UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ANALYSIS OF THE RELATIONSHIPS BETWEEN EYE TRACKING INFORMATION AND TEXT COMPREHENSION LEVELS IN HEALTHCARE

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

LUCAS MOSONI
Norman, Oklahoma
2017

ANALYSIS OF THE RELATIONSHIPS BETWEEN EYE TRACKING
INFORMATION AND TEXT COMPREHENSION LEVELS IN HEALTHCARE

A THESIS APPROVED FOR THE
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING

BY

____
Dr. Ziho Kang, Chair

____
Dr. Charles Nicholson

____
Dr. Randa Shehab

To my family who supported me all along in this life journey and to my friends who became my family away from home.

Acknowledgements

First, I would like to thank my thesis advisor at the University of Oklahoma, Dr. Ziho Kang, for giving me the opportunity to work on this very interesting data set. More than an advisor, he was also a professor who taught me numerous concepts of cognitive engineering and data analysis. Many thanks to the HFS (Human Factor & Simulation) lab for welcoming me in their team and especially Jiwon Jeon for helping at the beginning of the project to understand the procedure of the experiment that she helped organize, and to discover the eye tracker and its software.

I also thank Dr. Randa Shehab and Dr. Charles Nicholson for agreeing to be in my committee. Dr. Charles Nicholson was also a great professor teaching many data analysis and machine learning concepts that I have used in my thesis and that I will probably use in my future professional life. Thank you Dr. Randa Shehab for being a great help with the multiple administrative problems that I encountered when applying and arriving to OU.

Thank you to the School of Industrial and Systems Engineering and my school ISIMA for keeping this dual degree alive which allows a few lucky students like me to discover the American college life through the beautiful University of Oklahoma.

Table of contents

List of Tables

List of Figures

Abstract

Health information can be found more easily than ever through a variety of digital texts: from health forums online, informative websites, or phone apps that monitor your health. It is important that the patients are able to comprehend this information correctly as it can lead to important decisions regarding their health. However, the average patient will have difficulties understanding digital texts as they contain medical terminology. Reading such content requires multiple cognitive processes and memory to fully grasp the communicated information.

The data used in this project comes from an eye tracking experiment involving human subjects reading health related texts and answering questions about these texts. The eye tracking data includes information about the number of eye fixations, fixations' durations and positions in different areas of interest of the screen. The read paragraph also had different complexities and there were different types of questions. First, a quantitative analysis involving multiple ANOVAs was performed to uncover the effect of different chosen factors. After some feature engineering guided by the previous analysis, machine learning was used to classify whether a participant would answer correctly or not using eye tracking data and variables relative to the text and questions.

The initial analysis showed that increasing levels of complexity of the digital text resulted in more eye fixations and longer reading times and that different questions provoked different reading behaviors. More importantly, the random forest classification model was successful with a 96% accuracy thanks to the engineered features, showing that human decisions can almost be predicted using only our eyes' behaviors.

Chapter 1: Introduction

Understanding health related information is generally harder than understanding non-scientific information. Indeed, the presence of complex terms and processes in the digital text can confuse the reader. In addition, medical information can be found more and more online: multiple organizations and people publish on the Internet this type of information and 80% of Internet users already encountered some on their personal computers (Fox, 2011). However, the reliability and quality of this information cannot always be trusted (Eastin, 2001). It is then even more important that readers understand the content of those websites in order to sort what information can be trusted or not. It is true that this type of information is highly impactful and influential as readers can take decisions related to their own health. This is emphasized by the fact that healthcare in many countries is expensive and people use online information in order to solve their health problems without cost. Furthermore, not only this information found online or within healthcare applications can be used to educate ourselves, but it can also help doctors and patients to make decisions within a healthcare institution. One of the first step to making good decisions related to health is the comprehension of the patient's current medical situation.

It has been now established that understanding health information is essential in many ways. And that this comprehension is the product of an efficient read. But how can we evaluate if the reader was efficient and therefore understood the text correctly? In research, eye tracking is often used in order to test the usability of applications or different types of contents or layouts. Eye tracking could then also be used to analyze reading behaviors and then correlate them with the textual comprehension.

In the context of this research, an experiment involving human subjects was carried out. During this experiment, the participants read texts about different health conditions and then answered questions to evaluate their comprehension while having their eye movements tracked by an eye-tracking device. Once the data was collected and cleaned, a thorough analysis involving statistical tests, feature engineering and machine learning, was performed in order to correlate reading behaviors to the comprehension of health related texts.

This research’s goal is to try answering the following questions: What are the eyes’ behaviors related to reading health related texts and answering questions about them? How the visual exploration has an impact on textual comprehension?

To answer those questions, Chapter 2 presents the background and the literature review which will present related research. Then, Chapter 3 introduces the methods of analysis used in this research including every step before the results were obtained. Chapter 4 presents the results of every analysis performed. Chapter 5 discusses the obtained results and evaluates if they are relevant. Chapter 6 introduces some limitations of this analysis and the future research that could be done on this subject. Finally, Chapter 7 concludes this research.

Chapter 2: Literature review

In this chapter, I will present research in the field of healthcare, eye tracking used to assess online or desktop applications, visual exploration and textual comprehension.

2.1 Health information online

As said in the introduction, the internet has become a source of health information and advice for many adults. Sillence et al. (2006, 2007) studied how fifteen women looked for information about menopause online over four consecutive weeks. The results showed that when the design of the website was rich and responsive, the participants trusted more the website and therefore consulted them more. Some websites with high quality health information but poor design, were often rejected. They combined the online advice with the advice of friends, family and physicians to take their final decision. The participants judged that the Internet helped them in their decision process. This results are very positive for the Internet but this is in the case that Internet does not replace the visit to a physician. Therefore, reading online healthcare information could be considered the first step to the decision process with the last step being the doctor's visit. If the patient does not intend to go to the doctor, it is extremely important that the read information is understood.

2.2 Reading comprehension and eye tracking

The article written by Just & Carpenter (1980) is relatively old but it gives interesting information about reading comprehension. The authors studied the eye fixations of 14 college students reading scientific passages. They also provided a theory concerning reading comprehension which can be seen in Figure 1.

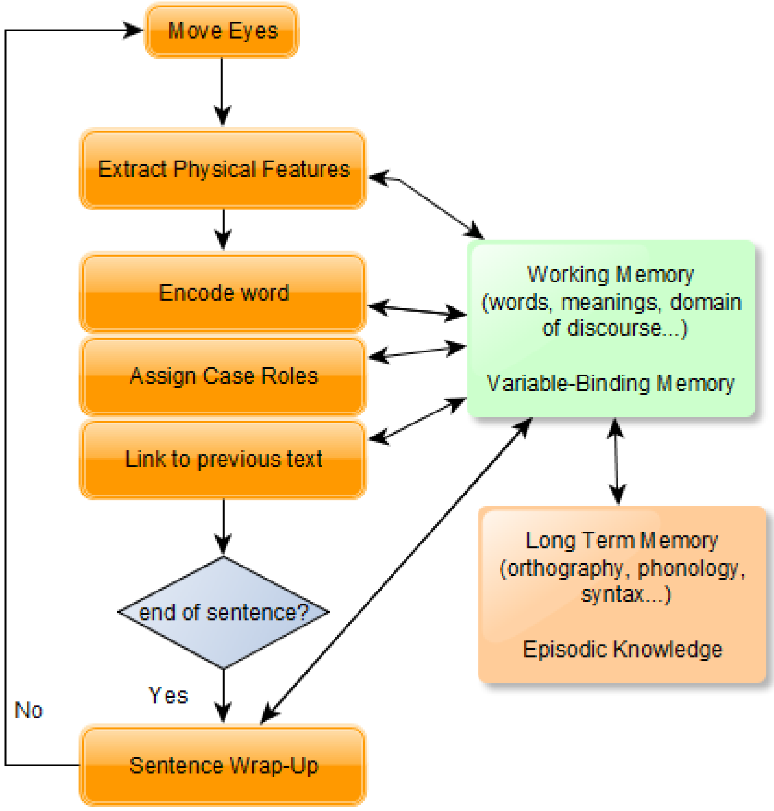


Figure 1 - Major processes and structures in reading comprehension (condensed version extracted from Just & Carpenter, 1980)

Figure 1 shows that while reading a text, the reader uses both his working memory and long term memory. His working memory is used to activate representations and bind words with easy concepts to make sense of what has been read. The long term memory allows the reader to read the word correctly using orthography and phonology. We can expect the long term memory processes of every participants to be quite similar as participants know how to read correctly even though some of them might be more efficient than others. Reading a complex term (for example a scientific word) probably uses the long term memory to spell it, pronounce it and check if the reader knows it or knows a similar word. The working memory is where the textual comprehension really happens and where the words become concepts known by the reader.

For their experiment, the authors used an eye tracking device to measure the gaze duration on each word of the passage. Their results showed that readers made longer pauses where there was a larger load of information to process. This happens when the reader is reading infrequent words, integrating information from important sentences of the text, and making inferences when reaching the end of a sentence. This is probably when the reader is using his long term memory to understand a word.

Mcinnes and Haglund (2011) evaluated the readability of several websites using different readability tests (Gunning FOG, SMOG, Flesch-Kincaid and Flesch Reading ease). The names of 22 health conditions were entered into five search engines and the readability of the first ten results for each search were evaluated. The tests showed that those websites were considered difficult to read. Websites on certain topics were found to be even harder to read. For health conditions with several names, the easiest name gave the most readable results. The difficulty tended to increase when reaching the concluding paragraphs. The authors found that the most frequent search results such as Wikipedia were some of the most difficult. This articles gives a first proof that health related texts are harder to read. Leroy et al. (2008) reached the same conclusion but their method, which used a toolkit of linguistic metrics was tested with real user in order to check if their method was accurate with real readers. Kim et al. (2007) created a metric which is more appropriate for medical/consumer health domain, that is to say that they tested their method on 4 different materials instead of just health related websites: consumer health texts, electronic health records, health news articles and scientific biomedical journals. Their results were more accurate than general readability methods as the results differed more for different types of content that the general methods judged similar. Estey, Musseau & Keehn (1991) found out that the health printed documents provided to patients are often way too complicated for the patients to understand them. That is why they conclude that those documents should be simplified.

Granka, Joachims & Gay (2004) proved that several factors can have an influence on eye movements by attracting the attention of the reader with a specific design. In their study, they observed sequences of eye fixations and they found out that the size of the text and the use of colors improve the quality of lecture. In the case of our experiment, the interface's design presenting the paragraphs and questions has been homogenized so that there is no unwanted expect of purely design factors. The impact of design factors was also studied by Kules & Xie (2011). They demonstrated that adding graphical elements and a dynamic interface where elements to be read moved and interacted with the reader, were favorable for a better comprehension and memorization of the information.

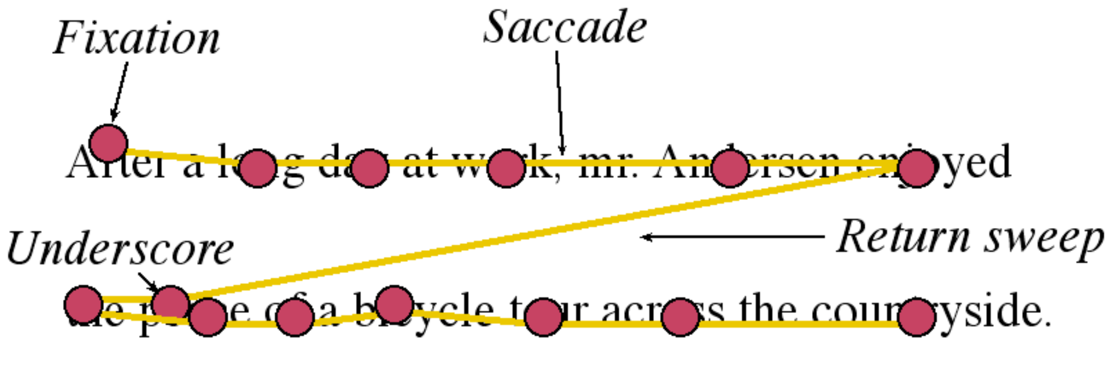

Fixation
Saccade
After a long day at work, mr. Andersen enjoyed
Underscore
Return sweep
the pleasure of a bicycle tour across the countryside.

Figure 2 - Eye movements when reading a text

Figure 2 shows typical eye movements associated with the reading of a text (Holmqvist & Wartenberg, 2005). When reading from left to right, it is logical that the fixations form a line sequence from left to right. The fixations follow each other quite fast as the human being is not able to cover a big quantity of words with only one fixation due to his memory and vision. It is also true that even if a fixation hit a specific point, the human vision allows to concentrate in the area of the point. Moreover, because the eye tracking device does not have a perfect precision, we should consider the fixations points more as the center of an area of attention, than think that the reader only focused on the fixation point. A saccade is a very fast eye movement when moving from a fixation to another. When the eyes reach the end of a line, the reader comes back to the left of the paragraph to the next line: this saccade can be called return sweep. After that, a lot of different things can cause the reader to fixate one word in particular: incomprehension, inattention, just randomly or a thought associated with this word. All of this reading behaviors apply to this project which gives ideas on what to expect in the results.

Bell (2001) studied extensive and intensive reading. Their goal was to compare the reading speed and textual comprehension of those two types of reading. Intensive reading means that the text read has been imposed and that the reader is conscious that there will be an interrogation on what has been read, which is often done in a limited time. On the opposite, extensive reading does not have any constraints: no time limit, the text was chosen, and no interrogation after the reading. Therefore, we could imagine that pressure would have a positive impact on reading speed and comprehension. On the contrary, they showed extensive reading obtained a higher reading speed as well as a better comprehension. In the case of our project, the reading might be intensive as there is an interrogation but this interrogation has no impact and there is no time limit. However, we can imagine that if the subject of the paragraph is particularly interesting to the reader, then the reading can become more extensive and the reader might do better.

The authors of this article also showed that comprehension and reading speed are strongly correlated. For example, a slow reader might have a poor comprehension as he forgets the text as he reads it. However, reading fast is also not very helpful. Therefore, it is important to read at a natural speed and to find the right balance.

2.3 Comprehension of health information

Soederberg et al. (2011) studied the mechanisms underlying the comprehension of health information. They were specifically interested in prior knowledge, working memory capacity, the integration of concepts in the text and the age of the participants. They performed an experiment involving more than 200 participants aged from 18 to 81. The participants read two paragraphs about nutrition. The working memory capacity was assessed by a test checking how many statements the participants could memorize. The participants were then tested on the newly acquired knowledge. The prior knowledge was evaluated using a 38-item multiple choice nutrition test. The results showed that prior knowledge was a significant factor of the acquisition of new knowledge especially for older adults. Both prior knowledge and working memory capacity promotes information processing. It also showed that older adults with a declining working memory capacity relied more on their prior knowledge. In the frame of this thesis, the prior knowledge and working memory capacity are also factors which might have an effect on the comprehension. The prior knowledge will be decreasing when the health condition described is more complex and less known. The participants of our study are relatively young so the effect of the working memory capacity can probably be ignored as it should be approximately the same for all participants.

Now focusing on healthcare application, more and more are created in order to help doctors provide health care. For example, Eghdam et al. (2011) studied the efficiency and the satisfaction of doctors using the prototype of an application called Infobiotika. The goal of this application is to help doctors administer antibiotics in the context of intensive care. For this type of care, antibiotics are often used more than they should because of a lack of knowledge of the patient’s current state. The research idea was to create a prototype where a lot of information on the patient’s file was available such as radios or bacteriological reports. This application could support doctors in their decisions. For the experiment, the authors compared results for two different types of populations: specialists and unexperienced physicians. They were asked to navigate within the application to find information, and then perform clinical tasks. The goal was to measure the comprehension of the patient’s file by the physicians.

