-subjects design could control for this.

We never investigated the case where no distractions are given to the reader. Including this case would allow us to investigate how, or if, distractions alter reading behaviour of participants. Additionally, this would allow us to investigate further the effects of text readability on distraction rates.