As for our own experiment, eye tracking was used to analyze the participants’ eyes movements. The procedure is also similar as the participant first find and read information and then their comprehension is tested when they perform the task. The difference is that the participants are also tested on how they are able to navigate in the application to find the right information. The results showed that less experienced physicians explored the screen more, and looked at visualizations more than the specialists who were faster and spent more time looking at the data directly. For our experiment, it would have been interesting to study different populations like medicine specialists and novices. However, it is possible for our experiment to study the effect of age. It would maybe reveal that older participants gained more knowledge about different health conditions.

Some new applications also allow the patients themselves to enter their own medical information in a regular manner. It is the case in the article presenting an application enabling patients with cancer to report their symptoms regularly (Bansback, 2014). Therefore, the patients and their doctors could follow the evolution of their symptoms.

Eye tracking was used again for usability testing. Indeed, by observing the visual exploration patterns, the authors were able to find critical paths or tab names or buttons which made the participants confused. A "think aloud" method was used meaning that the participants were asked to share their impressions and feeling while using the application. The participants were then able to share their difficulty to understand some scientific medical terms encouraging the designers to add some help bubbles showing the definition of those terms.

This aspect of the article is very similar to what we can expect in our experiment. Indeed, the presence of complicated terms will be directly linked with the textual comprehension and it will probably show in the questions' results. A limitation of this research which is also present for this thesis is that because the participants were recruited within a University, participants are relatively well educated and used to technologies. It is possible that with a more representative sample of the real population, participant would be less able to understand the information and their eye movements and analysis results could be entirely different.

2.4 Machine learning and eye tracking

Judd et al. (2009) tried to predict where people look into a picture. They first created a large database of eye tracking experiment with labels and analysis. Their second contribution was to implement a supervised machine learning model which use the picture info (bottom-up image-based saliency cues and top-down image semantic dependent cues) to predict where the eye fixations will be located. The final output takes the shape of a heat map. Their machine learning model was a support vector machine. The machine learning analysis done in this research is different from ours. Indeed, they take image information to predict eye fixation information. We intend to use eye tracking info to predict comprehension. But it is still interesting to see that some machine learning was done in the field of eye tracking.

It is true that machine learning is not yet often used combined with eye tracking, and even less in the field of reading comprehension. Kunze et al. (2013) tried to detect what was the type of document a participant was reading given his eye fixations information. The experiment involved eight participants and five different document types (a novel, a manga, a fashion magazine, a newspaper and a textbook). They achieved a recognition performance of 74% using user-independent training meaning that the information of the identity of the reader was not given to the algorithm. For this thesis, the machine learning was also done user independently as we want a general method describing behaviors found for all participants. Concerning the results, the best performances were obtained by the novel and the textbook because participants read the structured text line after line. The magazine recognition caused a problem for male participants as they showed less interest in pictures and just read the text fast. Moreover, the newspaper and manga were fixated quite similarly. In a general way the results were very dependent on the presence of pictures or not.

Chapter 3: Methods

This chapter presents the experiment, the data and how it was obtained and the methods of analysis used in this project. First, the experiment providing the data will be explained including participants' information, what eye tracking device was used and the procedure of the experiment and its purpose. Once the experiment done, the data was extracted and cleaned. The first step of the analysis was to perform a statistical analysis to have a first impression of what the data hid including which variables had an effect on the fixations. The next step of the analysis was to find features which could be particularly correlated with the textual comprehension and which might give interesting insight on reading patterns. Finally, using the engineered features, the fixations' information and information relative to the content read and the participant, machine learning such as regression and classification was used to predict comprehension of participants.

3.1 Participants

The experiment has been done within the HFS (Human Factors and Simulation) laboratory. 26 participants were recruited through email and flyers at the University of Oklahoma. The participants were relatively young with an average age of 23 years. However, the data coming from one participant was not used because it was highly corrupted and of a very bad quality. The participants filled a consent form given at the time of their individual appointments. If the participants were American citizens or if they had a Green Card, they were given a $10 Wal-Mart gift card.

3.2 Apparatus

The eye tracking device used for this experiment was a Tobii TX300 (Figure 3). It collected the eye movements with a sampling rate of 300Hz and a precision of 0.4° for binocular visions, it is equivalent to an error of 4.8 mm at a distance of 65 cm.

Figure 3 - Appearance of the Tobii TX300

The eye tracker has a software called Tobii Studio 4 which can be used to extract the data and do a preliminary analysis of the experiment. For this project, Tobii Studio 4 was used to review the recordings by looking at the eye movements of each participant during the experiment. It was also used to define the AOIs (Areas Of Interest). AOIs are zones of the screen in which we are particularly interested and especially interested in the eye movements occurring in those regions.

3.3 Procedure

The goal of the experiment was to analyze the correlation between the different methods of visual exploration while reading and the textual comprehension. Three scenarios involving different health conditions were described in different paragraphs, of increasing complexity. In order, those scenarios were an easy scenario about blood pressure, an intermediate scenario about ringworm infection and a difficult scenario about neuropathy associated with Lyme disease. For each scenario, the text included two paragraphs appearing one after the other: the first one was an introductive paragraph describing the symptoms written by a patient and the second one was written by a doctor who explained the health condition with its symptoms and mechanisms and then gave recommendations on how to cure this sickness. Once those two paragraphs were read, the participant could begin answering the question by making the first question visible. It is important to note that the doctor's paragraph appearing in second position stays on the screen while the questions appear as the questions are on the information given in the doctor's paragraph.

The questions all had five possible answers each and one of those answers was always "I don't know". Each question was about a different aspect of the text: the first question simply asked what symptoms the patient has, the second asked about the cause, the third one asked to select all possible answers among different diagnoses, the fourth one asked to define a scientific term and the last one asked to select the best recommendation the doctor gave.

The texts were extracted from a medical online forum on which users can ask medical question to doctors and get some help from them. The text was slightly modified in order to correct spelling mistakes but also to adjust the size of doctor's paragraphs so that the sizes of the paragraphs from the three scenarios have similar sizes. There was no time limit to the experiment and participants could navigate freely between questions.

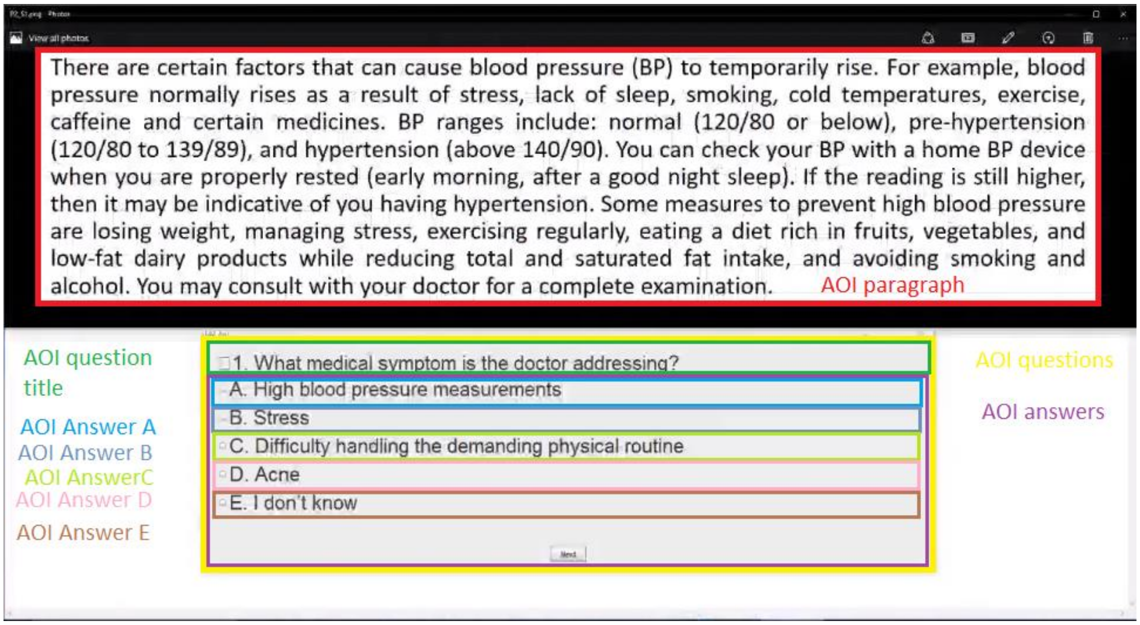
The experiment followed a within-subject design. Indeed, the 26 participants read about the three scenarios. However, to avoid any confounding effects, the order in which they read and answered questions about a scenario was randomized (Charness et al., 2012). If the participants had read about the three scenarios in the same order, this order could have had an unwanted effect such as being tired or confused when reaching the third scenario.

3.4 Data extraction and cleansing

Each participant had a recording and recordings were gathered into Tobii projects. Ideally, all recordings could have fit in a single project. However, recordings were quite long (between 10 and 15 minutes) so only three to four recordings fitted in one project. The extraction into an Excel file works with one project, therefore, several Excel files resulted from the extraction. Those Excel files were badly designed so they needed to be pre-processed later on.

The first step of the data analysis was to use Tobii Studio to watch the recordings including the eye movements on the screen. Watching those eye movements for each participant was useful to detect the fixations' deviations. Indeed, all recordings had a vertical deviation with a size depending on the participant. For example, when participants were reading the first line of the paragraph, their eye fixations appeared with Tobii Studio on the second or third line of the paragraph or on top of it. There was however no horizontal deviation.

The sizes of those deviations were noted for each participant so that the AOIs could be located at the right position: a little bit higher or lower than where the real area was located on the screen. The AOIs also needed to be located at the right moment. For example, the paragraph's AOI need to be activated only if the paragraph is visible by the participant. Once the data was extracted, each fixation was associated with one or several AOIs.



There are certain factors that can cause blood pressure (BP) to temporarily rise. For example, blood pressure normally rises as a result of stress, lack of sleep, smoking, cold temperatures, exercise, caffeine and certain medicines. BP ranges include: normal (120/80 or below), pre-hypertension (120/80 to 139/89), and hypertension (above 140/90). You can check your BP with a home BP device when you are properly rested (early morning, after a good night sleep). If the reading is still higher, then it may be indicative of you having hypertension. Some measures to prevent high blood pressure are losing weight, managing stress, exercising regularly, eating a diet rich in fruits, vegetables, and low-fat dairy products while reducing total and saturated fat intake, and avoiding smoking and alcohol. You may consult with your doctor for a complete examination.

□ 1. What medical symptom is the doctor addressing?

○ A. High blood pressure measurements

○ B. Stress

○ C. Difficulty handling the demanding physical routine

○ D. Acne

○ E. I don't know

Figure 4 - Areas Of Interest without deviations

There were nine AOIs created for this analysis: the paragraph, the question title, the question area (including question title and answers), answer area and each answer individually (Figure 4). Once those AOIs were created and moved for each participant. The data was extracted as eight big Excel files. Each file contains fixation information (one fixation by line) of three or four participants including the recording time in milliseconds, the position on X and Y axis in pixels and if the fixation occurred in an AOI. Those last variables are represented by columns with -1 if the AOI was deactivated when this fixation occurred, 0 if the fixations was not in the activated AOI and 1 if the fixation hit the concerned AOI.

The R language was used for a majority of the data wrangling and data analysis. The R language is a free and open source statistical scripting language (Team, 2000). A lot of packages were implemented in this language with many functionalities in data analysis, data visualization and machine learning. Some visualizations (heat maps and gaze plots) were done using the software Tableau. Tableau is a software focused on data visualization and business intelligence (Heer et al., 2008). It offers many options in visualization and it is very interactive and easy to use.

The data cleansing included several steps. First, the eight Excel files were read in R. Then, the data was re-organized as the design of those files was bad. Each file was converted in a data frame. A data frame is a data structure in R which is composed of same size vectors. Those vectors can contain data of different types: string, integers, date… The "participant" column (participant number) was added to each data frame because it was not included by Tobii Studio. The eight data frames were united in a single one to have all the data in one place. The fixations considered not relevant were deleted: fixations which did not hit any AOI. The variables corresponding to the presence in an AOI were turned into binary variables: when the AOI is deactivated (-1) or when the fixations did not hit the AOI (0) took 0 values.

The hardest part of the data cleansing was to sort the fixations into the different scenarios and questions. For this, it was necessary to write manually the time at which each scenario and question began and ended for each participant. Indeed, even if it was known that the paragraph AOI was hit, it was not possible to know which scenario the paragraph was from, and which question, the participant was answering at that time. A sorting algorithm was implemented: it used the times of the beginning and end of the scenarios and questions converted in milliseconds and the recording time variable of the fixations to sort the fixation in the right category. An additional difficulty was that the order of the scenarios was randomized, so, for each participant different scenarios corresponded to the first, second or third scenario.

Because of the deviations, the positions given by Tobii Studio were incorrect. Therefore, positions needed to be normalized: the fixations which hit an AOI for each participant were normalized using the minimum and maximum on each axis and the AOI real measurements. In other words, the participants' own spaces were turned into a more homogenate normalized single space.

3.5 Statistical analysis

The ANOVA was applied in order to analyze the effect of independent variables on the dependent variables presented later on. The ANOVA allows to see if factors or stimuli and their interactions have a significant effect on some measures coming from a population based on a representative sample. An interaction between two factors means that the effect of a factors varies according to the levels of another factor.

In other words, the analysis of variance checks if the effect of independent variables is significantly higher than the natural variance of a population. The more this effect will be high compared to the natural error, the more this factor will have a significant effect on the dependent variable which will translate in a low value for the p-value. The maximal threshold for considering a factor significant is a p-value of 0.05

because it means that it is 95% sure that the factor has an effect on the dependent variable.

The experiment was designed in order to study the effect of two independent variables on the number of fixations and the sum of fixations durations inside different AOIs. The first independent variable is the scenario with three different levels: easy scenario (i.e. blood pressure), medium difficulty scenario (i.e. ringworm infection) and difficult scenario (i.e. neuropathy associated with Lyme disease). The second independent variable is the type of question with five levels corresponding to the five questions asked for each scenario: the first question about the symptoms, the second about the cause, the third one asking to select all possible answers among different diagnoses, the fourth one asking to define a scientific term and the last one asking to select the best recommendation the doctor gave.

In addition to this independent variables, a few other factors of interest were studied to find some correlation effects. The participants were aged between the age of 18 and 32. Within a University, younger participants are just starting college and are undergraduate students whereas older participants probably are graduate students pursuing a Master’s degree or a PhD. Including a variable investigating differences of eye fixations information between those two types of students might reveal some insight. This variable was divided into two levels with the threshold of 22 years old in order to have a balanced number of participants in each category. A variable indicating how well the participant did on the question(s) might have an effect on the eye fixations information. Having an idea of how the fact of selecting an answer over five could also have such an effect.

In order to analyze the effect of these variables, three different cases of ANOVA were performed. For the first case, the eye fixations information was aggregated into each scenario (case 1). In other words, the eye fixations information was summed during the time of a scenario on different AOIs. For the second case, the fixations information was aggregated by question (case 2). Those two cases allowed to perform the tests on several AOIs: paragraph (initial reading and returns to paragraph while answering), answer area and questions area and question title. Finally, for the third case, the fixations information was aggregated for each individual answer AOI during the time of a question (case 3).

In order to fully trust the results of an ANOVA, three assumptions need to be respected. The first assumption is the independence of the observations. Observations should not be directly linked and there should be no pattern in the data. The fixation information of a participant is definitely not directly related to the fixations information of another participant. Furthermore, the order of the scenarios was randomized so the scenarios can be considered independent as there are no effect of the order. However, for the cases 2 and 3, there are several samples coming from the same participant and the same scenario but we will assume the independence as they were reading different questions or answers. Indeed, fixations information stays quite random and it is hard to imagine direct dependence between a question and another.

The second assumption for ANOVA to function properly is the constant variance assumption. Indeed, the variance of the dependent variables should be similar across the different groups. When doing an ANOVA with R, it is possible to create a plot verifying this assumption.

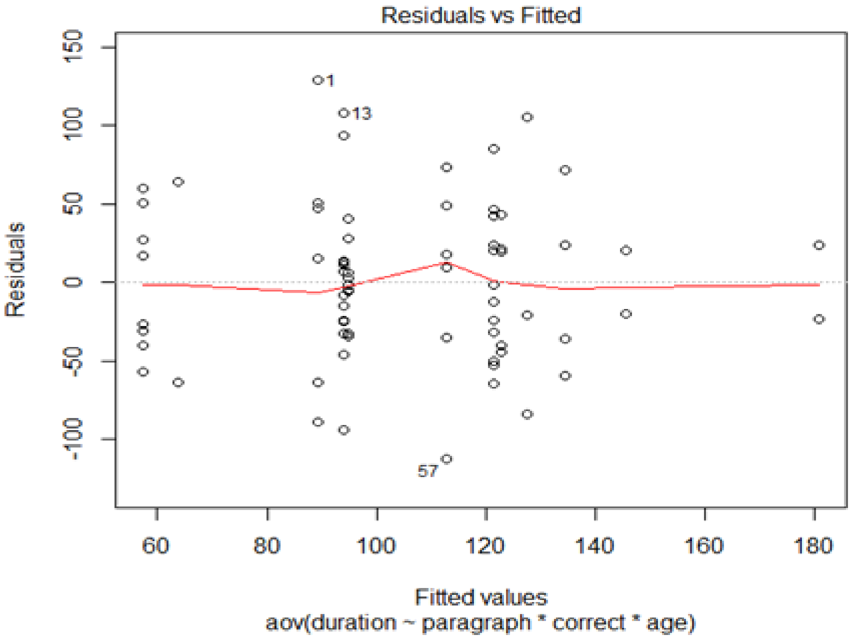


Figure 5 - Plot verifying the constant variance

If the red line visible on the plot is flat enough then it can be considered that the constant variance assumption is verified (Figure 5).

Finally, ANOVA needs the observation to follow a normal distribution. The risk of not respecting this assumption is an increased false positive rate, meaning that some effects can be found significant even though they are not. Before each ANOVA, a Shapiro-Wilk test was done in order to check whether the data was normally distributed.

There are three cases of ANOVA in this project: the number of samples increasing with the precision of the case. With a large number of samples, the means of those samples are approximately normally distributed even when the population is not normal. So for the first case of ANOVA with the scenarios, it is necessary to normalize the data using the Box Cox transformation presented later.

For cases 2 and 3, the normality assumption was again not respected. For a majority of tests, this non normality was due to skewed data: a majority of variables had the measures close to zero. One solution could be to delete some observations which are considered outliers, for example, delete all observations which have a measure of zero. This solution is not the best. The data comes from an experiment involving human subjects so we have a limited amount of data. Moreover, the observations with null measures have meaning: a participant did not look at an AOI for a reason.

In order to reach normality, the measures with zero values were deleted and the normalization was performed. The results were then compared with the ANOVA results without deleting the zeroes. The effects found significant were the same in both situations with a decreasing p-value for the normalized data. So there were no false positive effects found with the non-normalized data. Therefore, the results presented for the cases 2 and 3 were obtained with non-normalized data.

The normalization method called the Box Cox transformation was applied (Sakia, 1992).

$$\begin{cases} y = \dfrac{x^{\lambda} - 1}{\lambda} \text{ where } \lambda \neq 0 \\ y = \ln x \text{ where } \lambda = 0 \end{cases}$$

Figure 6 - Box Cox transformation

The “y” is the new value of the dependent variable calculated from the old value “x”. The “boxCox” function in R allows to find the optimal lambda in order to have the most normal data. If the optimal lambda found is 0 then the algorithm applies the natural logarithm to the data.

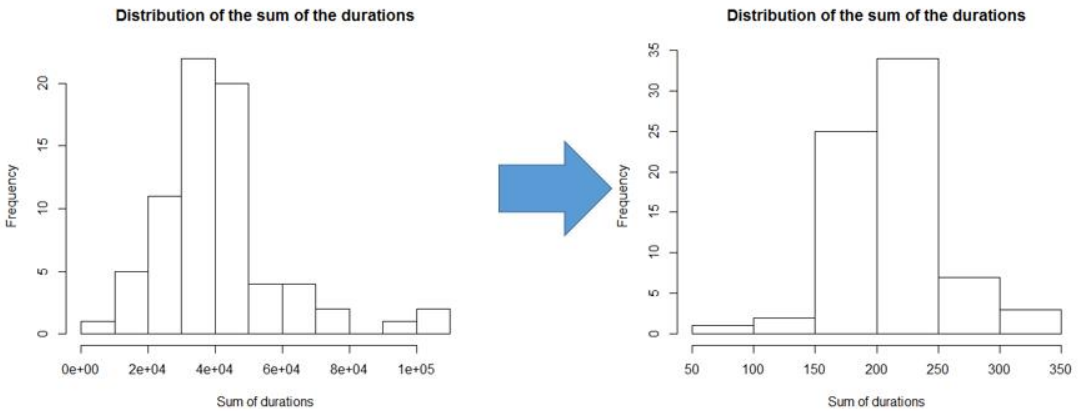


Figure 7 - Example of distributions before and after Box Cox transformation

On the left distribution of Figure 7, the data approximately follows a normal distribution except for some outliers on the right which are not present on the transformed data on the right distribution.

3.6 Feature engineering

In order to perform machine learning, doing feature engineering is a very important step to find meaningful variables which have an impact on what we are interested in. The statistical tests and visualization analysis done previously, gave a first impression of what was significant. The goal is to have some kind of insight on whether the participant is going to do well at the questions so that we are able to link the reading behaviors to the text comprehension.

The statistical analysis allowed to uncover which variables were the most significant on the fixations information. These variables will probably be decisive for the machine learning (for example, the scenario and question). Then, of course, the fixations information will provide the reading behaviors needed. This will include fixations duration and number of fixations in any of the following AOI: paragraph, question title and the answer area. We decomposed the paragraph AOI into two different time periods: before the questions appear that we called initial reading and while the question are visible which we called returns to the paragraph.

One feature that we already found is that long fixations were probably meaningful. Indeed, long fixations tend to focus on some terms like seen in the literature. However, it was first necessary to determine from which threshold a fixation can be considered long.

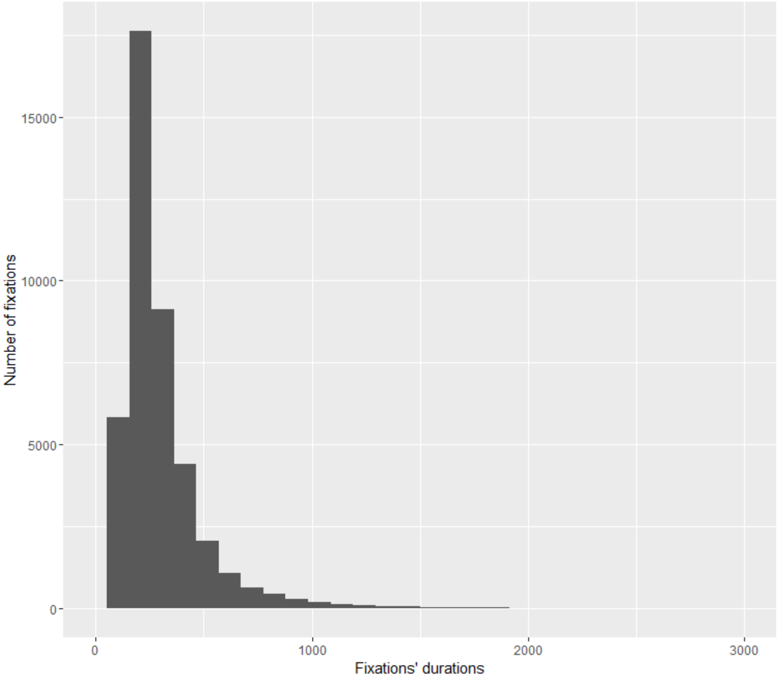


Figure 8 - Distribution of the durations of all fixations

On the above figure, most of the fixations are less than 300 milliseconds (Figure 8). Of course, this concerns all the fixations. In order to really take advantage of this kind of analysis, it is more appropriate to consider fixations which are in a single type of AOI. That is why the found threshold should be different for every AOI.

In order to solve this problem, an algorithm was developed. This algorithm takes as an input the durations of the fixations in a specific AOI.

Pseudo code: Find threshold of long fixations

```
Input : fixationsDurations
maxDuration ← maximum(fixationsDurations)
For threshold from 1 to maxDuration do
    beforeThresh[threshold] ← Number of fixations before threshold
End
Find elbow of curve : calculate double derivative beforeThresh
Return which threshold had the maximum derivative
```

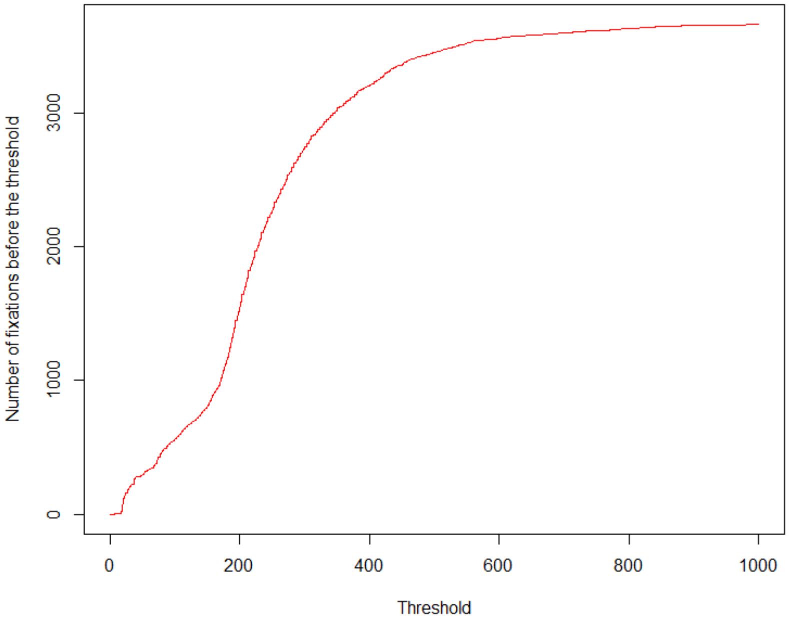


Figure 9 - Number of fixations before threshold VS threshold

This algorithm draws a curve which gives similar information than the distribution histogram shown before (Figure 9). Indeed, this curve shows the cumulative frequency. Calculating double derivative helps to find where the slope is the strongest so that we have an idea of where is the elbow.

After having calculated the different thresholds, we have new features which are the number of long fixations during the initial reading, when returning to the paragraph and while reading the question title and the answer area.

For now the positions of the fixations were not used. It is however very important to find features which include the fixations' positions. Indeed, knowing where the participant looked at, has to be meaningful and useful to deduct if the participant will answer correctly. A first idea was to perform a clustering of the fixations using their positions X and Y and their duration. The first step would be to find some clusters using the fixations of all participants to find zones of interest inside an AOI. Indeed, performing clustering for each participant would be useless as everyone has different gazing behaviors. After having collected those clusters of fixations, the features could be the number of clusters fixated, which ones, how many fixations and what total time in each of the fixated clusters while a participant answers a question or the five questions of a scenario. Hierarchical clustering and the K-Means algorithm were used to find clusters. For hierarchical clustering, the Hartigan's rule was used to find the right number of clusters and the function in R of K-Means provide a way to find the optimal number of clusters (Mirkin, 2011). The second idea was that if all the fixations were selected for the clustering analysis, there would be too many. Instead, the clustering analysis was performed on long fixations.

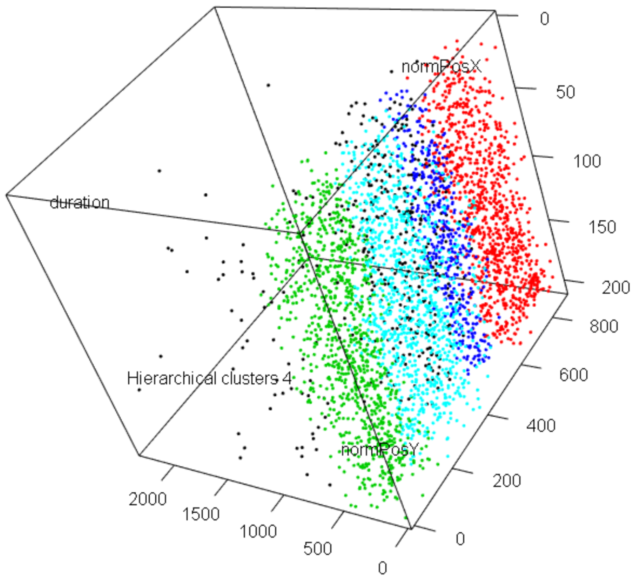


Figure 10 - Example of clusters found with hierarchical clustering for the paragraph AOI

Nonetheless, the clustering analysis was not conclusive because long fixations were still quite spread on the AOIs and the clusters found were just rectangles dividing the AOI. There was no cluster representing a specific reading behavior (Figure 10).

In reaction to this disappointing results, another way to take advantage of the fixations' positions was found. Instead of finding clusters, a virtual grid was traced over the concerned AOI. The size of the grid depended of the AOI. The grids were made so that each line of the grid approximately match with a line of text and the columns were sized so that an average sized word could fit inside. After that, the algorithm just counts for each question or scenario, how many fixations fell into each box of the grid. Then, the features calculated from this grid are the maximum value contained in the grid, the mean of all the values and the positions X and Y of the box where there was the maximum of fixations. The maximum value will represent whether the participant really focused on a box of the grid. The mean is a way to know if the participant fixated a lot the AOI and if he globally explored the all AOI. The position X and Y of the box with the maximum value will allow to see where the participant did look. The grid shape and size and which features should be extracted from it was decided when performing the machine learning algorithms.

3.7 Generation of synthetic data

For both regression and classification methods, there was a need to generate synthetic data. However, these needs were different for regression and classification. For the regression, the goal was to predict the percentage of success on a scenario. Therefore, as the fixation data was used for the first case of ANOVA, the number of observations was limited. In that case, the generation of synthetic data was used to add more observations to the existing ones.

The R package synthpop (Nowok, Raab & Dibben, 2015) enables to mimic the original data and generate new synthetic data which keeps the relationships between variables found in the original data set intact but does not contain any disclosive records. The simple synthesis was performed for only one data set. In that case, the synthesizer fits the data to the assumed distribution and calculates estimates of its parameters. More precisely, the distributions of the variables or columns that need to be synthesized are estimated thanks to the distribution of the original data to which is added the synthesized data when being processed.

For the binary classification, the goal was to predict whether the participant will answer correctly or incorrectly a question. Naturally, the fixations information came from the corresponding question so more observations were available in that case. However, a majority of the questions were answered correctly making the learning harder for the classification algorithms with such unbalanced data.

Therefore, the ROSE package was used in R (Lunardon, Menardi & Torelli, 2014). This package enables to deal with binary classification problems when one of the class is rare. This is exactly what would solve the problem as there is a minority of observations with an incorrect answer. ROSE generates artificial balanced samples according to a smoothed bootstrap method. The method used here is similar than the one used for synthpop except that the distributions are now from different classes and ROSE adds randomness to the process. For example, if ROSE generates an observation, the process will be the following:

- Select a class depending on the needed class (in our case, the incorrect class),
- Select randomly an original observation of the incorrect class with uniform probability,
- Sample a new observation from a probability distribution centered at the chosen original distribution and a covariance matrix H.

The probability distribution and covariance matrix come from the kernel density estimate from the linear mapping from the variables to the class. By repeating this process several times, a balanced data set can be reached with half of the observation being correct and the other half incorrect.

For both regression and classification, 70% of the data was used to train the models and 30% of the data was used for testing.

Chapter 4: Results

This Chapter details the findings of the different methods of analysis used to uncover the link between eye movements and textual comprehension. Those different methods of analysis are statistical tests studying the significance of different factors, analysis of visualizations showing eye fixations and their durations on the paragraph and finally, machine learning analysis with the goal to predict if the participant will answer correctly the questions proving that he understood the text.

4.1 Statistical tests

The statistical test performed is the ANOVA. As first mentioned in the Methods chapter, there are three cases of tests. The dependent variables used for those tests are the sum of the fixations' durations and the number of fixations for a specified time lapse (the time of an entire scenario or a question) which occurred in an AOI.

For the first case, in addition to the difficulty of the scenario, a few other factors were studied in order to analyze correlation effects. The age of the participant variable presented in the Methods chapter was included. The score of the participant when answering the questions of a scenario was included. This score was divided in two levels: first, whether the participant answered less than 4 questions correctly or if he had 4 or 5 correct answers. There were tests performed for the fixations which occurred in the paragraph AOI during the initial reading, the question title AOI and the questions and answer areas AOIs.

Concerning the second case, the two independent variables, difficulty of scenario and type of question were studied. Moreover, the score variable was included in this analysis with two levels whether the participant answered the question correctly or not. The age variable was not included as it was not found significant in the first case. There were tests performed for the fixations which occurred in the paragraph AOI during the questions, the question title AOI and answer area AOI. The questions area AOI was not studied as it was found redundant with the answer area AOI as explained later on.

For the third case, the difficulty of scenario and type of questions were studied again. In addition, a variable indicating whether an individual answer had been chosen or not was included in this analysis. This ANOVA was a single test as the only concerned AOIs were the individual answer AOIs (A, B, C, D or E).

The following table summarizes the effects including interaction effects found significant with all the ANOVA performed.

Table 1 - Significant effects found with ANOVA

Case	Variable(s)	AOI	Dependent variables
1	Difficulty of scenario	Questions area	Count (p-value = 0.02)
1	Difficulty of scenario	Answer area	Count (p-value = 0.04)
1	Difficulty of scenario & age	Paragraph (initial reading)	Count and duration (p-value = 0.01)
1	Difficulty of scenario & score	Paragraph (initial reading)	Count and duration (p-value = 0.04)
1	Score	Question title AOI	Count (p-value = 0.02)
2	Type of question	Question title AOI	Count (p-value = 6e-5) and duration (p-value = 0.03)
2	Difficulty of scenario	Question title AOI	Count (p-value = 0.005) and duration (p-value = 0.03)
2	Difficulty of scenario	Paragraph (returns to paragraph)	Count (p-value = 0.02)
2	Type of question	Paragraph (returns to paragraph)	Count (p-value = 0.002) and duration (p-value = 0.02)
2	Difficulty of scenario	Answer area AOI	Count (p-value = 2e-5) and duration (p-value = 5e-4)
2	Type of question	Answer area AOI	Count and duration (p-value = 1e-8)
2	Difficulty of scenario & score	Answer area AOI	Duration (p-value = 0.04)
2	Difficulty of scenario & question	Answer area AOI	Count and duration (p-value = 1e-6)
3	Type of question	Individual answers AOIs	Count and duration (p-value = 1e-16)
3	Difficulty of scenario & question	Individual answers AOIs	Count and duration (p-value = e-11)
3	Difficulty of scenario & chosen	Individual answers AOIs	Count (p-value = 0.03)

In order to analyze these significant effects, a post hoc analysis was done using the Tukey test and some plots showing the means of each level of the variables and the error intervals were drawn.

For the first case ANOVA, the difficulty of the scenario was found significant for the questions AOI and the answers AOI. As a reminder, the questions area AOI is uniting the answer AOI and the question title AOI. These tests allowed to find out that the questions area AOI is redundant as the same effect was found for both questions and answers area AOI. Therefore, the questions area AOI will no longer be used for further tests.

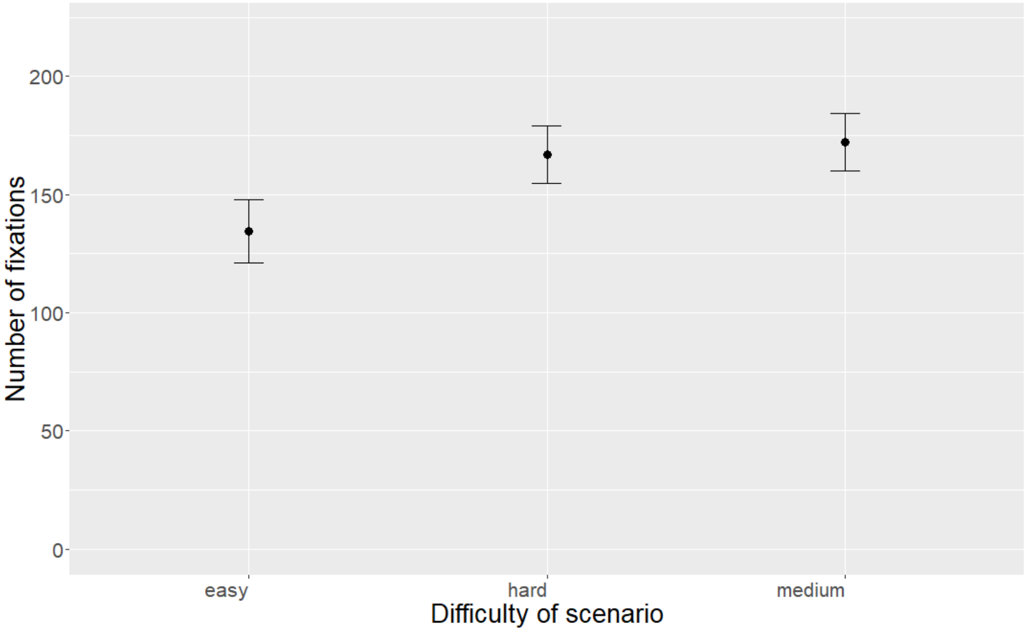


Figure 11- Means of the number of fixations for each scenario for the Answers AOI

The Tukey test showed a significant difference between the easy scenario and the intermediate and hard ones. Indeed, the medium and hard scenarios had a similar number of fixations which was higher than for the easiest scenario (Figure 11).

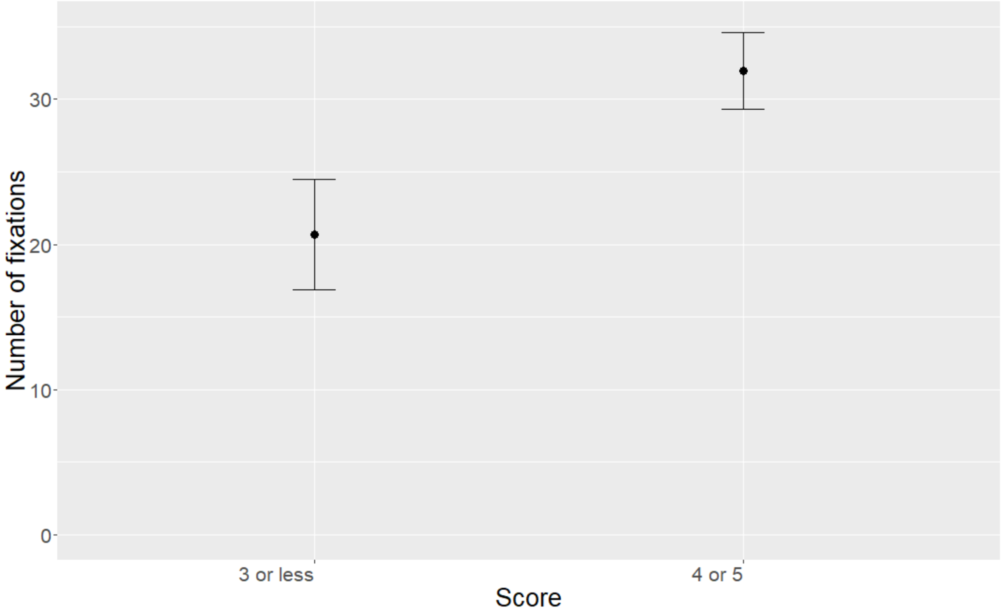


Figure 12 - Means of the number of fixations for different scores for the question title AOI

Concerning reading the question title, for the fixations' duration, nothing was significant. However, for the number of fixations, the participant's score was significant (Figure 12). This finding suggests that participants with a better score read the question title (for all scenarios) more than participants who had more incorrect answers.

The ANOVA showed that the interaction between the age of the participant and the complexity of the scenario was the most significant. The interaction between the complexity of the scenario and the score was also significant. Those results were found for both dependent variables. All participants had approximately the same number of fixations with the same overall duration for the easiest scenario, however for the medium and hard scenario, the participants with lower scores spent a lot more fixations and time on the initial reading.

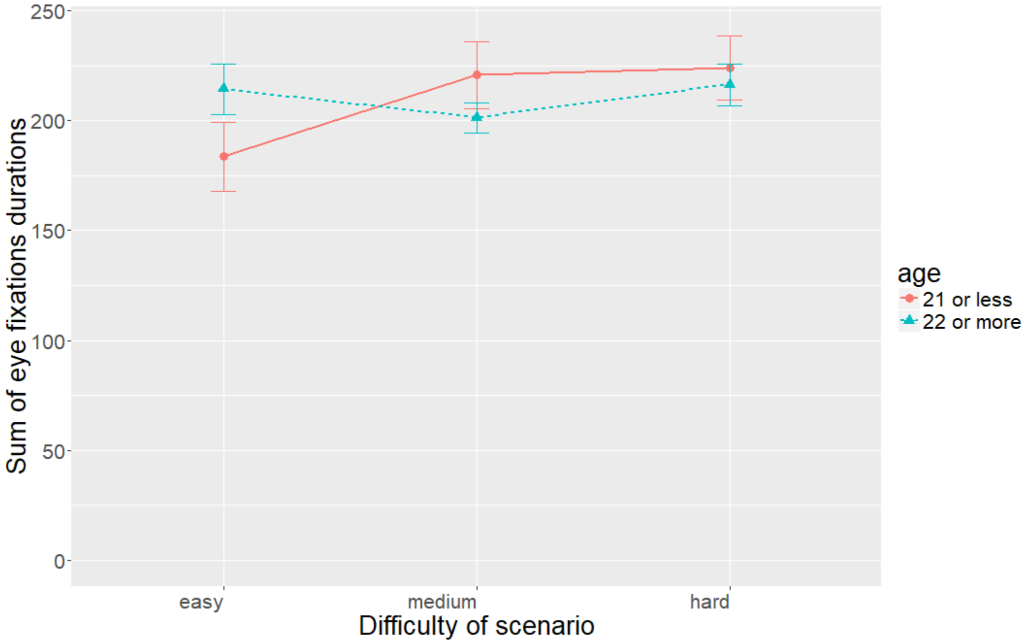


Figure 13 - Interaction plot between age and scenario for the initial reading

For the moderate and hard scenarios, youngest participants spent more time reading than older participants (Figure 13). For the easy scenario, younger participants were much faster. One explanation could be that the easy scenario was blood pressure and the patient was an 18 years old student mentioning his life habits. This could mean that younger reader maybe felt more concerned about this scenario, so they were more efficient when reading because more interested.

For the reading of the paragraph while answering the question, nothing was significant. Indeed, it is probable that the number of fixations and their durations follow more complex patterns. Going further into the details, for questions and answers, might be more useful for this case.

For the second case of ANOVA, the correctness or score on one question was often found significant. The effect was often involving less fixations and less long for questions answered correctly. However, because 78% of the questions were answered correctly by the participants, I chose to ignore these effects.

The difficulty of the scenario was significant for both dependent variables in the question title AOI. The Tukey test showed that the question title is fixated more for questions about the hardest scenario compare to the two other ones.

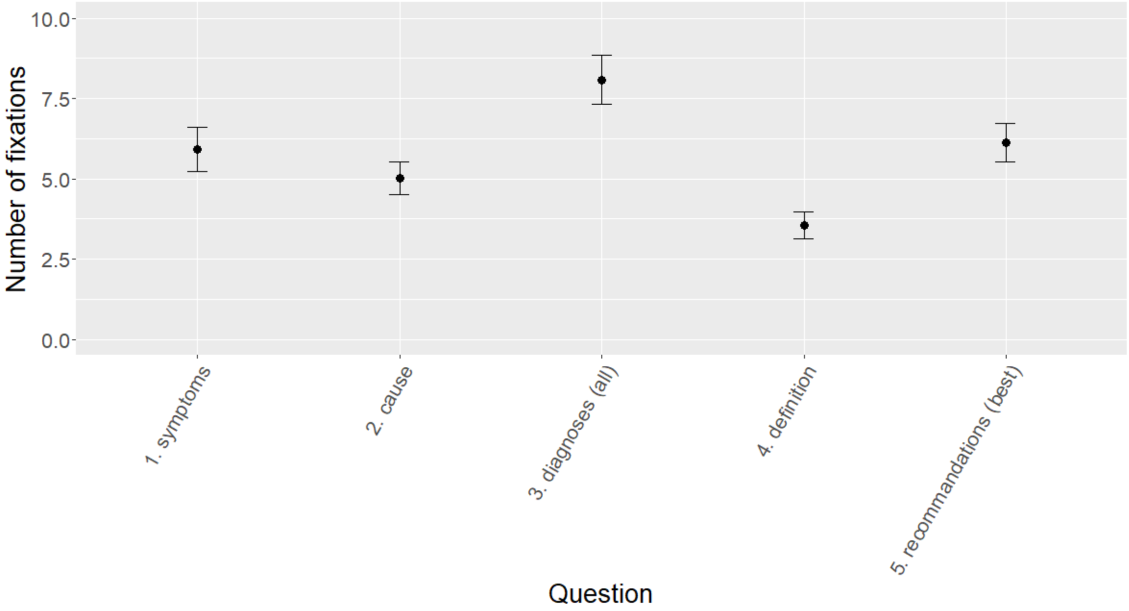


Figure 14 - Means of the number of fixations for different types of questions of the question title AOI

The type of question was also significant for the fixations' duration and for the number of fixations. Question titles about all the possible diagnoses were fixated the most and for the longest time while the question titles about the definition were fixated the least and for the shortest time (Figure 14).

Concerning the paragraph AOI corresponding to the returns to paragraph while answering, the difficulty of the scenario was again found significant but only for the number of fixations. There were a lot more returns for the hard scenario than for the easy and medium which had a similar number of returns. The type of question was again very significant for both dependent variables

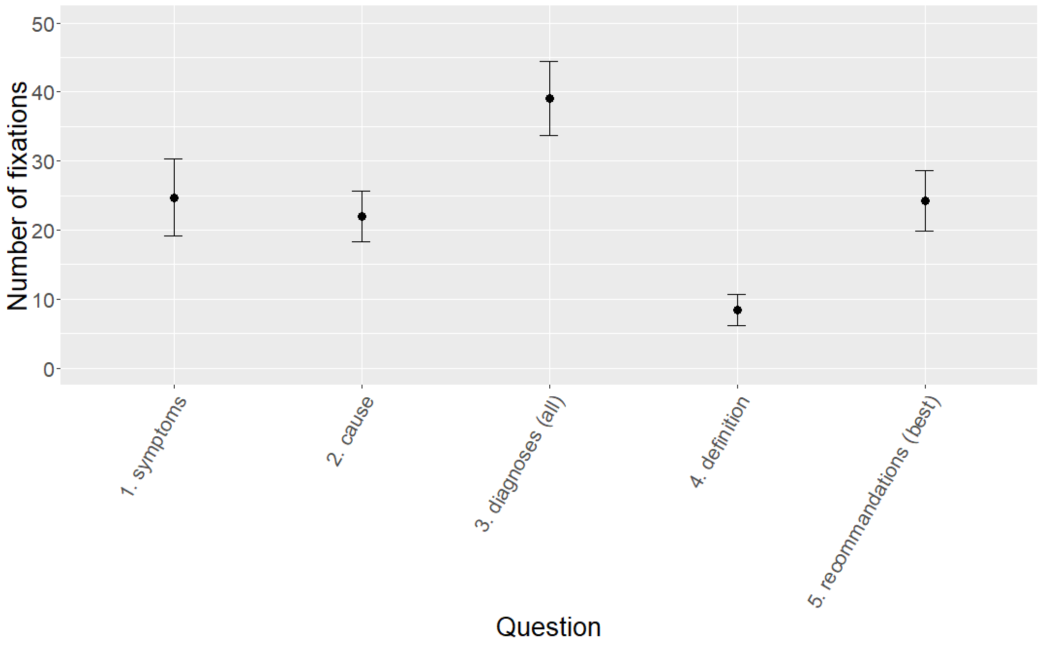


Figure 15 - Means of number of fixations for each type of question for the paragraph AOI

The question asking to select all possible diagnoses needed more returns than the others (Figure 15). This may be due to the fact that several answers are possible, therefore, the participants did not want to forget anything and they had to check the paragraph more. However, the definition of a term might already be known by the participant so he does really need to come back to the paragraph to check.

For the answer section AOI, there were again a lot of significant factors. The difficulty of scenario was significant again. This time, the hardest and the intermediate scenario had more fixations than the easiest one.

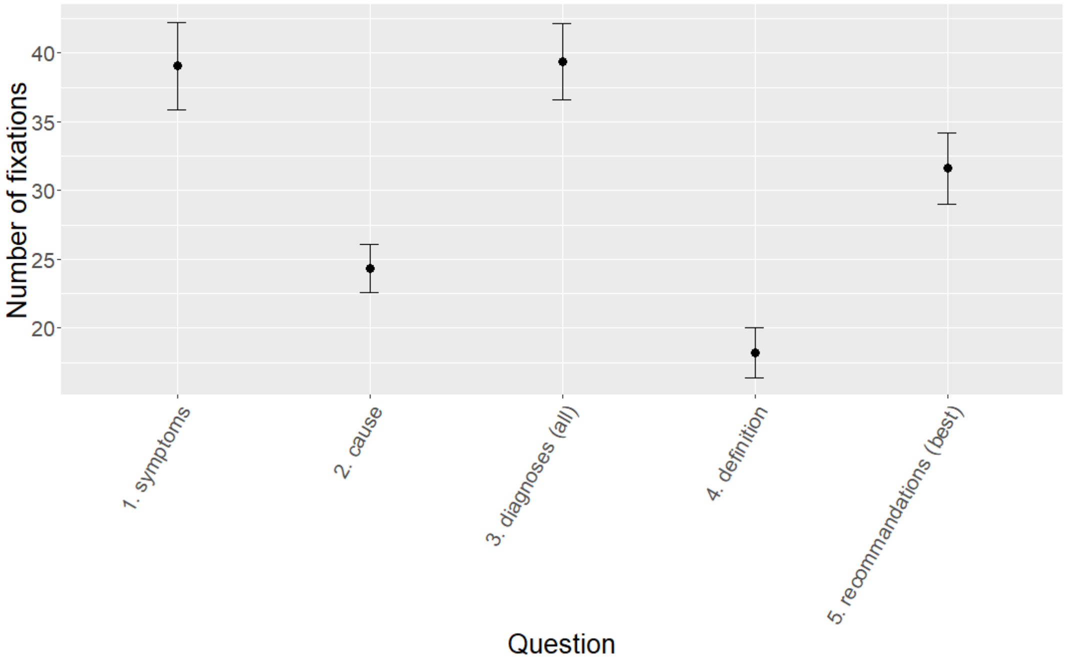


Figure 16 - Means of number of fixations for each type of question for the answer section AOI

The type of question is significant for both measure variables. The questions with fewest fixations in the answer sections are the questions about the cause and the definition (Figure 16). The question about the symptoms and the diagnoses requires a lot of fixations.

There was an interaction effect only for the fixations' durations between the correctness and the complexity of the scenario.

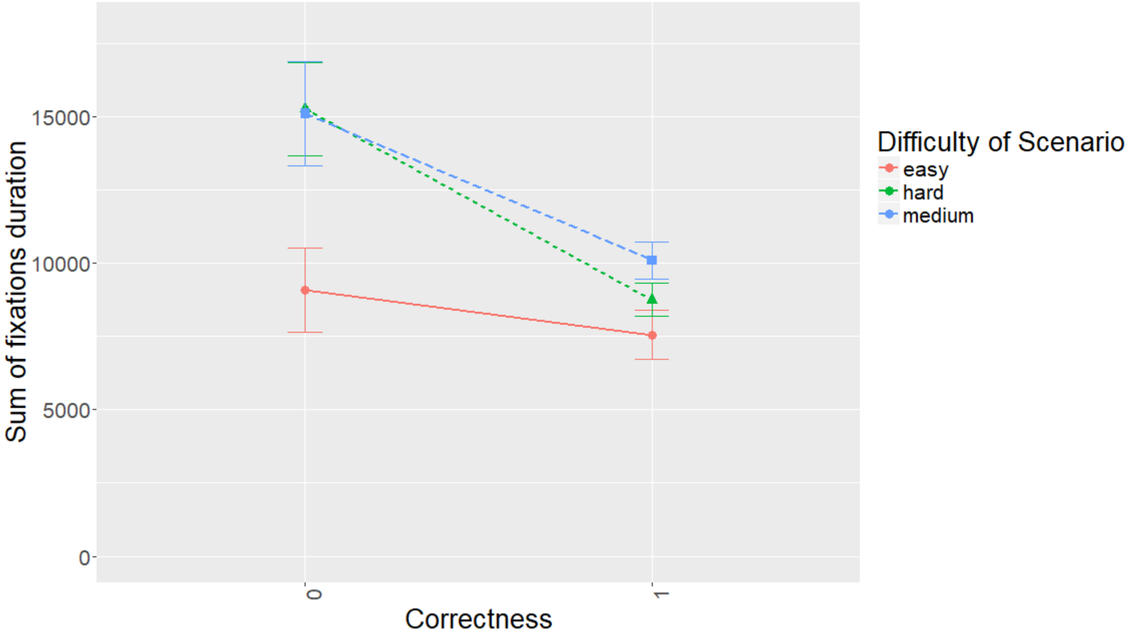


Figure 17 - Interaction plot between the correctness and the scenario for the answer section

Even though, the results cannot be fully trusted because of the biased data, some insights can be found in this plot. Indeed, for the medium and hard scenarios incorrect participants had longer fixations than with the easiest scenario (Figure 17). It is possible that staring at the answers could be a proof of confusion. When the participant knows the answer, he would have shorter fixations. This is related to the complexity of the scenario: harder scenarios may go with more confusion.

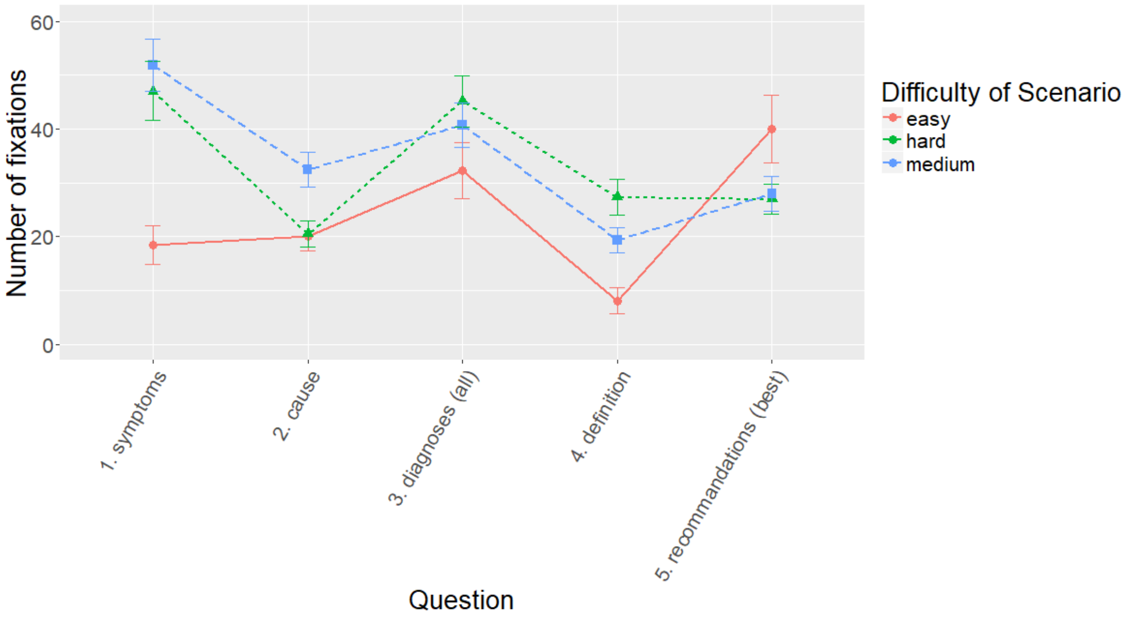


Figure 18 - Interaction plot between the type of question and the scenario for the answer section

There was also a very significant interaction effect between the type of question and the scenario. This interaction plot might seem very complex, but it is very useful to compare the questions' complexity, which can be considered proportional with the fixations' duration and number of fixations. Therefore, for the questions about the diagnoses and the definition, the order of complexity of the scenarios is respected. However, for some other questions it is not the case. For example, it seems that the recommendations question of the easiest scenario required more fixations in the answer section than for harder scenarios (Figure 18).

For the third case, the most significant variable was again the type of question. Concerning which questions were fixated the most or the least, the results are the same as for the answer section. This is logical, as the answers aggregated represent the answer section. Moreover, the type of question concerns all five answers of a question.

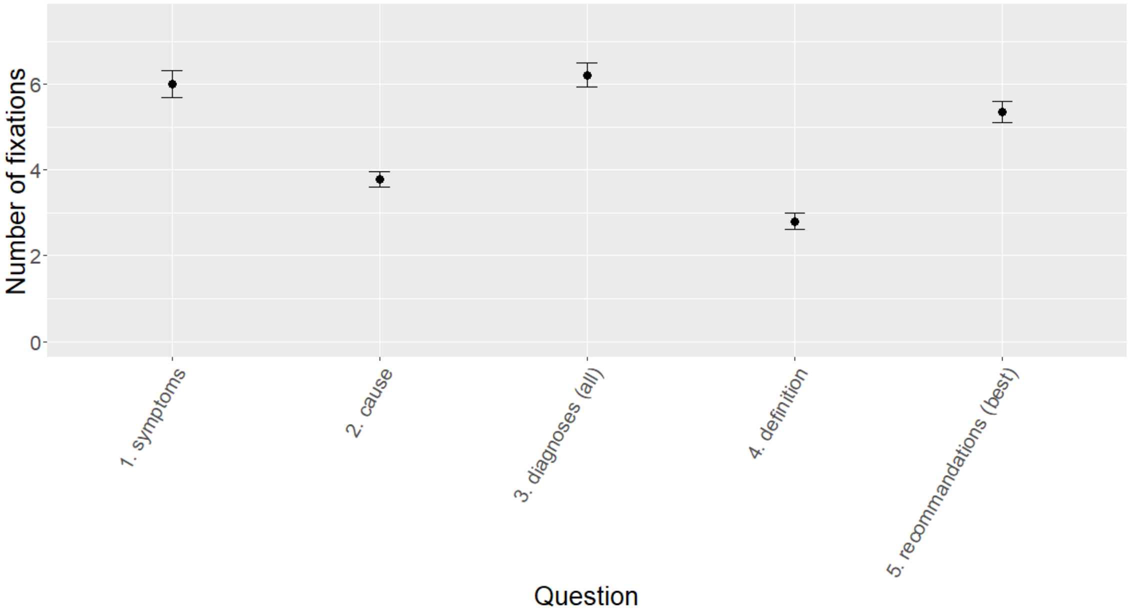


Figure 19 - Means of fixations' duration for each type of question for the answers AOI

The plot is indeed very similar to the previous one. One difference is that the error bars are much smaller (Figure 19). The p-value is much lower than before meaning that the effect stands out much more from the natural variation. The two binary variables indicating whether the answer was chosen or correct were not found significant. The interaction between the scenario and the type of question was again found significant and when plotting the interaction effect, it is the same as in the answer section in the second case of ANOVAs. This confirms the relevance of the results.

Another interaction effect was found significant for the fixations' duration: the interaction between the chosen variable and the scenario.

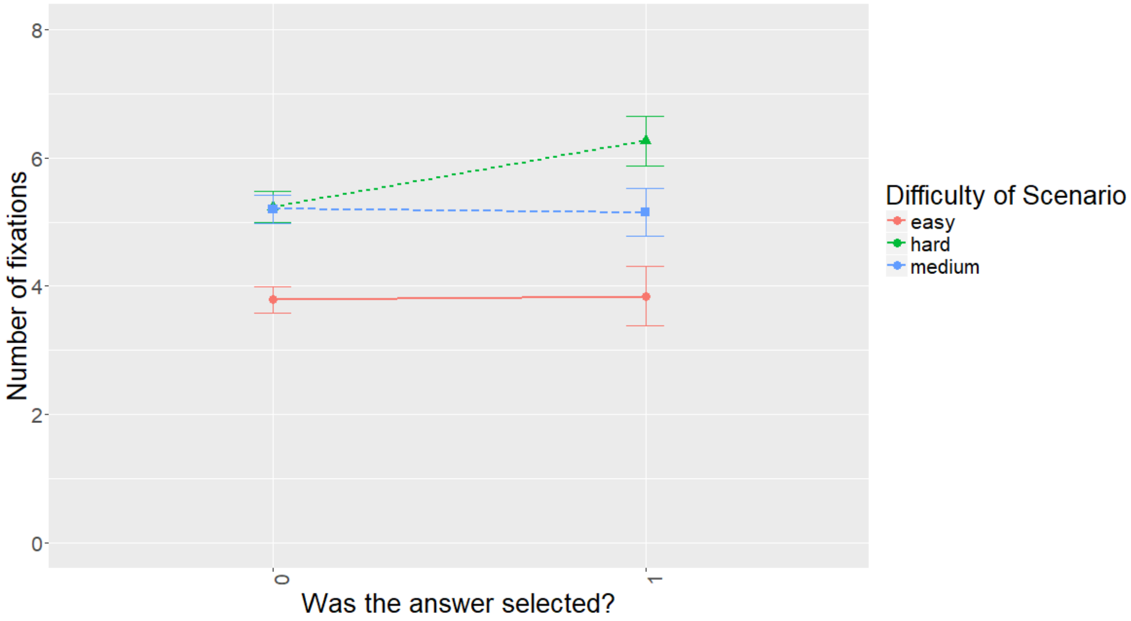
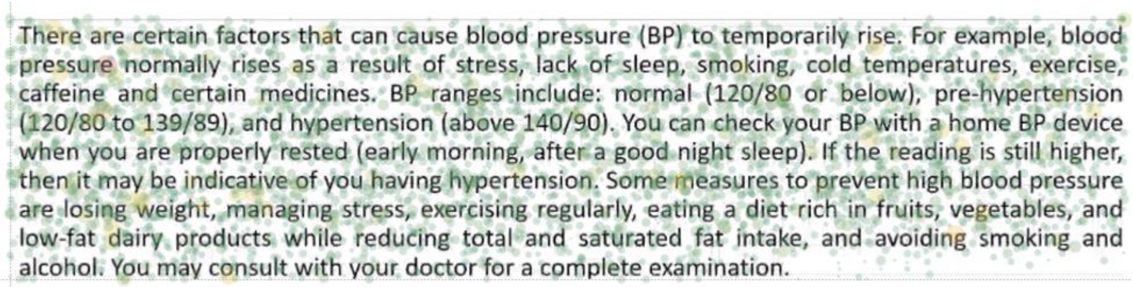


Figure 20 - Interaction plot between the scenario and the “chosen” variable

This plot shows that there was no difference in number of fixations whether the participants had selected the answer or not for the easy and medium scenarios. However, for the hard scenarios, the selected answer had more fixations (Figure 20). Therefore, it is possible that with a harder scenario, selecting an answer requires looking at it more.

4.2 Visualizations

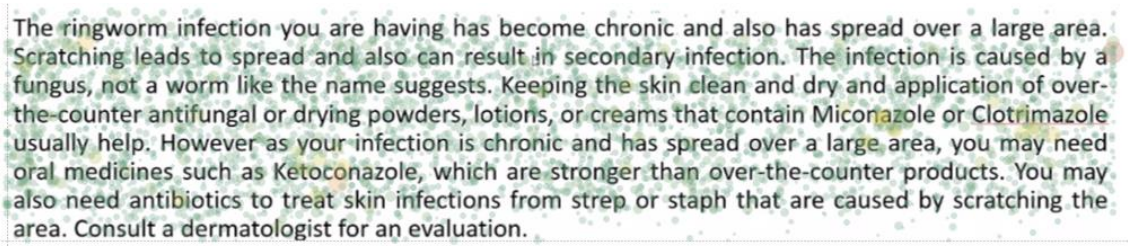
Another good way to analyze gaze data is to use visualizations, especially gaze plot or heat maps. In the first set of visualizations, the initial reading for the paragraphs of each scenario will be studied using a gaze plot combined with a heat map. Indeed, each point will represent a fixation and the duration will be represented by both the size of those circles and their color. The color will vary through a temperature scale going from green to red. Each heat map contains the data aggregated of all 25 participants which gives the general behavior while reading.



There are certain factors that can cause blood pressure (BP) to temporarily rise. For example, blood pressure normally rises as a result of stress, lack of sleep, smoking, cold temperatures, exercise, caffeine and certain medicines. BP ranges include: normal (120/80 or below), pre-hypertension (120/80 to 139/89), and hypertension (above 140/90). You can check your BP with a home BP device when you are properly rested (early morning, after a good night sleep). If the reading is still higher, then it may be indicative of you having hypertension. Some measures to prevent high blood pressure are losing weight, managing stress, exercising regularly, eating a diet rich in fruits, vegetables, and low-fat dairy products while reducing total and saturated fat intake, and avoiding smoking and alcohol. You may consult with your doctor for a complete examination.

Figure 21 - Gaze plot / heatmap of fixations during initial reading of the first scenario

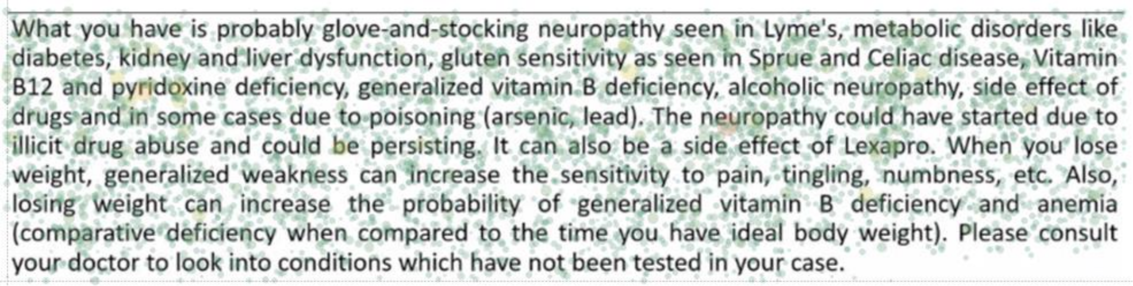
For the first scenario, the maximum duration is approximately 2 seconds. The integrality of the paragraph was hit by fixations with some variations that I will intend to explain (Figure 21). There are less fixations approaching the end of the paragraph which could mean a loss of interest or just tiredness from the participant. It is hard to conclude anything concerning reading patterns except witnessing visually that some terms seem to have more fixations than others and longer fixations such as maybe “pressure” and “hypertension”.



The ringworm infection you are having has become chronic and also has spread over a large area. Scratching leads to spread and also can result in secondary infection. The infection is caused by a fungus, not a worm like the name suggests. Keeping the skin clean and dry and application of over-the-counter antifungal or drying powders, lotions, or creams that contain Miconazole or Clotrimazole usually help. However as your infection is chronic and has spread over a large area, you may need oral medicines such as Ketoconazole, which are stronger than over-the-counter products. You may also need antibiotics to treat skin infections from strep or staph that are caused by scratching the area. Consult a dermatologist for an evaluation.

Figure 22 - Gaze plot / heatmap of fixations during initial reading of the second scenario

For the second scenario, maximum's fixation's duration, was almost 3 seconds which is more than for the first scenario. There are also less fixations at the end of the paragraph. It is also interesting to notice some warmer areas (more red) around the terms “Miconazole”, “Clotrimazole”, “Ketoconazole”, “ringworm” and “scratching”. The first three terms are scientific compounds which are complicated terms and difficult to pronounce and the last two are very important terms of the paragraph as the paragraph is about the ringworm infection and one of the main symptom of this health condition is scratching (Figure 22).



What you have is probably glove-and-stocking neuropathy seen in Lyme's, metabolic disorders like diabetes, kidney and liver dysfunction, gluten sensitivity as seen in Sprue and Celiac disease, Vitamin B12 and pyridoxine deficiency, generalized vitamin B deficiency, alcoholic neuropathy, side effect of drugs and in some cases due to poisoning (arsenic, lead). The neuropathy could have started due to illicit drug abuse and could be persisting. It can also be a side effect of Lexapro. When you lose weight, generalized weakness can increase the sensitivity to pain, tingling, numbness, etc. Also, losing weight can increase the probability of generalized vitamin B deficiency and anemia (comparative deficiency when compared to the time you have ideal body weight). Please consult your doctor to look into conditions which have not been tested in your case.

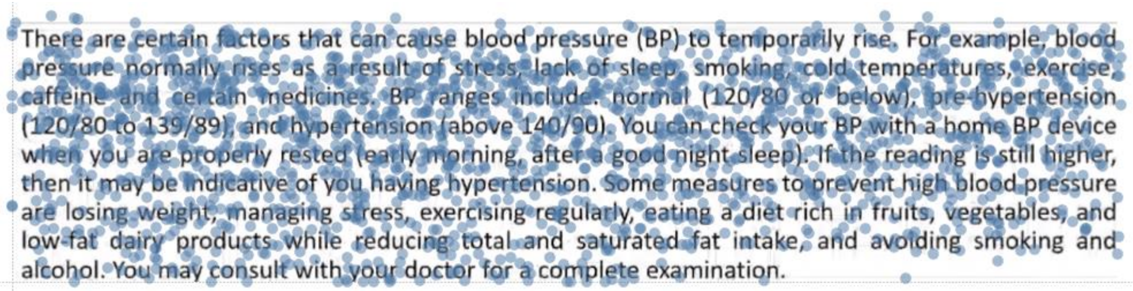
Figure 23 - Gaze plot / heatmap of fixations during initial reading of the third scenario

For the third scenario, the fixation with the longest duration was more than 3.6 seconds long. This scenario had a lot more fixations and it is possible to clearly see some agglomerates of fixations around many terms (neuropathy, Sprue, Celiac, Lexapro…). Moreover, there are also long fixations on some specific terms like “neuropathy” and “pyridoxine” (Figure 23). We can have similar conclusions as for the previous paragraph. Pyridoxine is a complex scientific term and harder to read. And neuropathy is the health condition mentioned in this scenario.

The second visualization idea is to visualize the returns to the paragraph but filter them for each question. Indeed, these gaze plots will allow us to visualize reading behaviors depending on the scenario and type of question for all participants. Only the gaze plots for the first scenario will be shown as the behaviors for different scenarios were quite similar except for an increasing number of fixations with the complexity of the scenarios.

- **Question 1: What medical symptom is the doctor addressing?**

  The correct answer to the question is “High blood pressure measurements”.



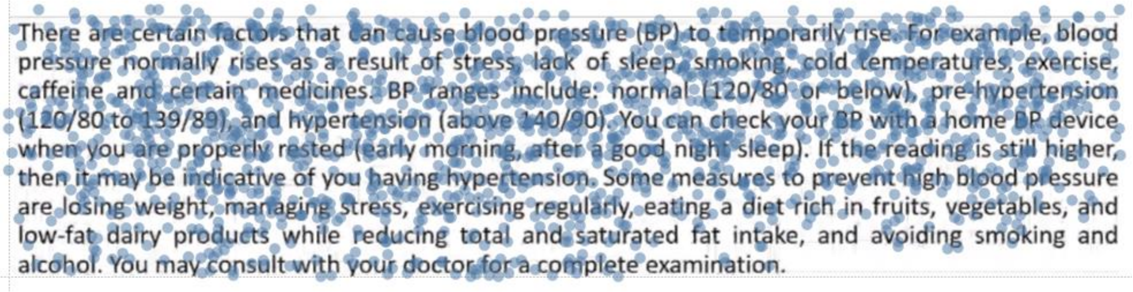
There are certain factors that can cause blood pressure (BP) to temporarily rise. For example, blood pressure normally rises as a result of stress, lack of sleep, smoking, cold temperatures, exercise, caffeine and certain medicines. BP ranges include: normal (120/80 or below), pre-hypertension (120/80 to 139/89), and hypertension (above 140/90). You can check your BP with a home BP device when you are properly rested (early morning, after a good night sleep). If the reading is still higher, then it may be indicative of you having hypertension. Some measures to prevent high blood pressure are losing weight, managing stress, exercising regularly, eating a diet rich in fruits, vegetables, and low-fat dairy products while reducing total and saturated fat intake, and avoiding smoking and alcohol. You may consult with your doctor for a complete examination.

Figure 24 - Gaze plot of the returns to the paragraph for question 1

We can see that the participants deliberately avoided the first line. They may have remembered from the initial reading that this sentence did not contain symptoms information. The fixations are then focused on the first part of the paragraph where the symptoms are actually described. This is interesting to see how the participants were pretty much aware of where this information was available in the text. There are still some returns going to the end of the paragraph (Figure 24).

- **Question 2: What does the doctor think is causing the problem?**

  The correct answer to that question is “stress and lack of sleep”.



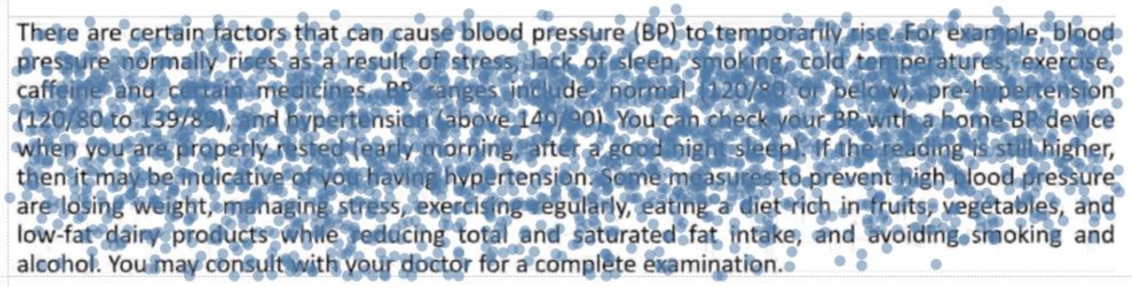
There are certain factors that can cause blood pressure (BP) to temporarily rise. For example, blood pressure normally rises as a result of stress, lack of sleep, smoking, cold temperatures, exercise, caffeine and certain medicines. BP ranges include: normal (120/80 or below), pre-hypertension (120/80 to 139/89), and hypertension (above 140/90). You can check your BP with a home BP device when you are properly rested (early morning, after a good night sleep). If the reading is still higher, then it may be indicative of you having hypertension. Some measures to prevent high blood pressure are losing weight, managing stress, exercising regularly, eating a diet rich in fruits, vegetables, and low-fat dairy products while reducing total and saturated fat intake, and avoiding smoking and alcohol. You may consult with your doctor for a complete examination.

Figure 25 - Gaze plot of the returns to the paragraph for question 2

There are less fixations than for the previous question which is probably due to the fact that the participants already found the information previously. The fixations are concentrated in the middle of the paragraph. They are not exactly where the right answer is located (second line). The fixations are more focused on the next lines which is surprising (Figure 25). This might mean that some participant did not find the information and kept reading.

- **Question 3: What are all the possible diagnoses listed by the doctor? (Select all correct answers)**

The correct answer to that question is “temperature” and “temporary high blood pressure”.



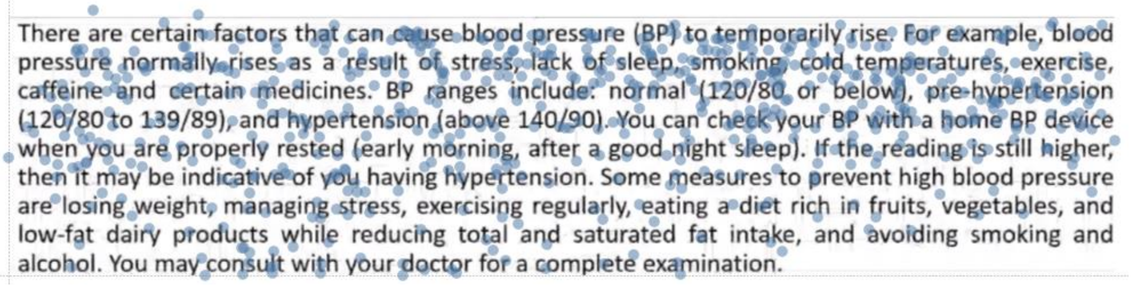
There are certain factors that can cause blood pressure (BP) to temporarily rise. For example, blood pressure normally rises as a result of stress, lack of sleep, smoking, cold temperatures, exercise, caffeine and certain medicines. BP ranges include: normal (120/80 or below), pre-hypertension (120/80 to 139/89), and hypertension (above 140/90). You can check your BP with a home BP device when you are properly rested (early morning, after a good night sleep). If the reading is still higher, then it may be indicative of you having hypertension. Some measures to prevent high blood pressure are losing weight, managing stress, exercising regularly, eating a diet rich in fruits, vegetables, and low-fat dairy products while reducing total and saturated fat intake, and avoiding smoking and alcohol. You may consult with your doctor for a complete examination.

Figure 26 - Gaze plot of the returns to the paragraph for question 3

This is the question which had the most fixations as found using ANOVA and it is visible on this visualization (Figure 26). For the three scenarios, for this question, the participants explored a lot more in the paragraph. This is possibly due to the fact that several answers are possible so the participants want to explore the integrality of the paragraph so they do not miss anything.

- **Question 4: What is “BP”? (definition question)**

The correct answer to this question is of course blood pressure.



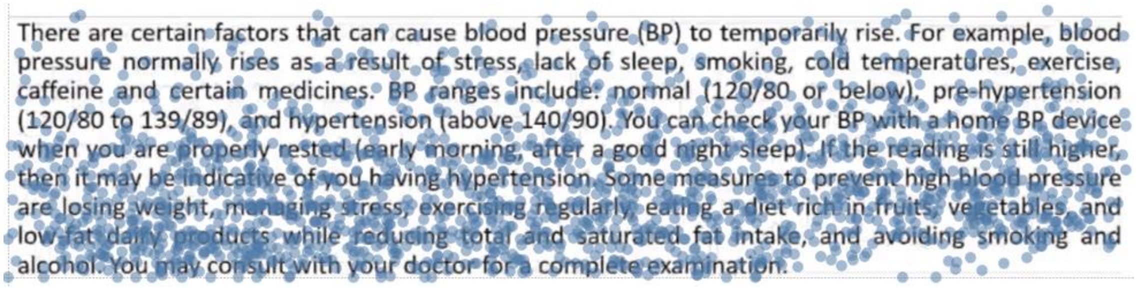
There are certain factors that can cause blood pressure (BP) to temporarily rise. For example, blood pressure normally rises as a result of stress, lack of sleep, smoking, cold temperatures, exercise, caffeine and certain medicines. BP ranges include: normal (120/80 or below), pre-hypertension (120/80 to 139/89), and hypertension (above 140/90). You can check your BP with a home BP device when you are properly rested (early morning, after a good night sleep). If the reading is still higher, then it may be indicative of you having hypertension. Some measures to prevent high blood pressure are losing weight, managing stress, exercising regularly, eating a diet rich in fruits, vegetables, and low-fat dairy products while reducing total and saturated fat intake, and avoiding smoking and alcohol. You may consult with your doctor for a complete examination.

Figure 27 - Gaze plot of the returns to the paragraph for question 4

This type of question was found with ANOVA to be the one with the least fixations. This is verified using this visualization (Figure 27). It is even more visible for the first scenario as the question is very easy and participants probably do not even need to come back to the paragraph to answer it. Moreover, the fixations focus on the first part of the paragraph where the term “BP” is.

- **Question 5: How does the doctor suggest the patient initially handles the problem? (Select 1 best answer)**

The correct answer to that question is “try another measurement in the morning”.



There are certain factors that can cause blood pressure (BP) to temporarily rise. For example, blood pressure normally rises as a result of stress, lack of sleep, smoking, cold temperatures, exercise, caffeine and certain medicines. BP ranges include: normal (120/80 or below), pre-hypertension (120/80 to 139/89), and hypertension (above 140/90). You can check your BP with a home BP device when you are properly rested (early morning, after a good night sleep). If the reading is still higher, then it may be indicative of you having hypertension. Some measures to prevent high blood pressure are losing weight, managing stress, exercising regularly, eating a diet rich in fruits, vegetables, and low fat dairy products while reducing total and saturated fat intake, and avoiding smoking and alcohol. You may consult with your doctor for a complete examination.

Figure 28 - Gaze plot of the returns to the paragraph for question 5

The most visible pattern on this visualization is that the fixations focus on the end of the paragraph where the recommendations actually are (Figure 28). This means that the participants either used their memory from the initial reading or just thought that the answer to the last question was probably at the end.

4.3 Machine learning

The machine learning analysis will have for goal to link the reading behaviors of the participants to the comprehension of the text. In other words, the objective is to correlate the fixations information and eye movements of the participants to the score of the participants on the questions asked during the experiment.

As explained in the methods part, our data is unbalanced and most of the participants did well when answering the questions. That is why for both regression and classification, R packages were used in order to make the data more balanced. Those packages were also used to simulate new data because as our experiment involved human subjects, the number of observations is limited.

For every machine learning problem, it is important to keep some data aside in order to test the performance of the algorithm on new data. That is why in our case one third of the data was kept for testing and the rest was used to train the algorithm.

The first machine learning method used was regression. Predict the score (percentage from 0 to 100%) of a participant on an entire scenario would allow to have aggregated data on each scenario with the initial reading and all other AOIs. Moreover, this analysis would be like a preliminary analysis to the classification problem as the objectives stay very similar. Indeed, the first goal is to predict the score of a participant on the 5 questions of a scenario using his fixations information in the current scenario and other informative variables (scenario complexity, type of question and participant's age). This allows us to have 75 observations. Using simulation of new data from the existing data, in total 500 observations were used for the analysis.

The different methods used for regression were linear regression, step wise regression, Lasso and Elastic net regressions, and regression tree and random forest.

Linear regression is a great way to have a first idea of which variables have an impact on the dependent variable. The step-wise regression allows to sort which variables are very meaningful from the ones which are not. Lasso and Elastic net are regression methods which improve and optimize the linear regression by introducing new parameters. Finally, the regression tree and random forest are tree based regression methods.

There was not a method significantly better than another. The metrics used were the RMSE (Root Mean Square Error) which is a good measure to calculate differences between real and fitted values and the BIC (Bayesian Information Criteria) which adds a penalty for adding too many parameters. The results were very homogeneous with RMSEs around 15 and BICs around 2000. The Elastic Net model did slightly better. The RMSE is almost satisfying because the score being between 0 and 100, 15 would be quite a low error. However, the BIC around 2000 let us conclude that our models are not good as a perfect model would have a BIC of 0.

The results were not satisfying. One of the reasons for that might be the limited number of observations even after simulation of data. Another reason is probably that there are 5 possible values that the score of a participant can take: 20%, 40%, 60%, 80% or 100% correct answers. And there are not decimal values which obviously suggest that a classification algorithm would be more efficient.

Therefore, for the second machine learning analysis, the goal was to do a binary classification on a single question, that is to say, predict if the participant will answer the question correctly or not using his fixations information from the current scenario and question and the same informative variables as for the regression analysis. With 5 questions for each scenario and 25 participants, this gives a total of 375 observations which is much more than for the regression analysis. However, approximately 80% of the questions were answered correctly. Therefore, ROSE was used to simulate new data and especially observations with an incorrect answer. With 375 observations, there was a reasonable amount of data so ROSE was used only to create more observations with incorrect answers which produced a total of 700 observations. This way, the amount of simulated data was reduced to its minimum and at least for the observations with a correct answer, the algorithms would work mostly with real data.

The data contained in the end more than 50 variables including the engineered features, the fixations information on different AOIs, the question and scenario, and even the age of the participant.

The different classification methods used were the logistic regression, decision tree, random forest, bagging tree method, Ada-boosted tree, Support Vector Machine and neural networks.

Logistic regression is the equivalent of linear regression except that the target variable is categorical and not continuous, so its functioning is different but the theory is similar. This method is one of the easiest classification method and it can give a first idea of the variables importance and if the classification problem is feasible.

Tree based classification methods are really appropriate for the classification problem of this project. Indeed, tree based methods are perfect for data with lot of attributes with unknown importance. It also makes no assumption on the data being normal or correlated. Some of the variables used here are much correlated (long fixations and all fixations, counts and durations…).

A decision tree is used in classification problems in order to divide a large heterogeneous population into smaller sets which are purer in respect to a target variable. A decision tree is a hierarchical structure and bagging and boosting methods are based on decision trees. Bagging methods fit several big trees to bootstrap-resampled versions of the training data and then classify by majority votes. Random forest are similar to bagging methods with an additional layer of randomness. Finally, boosted trees introduce weights to the classification process.

Support vector machine is a method which finds the optimal hyperplane, that is to say the hyperplane that separate the training examples of different classes by the highest distance. Finally, neural networks inspired by the biological neural networks are composed of different layers of connected neurons which let a bit pass if a certain threshold is reached and gives an output.

The metrics used were the accuracy which is the number of correctly classified divided by the total number of examples. The Kappa value introduces randomness to the observed accuracy which makes it a more robust metric. The true positive rate and true negative rate measures the rate of positive or negative observations which have been correctly classified as such. The Area Under the Curve (AUC) is a metric which takes into account the true positive rate but also the false positive rate which are the observations classified as true while being false.

Table 2 - Results for different classification methods

Method	Accuracy	Kappa	True Positive Rate	True Negative Rate	AUC
Logistic regression	0.7	0.39	0.67	0.72	0.69
Decision tree	0.9	0.81	0.88	0.92	0.9
Random forest	0.96	0.91	0.96	0.95	0.96
Bagging	0.92	0.84	0.94	0.9	0.92
Boosting	0.95	0.89	0.97	0.93	0.95
SVM	0.78	0.57	0.78	0.79	0.78
Neural net	0.64	0.29	0.6	0.69	0.64

The no-information rate was of 0.53 meaning that if the classifier just chose the class with a majority of observations, the obtained accuracy would be 0.53. The method with the best results was the random forest with excellent measures for all metrics used (Table 2). In a general way, the results are very satisfying which proves that comprehension is highly linked to reading behaviors.

There is an opportunity while using the random forest algorithm to calculate variables' importance. This variable importance is also called the mean decrease in Gini which represents how each variable contributes to the purity of the nodes. An analysis of the variables' importance was done using the variables' characteristics.

The variables the most significant for this classification were the position X of the cell with the highest number of long fixations, and the mean of the long fixation counts of all cells from the position analysis of the question title and the type of question.

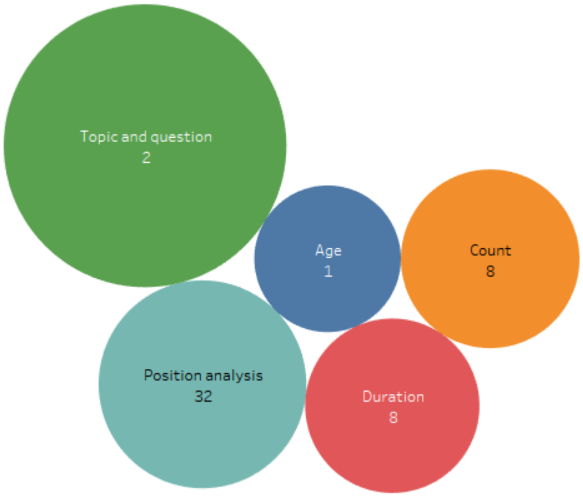


Figure 29 - Average importance for different types of variables

The variables were of different types: variables related to the gazing behaviors of participants, information related to participant, and information related to the content being read. Some variables were simple measures of counts and durations indicating how much and how long participants focused on an AOI. Moreover, the position analysis gives information on the location of those fixations within an AOI. The scenario and question were variables qualifying the text being read.

The scenario and question were important taking into account the number of variables concerned. The position analysis was also useful considering a total of 32 variables (Figure 29).

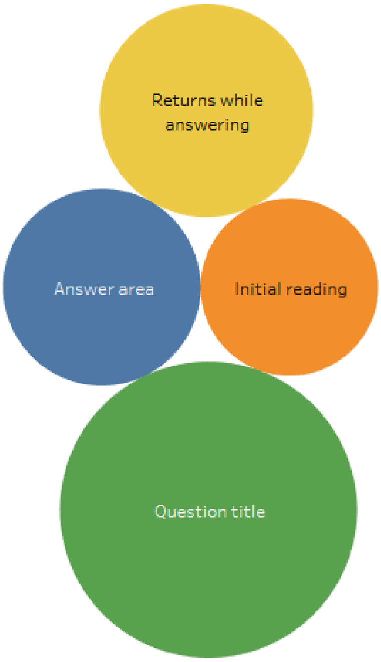


Figure 30 - Average importance for the different AOIs

The variables concerned different AOIs and the most useful for this analysis was the question title. This is surprising as the question title is also the smallest AOI of all. We can also see that the paragraph was also very useful as both the initial reading and the returns while answering was considered important (Figure 30).

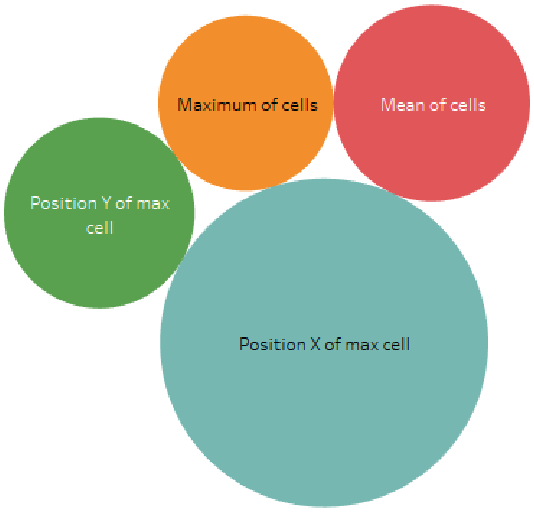


Figure 31 - Average importance for different position analysis variables

Within the position analysis, the position X was the most useful which makes sense as all the AOIS are much larger than longer in size (Figure 31).

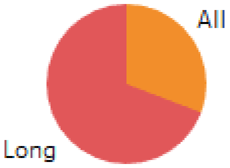


Figure 32 - Importance of variables for long and all fixations

The variables which concerned long fixations were considered more important during the classification process (figure 32).

This importance analysis supports that the feature engineering was conclusive as the long fixations and the position analysis allowed to determine some of the most useful variables for the classification.

Thanks to the package “inTrees” in R (Deng, 2014), the association rules which appeared the most and which were found to be true the most often were extracted from the trees of the random forest. However, the produced rules should not be considered as true as the decision trees use those rules combined with many other rules.

Some of those rules are:

- *If the question is about the symptoms, the diagnoses or the recommendations of the doctor, then the participant will answer incorrectly.*

With the statistical tests, those questions often had the highest number of fixations and longest fixation on almost all AOIs. This is another proof that those questions were considered more complex and therefore there were more chances of people answering incorrectly to those questions.

- *If the question is about the cause or the definition of a term then the participant will answer correctly.*

In the opposite way, the questions about the cause and the definition of a term were the ones which required the least fixations meaning that they were the easiest questions and that participants had less trouble answering them.

- *If the duration of long fixations returning to the paragraph while answering is inferior or equal to 8084 milliseconds, then the participant will answer incorrectly.*

This rule is quite straightforward as long fixation on the paragraph might mean more concentration so more chances to answer correctly.

- *If the position on the X axis of the cell with the highest number of long fixations in the question title AOI was approximately in the first part of the question then the participant answered incorrectly.*

This variable appeared in many rules as it is the most important variable in the classification model. The general behavior is that people focusing on the first half of the question have more chances to be incorrect whereas participants focusing on the second half have more chances of being correct.

- *If the maximum of a cell in the paragraph AOI (returns to paragraph while answering) is inferior to 1.8 fixations, then the participant will be incorrect.*

This means that participants not focusing on one part of the paragraph have more chance of being incorrect which goes with the rule found earlier about long returns to paragraph.

- *If participants had the mean of the cells in the answer area AOI inferior to 2.6 fixations then they were correct.*

In other words, if participant fixated a lot and quite globally the answer area they had more chances of being correct.

- *If the duration of long fixations on the answer area is inferior to 7482 seconds then the participant will be incorrect.*

Having long fixations on the answer area is also probably correlated with more concentration and understanding.

A lot of other similar rules were found. It can be concluded from this extraction of rules, that a random forest is a very complex classifier and that it is hard to understand the general behavior as many different rules are used combined with each other's. The random forest can be compared to a black box. But having those rules and the variables' importance gives some insight on what is correlated with text comprehension.

Chapter 5: Discussion

The quantitative analysis gave some interesting insights on the different factors studied. The first conclusion that can be drawn from all those tests, is that the two dependent variables, which are the sum of the fixations' durations or the fixations' count on an AOI, had different effects. For many tests, the same effects were found for both dependent variables. It is true that that there were only a few times when only one of the dependent variables was significantly affected. This shows that the fixations' durations and counts are linked. Indeed, the more fixations there will be, the more there is a chance that the sum of fixations' duration will be higher. That is not always the case: if the fixations are very long, then this causation effect is not always true. A set of numerous short fixations can have a lower sum of durations than only a few long fixations. This is why it is interesting to have both variables: the duration adds information to the number of fixations. However, the p-value of the fixations' duration is often lower than for the number of fixations. This is maybe a sign than differences in fixations' durations are not as easy to trigger as differences in number of fixations. This is probably due to the fact that long fixations are due to complex mechanisms affecting our memory and concentration, which are harder to detect at the AOI level, but those mechanisms should be easier to witness at the word level.

The statistical tests allowed to discover which variables had an effect on the dependent variables and then get a first and global idea of the reading behaviors.

It showed for the initial reading, that the participants did indeed allocate more attention to the paragraphs which were more complex. The quantitative analysis also showed that looking at the question title meant that the participants were more likely to get better results. This proves a known fact that reading the question is the first necessity in order to answer it correctly. However, the fixations’ durations did not matter in that case, maybe because staring too long at a question means that the participant does not understand the question and is confused.

Even though, it was not always significant, the complexity of the scenario was almost always recognizable when looking at the dependent variables (especially for the returns to paragraph and the answer area). This proves that a more complex paragraph with more scientific terms and a less known health conditions described, has to be read using longer fixations and more fixations. It is logical that a more complex scenario needs more time and reading to be understood. However, this could be a first conclusion that this higher number of fixations and higher time means that not only readers take more time to browse the text but also to think about it and access their memory to understand better. This is amplified by the fact that the paragraphs have similar sizes and share similar content. This goes also for looking at the other AOIs because with a harder scenario, the questions going with it have more chances to be harder to answer: choosing the right answer takes more time and fixations. Indeed, even though the questions were very similar for every scenario, the questions and associated answers were fixated more and longer with a harder scenario. Moreover, the number of fixations and fixations’ durations reflect the cognitive tasks associated with understanding the text, not only reading. Indeed, the questions and proposed answers were designed to be the same for each scenario which show that the scenario’s complexity is the reason for this increase of the dependent variables. This is probably due again to the access to the working memory to link the words to concepts and understand the described processes in order to try answering correctly the questions.

The age variable was never significant on its own. It was only significant within interaction effects which are hard to interpret. It was already suspected that the age was not going to be an important variable as all the participants were relatively young. In conclusion, in the context of this experiment, being a freshman or a PhD does not significantly alter reading efficiency or behaviors.

The second and third cases of ANOVAs allowed to uncover new variables which revealed to have very significant effects on the dependent variables. The question was the most significant factor with very low p-values for every single test for both dependent variables. The questions which had the higher number of fixations and longest durations was the questions asking for the diagnoses. This question was the only one which allowed the participant to select several answers. Because there were several possible answers, the participant had to browse the answers a lot more and also the paragraph with many more returns. The questions which also required more browsing of the different answers where the questions asking about the mentioned symptom and the “best” doctor’s recommendation for this health condition. Selecting the best answer might also be a reason for the higher values of dependent variables in this AOI. The question asking to define a term also gave interesting insight: because the participant might already know the answer, there were far less returns to the paragraph. For this question, there was also a lot less fixation in the answer area. Those measures give very interesting analysis on the questions’ complexity, even though as seen with the interaction between scenario and question, the question’s complexity also depends on the scenario for the reason explained previously.

The statistical test allowed to have a good idea on the AOI level of how the eye tracking measure varied with different variables. However, the analysis did not go to the word level and there was no information about the position of the fixations.

The gaze plots gave us more information about this aspect of the eye tracking data. First, those gaze plots allowed to verify many conclusions made with the ANOVAs: the increasing number of fixations and their duration with the scenario complexity, and the number of returns and their durations for different types of questions. For the initial reading, the participants went over the whole text even though they seemed to lose motivation and concentration when approaching the end of the text. It was interesting to see that some scientific and complex words were fixated for long periods of time. Those infrequent words required the long term memory of the participant to check if he could guess its meaning like found in the literature (Just & Carpenter, 1980). Some words which were not particularly infrequent but which were important in the context of the text and the questions, were also fixated longer meaning that the working memory was probably making sense out of the text. The gaze plots for each question showing the returns to the paragraph showed that the participants were aware of where the information was available. This is probably because of their initial reading that they were able to locate quite fast where the answers of the questions were.

The results of the machine learning analysis were interesting as the random forest was able to classify whether a participant would answer a question correctly with a 96% accuracy and excellent values for other metrics. The importance analysis allowed to see which AOIs were the most important. The question title AOI was the most important. Within this AOI, the variable indicating the position X of the grid cell with the most fixations was by far the most significant variable for the classification. It can be concluded from this, that the position of where the participant focused when reading the question is very meaningful. The results tended to show that focusing on the end of the questions means more chances to succeed in answering the question. This might mean that the participant who actually read with attention the question and reached the end and focused on the end understood the question better. However, this assumption is not completely accurate as the random forest is a complex algorithm combining several rules like this one. In a general way however, long fixations were very important for this classification problem. Nevertheless, it is hard to tell if long fixations are beneficial or not. Long fixation in the paragraph means incomprehension of a term or just confusion. But, in the answer area, an extracted rule showed that participants with longer fixations have more chances to answer correctly. Finally, for returns to paragraph, a rule showed that having long fixations back to the paragraph when answering was a good sign of success. Information such as the type of question and the scenario were also variables which were used by the random forest for the classification. For the question, it seems like questions requiring more fixations and more time as shown in the statistical analysis were questions for which the participant had a greater chance to answer incorrectly. It is interesting to see that the complexity of a question is indeed related to the number of fixations and their durations as they are harder to answer correctly. The random forest also gave insight on how a text has been well understood: if the participant explored the entire paragraph and the answer area (with long fixations preferably) without focusing too much on any part of the paragraph.

As mentioned in the Results chapter, the machine learning allowed to validate the engineered features. Indeed, the importance analysis extracted from the random forest model showed that the most important variable in this classification was an engineered feature from the position analysis. The position analysis was indeed very beneficial as many variables were used to create the different trees. The long fixations were considered more important to gain insight on how well participants will do. Using a classification algorithm like the random forest is very appropriate in order to analyze complex reading behaviors found in eye tracking data.

Chapter 6: Limitations & Future Research

6.1 Limitations

The first limitation of this research came from the eye tracking experiment which gave fixations with deviated vertical position. The deviations reduced the quality of the data set and even though the AOIs have been moved for each participant and that the fixations positions have been normalized if needed, there was still some approximation in the location of those fixations. Luckily, at the word level it is more important to have a correct horizontal position to know which word is being fixated. Then hopefully with the normalization, the fixation will be shown on top or at the bottom of the word, between the two lines.

Another limitation of this research is that the cases 2 and 3 of the ANOVAs had their independence assumed. Indeed, as the design of the experiment was made for the first case ANOVA, the independence could not be assured. However, most of the ANOVA results were then verified thanks to the visualizations.

The limitation of the machine learning was the unbalanced data with almost 80% of the questions answered correctly. If this data had been directly used, it would have been difficult to assess the performance of a classifier as the no information rate would have been 0.8 meaning that a classifier always saying that the participant will answer correctly would have an accuracy of 80%. This issue was resolved thanks to the simulation of new data from the existing data reaching a no information rate of 0.5. For the questions answered correctly, there is no problem as we had enough data but for the questions answered incorrectly, there were a lot of simulated data. The very high accuracy obtained could be explained by the negative class sharing similar characteristics due to the simulation. However, the classification results are still relevant as it can dissociate correct from incorrect. Thus, the real limitation here is that more real data would have allowed to avoid the simulation of data. Nevertheless, as all experiments involving human subjects and sensor data, it is expensive in time and often money to obtain large amounts of data.

6.2 Future research

Healthcare information is also often represented through graphs, plots and pictures. Performing a similar analysis on pictures and graphs instead of text would probably reveal interesting insight on what are the scanning patterns when fixating graphical elements. The cognitive processes would be different as a text is processed differently than a picture. An interesting question would be, is a picture more useful than a text in order to understand healthcare content? Or maybe a combination of both would be the perfect compromise?

Concerning applying machine learning to eye tracking data, there is a lot of future research which could be done in that field. Other than showing reading behaviors associated with a better textual comprehension of healthcare information, this classification problem could also be used in other contexts. Some applications of this classification would be to test the comprehension of students or people under some kind of training. Before submitting their answer, the computer could ask the students, who are being eye tracked, if they are sure of what they just answered and maybe guide them to areas of the screen they did not fixate enough, or just tell them to read again the question or the answers or the paragraph. Developing such a tool would need some kind of eye tracker and the context would be very similar to the experiment done in this thesis. The everyday-use of this tool would enable to collect a lot of data which would improve the classifier performances. When used multiple times, this tool could train users how to read more efficiently and even understand what they are reading better.

In the context of healthcare, such a tool could make sure the patient understands correctly his sickness and the associated treatment. Such a tool could also verify that a doctor understands accurately the treatment he is about to provide to a patient in order to avoid mistakes. In that case, the context would be slightly different as they would not be any questions to answer but only a text or content to go through which would probably need a different classifier.

Using such a classifier in other experiments involving eye tracking is useful to discover our eyes behaviors in other contexts: when taking decisions, when feeling different emotions or looking at different contents… Indeed, eye behaviors are very complex but so are the machine learning algorithms which can help understand better the underlying meaning of our eye movements like the random forest helped to show.

Chapter 7: Conclusion

Eye tracking is an excellent way to analyze reading behaviors and judge of the comprehension of human participants. The analysis performed for this thesis allowed to verify some known reading behaviors and uncover new ones. Healthcare information is hard to understand and it requires to be read with concentration and in a specific way in order to be understood correctly. Both working memory and long term memory are put to contribution to understand the meaning of the text or try to understand unknown words through context. The quantitative analysis allowed to verify this knowledge and gain insight on the behaviors for different types of questions and scenario of increasing complexity.

Applying machine learning to sensor data such as eye tracking data is something very new and only a few applications can be found in the literature. The research performed in this thesis is another proof that eye tracking data can be used to predict very accurately how humans make decisions and how well they understood a content. Even though, the random forest is a model hard to read, several rules were extracted which gave some guidelines of what makes a great reader.

References

- Bansback, N., Li, L. C., Lynd, L., & Bryan, S. (2014). Development and preliminary user testing of the DCIDA (Dynamic computer interactive decision application) for ‘nudging’patients towards high quality decisions. BMC medical informatics and decision making, 14(1), 62.
- Bell, T. (2001). Extensive reading: Speed and comprehension. The reading matrix, 1(1).
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. Journal of Economic Behavior & Organization, 81(1), 1-8.
- Deng, H., & Deng, M. H. (2014). Package ‘inTrees’.
- Eastin, M. S. (2001). Credibility assessments of online health information: The effects of source expertise and knowledge of content. Journal of Computer-Mediated Communication, 6(4), 0-0.
- Eghdam, A., Forsman, J., Falkenhav, M., Lind, M., & Koch, S. (2011). Combining usability testing with eye-tracking technology: evaluation of a visualization support for antibiotic use in intensive care. In MIE (pp. 945-949).
- Estey, A., Musseau, A., & Keehn, L. (1991). Comprehension levels of patients reading health information. Patient Education and Counseling, 18(2), 165-169.
- Fox, S. (2011). Health topics. Pew Internet & American Life Project. Washington, DC.
- Granka, L. A., Joachims, T., & Gay, G. (2004, July). Eye-tracking analysis of user behavior in WWW search. In Proceedings of the 27th annual international

ACM SIGIR conference on Research and development in information retrieval (pp. 478-479).

- Heer, J., Mackinlay, J., Stolte, C., & Agrawala, M. (2008). Graphical histories for visualization: Supporting analysis, communication, and evaluation. IEEE transactions on visualization and computer graphics, 14(6).
- Holmqvist, K., & Wartenberg, C. (2005). The role of local design factors for newspaper reading behaviour-an eye-tracking perspective. Lund University Cognitive Studies, 127, 1-21.
- Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009, September). Learning to predict where humans look. In Computer Vision, 2009 IEEE 12th international conference on (pp. 2106-2113). IEEE.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological review, 87(4), 329.
- Kim, H., Goryachev, S., Rosemblat, G., Browne, A., Keselman, A., & Zeng-Treitler, Q. (2007). Beyond surface characteristics: a new health text-specific readability measurement. In AMIA Annual Symposium Proceedings (Vol. 2007, p. 418). American Medical Informatics Association.
- Kules, B., & Xie, B. (2011). Older adults searching for health information in MedlinePlus–an exploratory study of faceted online search interfaces. Proceedings of the American Society for Information Science and Technology, 48(1), 1-10.
- Kunze, K., Utsumi, Y., Shiga, Y., Kise, K., & Bulling, A. (2013, September). I know what you are reading: recognition of document types using mobile eye

tracking. In Proceedings of the 2013 International Symposium on Wearable Computers (pp. 113-116). ACM.

- Leroy, G., Helmreich, S., Cowie, J. R., Miller, T., & Zheng, W. (2008). Evaluating online health information: Beyond readability formulas. In AMIA Annual Symposium Proceedings (Vol. 2008, p. 394). American Medical Informatics Association.
- Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. R Journal, 6(1).
- Mcinnes, N., & Haglund, B. J. (2011). Readability of online health information: implications for health literacy. Informatics for health and social care, 36(4), 173-189.
- Mirkin, B. (2011). Choosing the number of clusters. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 252-260.
- Nowok, B., Raab, G. M., & Dibben, C. (2015). synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software.
- Sakia, R. M. (1992). The Box-Cox transformation technique: a review. The statistician, 169-178.
- Sillence, E., Briggs, P., Harris, P. R., & Fishwick, L. (2007). How do patients evaluate and make use of online health information?. Social science & medicine, 64(9), 1853-1862.
- Sillence, E., Briggs, P., Harris, P., & Fishwick, L. (2006). Going online for health advice: changes in usage and trust practices over the last five years. Interacting with computers, 19(3), 397-406.

- Soederberg Miller, L. M., Gibson, T. N., Applegate, E. A., & de Dios, J. (2011). Mechanisms underlying comprehension of health information in adulthood: The roles of prior knowledge and working memory capacity. Journal of Health Psychology, 16(5), 794-806.
- Team, R. C. (2000). R language definition. Vienna, Austria: R foundation for statistical computing.