The Effects of Individual Differences and Linguistic Features on Reading Comprehension of Health-Related Texts

Michael Ratajczak

Doctoral Thesis Submitted in Fulfilment of the Requirements
for the Degree of Doctor of Philosophy in Linguistics

April 2020

Lancaster University

Department of Linguistics and English Language

Declaration

The work submitted in this thesis is my own and has not been submitted in substantially the same form towards the award of another degree or other qualifying work by myself or any other person. I confirm that acknowledgement has been made to assistance given and that all major sources have been appropriately referenced.

Name: Michael Ratajczak

Signature: M. Ratajczak

Date: 27/04/2020

Abstract

Background. Relatively little attention has been focused on whether or how the effects of reader characteristics, or of the linguistic properties of a text, predict reading comprehension of health-related information. In addition, there is little evidence for the utility of any of the writing guidelines promulgated by the National Health Service (NHS) in order to improve the comprehension of health information. Nonetheless, some previous research suggests that health-related texts could be adapted for different groups of users to optimise understanding. Thus, existing knowledge presents important limitations, and raises concerns with potentially far-reaching practical implications. To address these concerns, I investigated how variation in individual differences and in text features predicts the comprehension of health-related texts, examining how the effects of textual features may differ for different kinds of readers.

Method. The focus of this thesis is on Study 3, in which I investigated the predictors of tested comprehension, but I report preliminary studies where I examined the readability of a sample of health-related texts (Study 1), and the perceived comprehension of a sample of health-related texts (Study 2). In the primary study (Study 3), I used Bayesian mixed-effects models to analyse the influences that affect the accuracy of responses to questions probing the comprehension of a sample of health-related texts. I measured variation among 200 participants in their cognitive abilities, to capture the effects of individual differences, as well as variation in the linguistic features of texts, to capture the effects of text structure and content.

Results. I found that tested comprehension was less likely to be accurate among older participants. However, comprehension accuracy was greater given higher levels of education, health literacy, and English language proficiency levels. In addition, self-rated evaluations of perceived comprehension predicted comprehension, but only in the absence of other individual-differences-related predictors. Variation in text features, including readability estimates, did not predict comprehension accuracy, and there was no evidence for the modulation of the effects of individual differences by text features.

Discussion. Text features did not module the effects of individual differences to influence comprehension accuracy in any meaningful way. This suggests that adapting health-related texts to different groups of the population may be of limited practical value.

Implications. Individual differences really matter to comprehension. Thus, optimally, understanding of health-related texts amongst the end-users should be tested, and interventions to aid readers, such as those with relatively low health literacy levels, could be used to improve comprehension of health-texts. In the absence of sensitive measures of reader characteristics, and when testing of understanding is not possible, the use of end-user evaluations of health-related texts may serve as a useful proxy of tested comprehension. However, looking for text effects, and guidance focusing on text effects, seems less useful given the reported evidence. Consequently, the effectiveness of designing health-related texts with the consideration of NHS's text writing guidelines, is likely to be limited.

Acknowledgments

I thank my supervisors (Professor Judit Kormos, Dr. Rob Davies, and Dr. Megan Thomas) for being patient, having an open mind, supporting me academically and emotionally, and encouraging me to improve my work. I also thank Dr. Simon Taylor for teaching me statistics. I am happy and grateful that I got to this stage and that I have learned so much from all of you in the process. Unquestionably, this thesis would be of lower quality without your input.

This work was supported by the Economic and Social Research Council [award reference 1606173].

Contents

List of Tables

Table: **Page Number:**

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 7:

List of Figures

Figure: **Page Number:**

Chapter 1:

Chapter 4:

Chapter 5:

Chapter 6:

Chapter 7:

List of Abbreviations

Abbreviation:	**Definition:**
ANOVA	Analysis of Variance
BNC	British National Corpus
CI Model	Construction-Integration Model
CIs	Credible Intervals
DVC Model	Decoding, Vocabulary, and Reading Comprehension Model
EJC	Expected Judgement Change
ESL	English as a Second Language
Est.	Estimate
FRE	Flesch Reading Ease
H	Hypothesis
HLVA	Health Literacy Vocabulary Assessment
L1	First Language
L2	Second Language
L-95%	Lower 95%
LARRC	Language and Reading Research Consortium
LD	Lexical Decision
LKJ (prior)	Lewandowski-Kurowicka-Joe (prior)
LOO-CV	Leave-One-Out Cross-Validation
LOOIC	Leave-One-Out Information-Criterion
LQH	Lexical Quality Hypothesis
LSA	Latent Semantic Analysis
M	Mean
MCMC	Markov Chain Monte Carlo

Mdn	Median
NHS	National Health Service
NHST	Null Hypothesis Significance Testing
OR	Odds Ratio
ORF	Oral Reading Fluency
PPC	Posterior Predictive Check
PSIS	Pareto-Smoothed-Importance-Sampling
RDL2	Coh-Metrix L2 Readability Index
RQ	Research Question
SAHL-*E*	Short Assessment of Health Literacy-English
SD	Standard Deviation
SMOG	Simple Measure of Gobbledygook
S-TOFHLA	Short Test of Functional Health Literacy in Adults
SVR	Simple View of Reading
THLQ	Three Health Literacy Questions
U-95%	Upper 95%
UK	United Kingdom
US	United States
VIF	Variance Inflation Factor
WM	Working Memory

Introduction

Relatively little attention has been focused on whether or how the effects of cognitive abilities, and the effects of the linguistic properties of the text, influence the reading comprehension of health-related information. This limitation should be a concern to health service providers across the world because, in health settings, reading comprehension problems are associated with poor health status, more hospital admissions, and an increased risk of dying earlier (e.g., Baker et al., 2002; Bostock & Steptoe, 2012; Schillinger et al., 2002). Critically, it is estimated that 43% of working-age adults in England do not have literacy skills at a level which would allow them to understand and make use of health information (Rowlands et al., 2015). This is important as adults with low literacy skills tend to be less trusting, less informed, and on average suffer worse health than those with higher literacy levels (e.g., Bostock & Steptoe, 2012). However, given that adults vary in health literacy skills, we cannot assume that variation in textual properties influences everyone's comprehension in the same way. Thus, we need to investigate whether textual properties matter alongside individual differences.

In my PhD research, I examined the factors that are likely to predict the comprehension of printed health-related information in adults. Specifically, my research aimed to identify the factors that predict the variation in comprehension of printed health-related texts, in contexts where health-related information is presented primarily in written textual format without illustrations, and where adults are expected to read the text alone at the hospital or at primary care premises. My goal was to provide an answer to the question: How do adults with different characteristics understand printed health information? In answering this question, I aimed to furnish the basis for guidelines that can inform the production of health-related texts that are optimally comprehensible for adults with different individual profiles. Therefore, I investigated not only the effects of text characteristics on comprehension but also the effects of individual differences variation among adults in contributing to comprehension. In addition, I considered the possibility of the effects of individual differences being modulated by the effects of text features when reading health-related texts. Next, I briefly outline the structure of this thesis for the benefit of the reader.

Thesis Structure

This thesis consists of eight chapters. The first three chapters constitute the literature review. The aim of the literature review chapters is to illustrate the similarities and discrepancies between some of the theoretical accounts of reading comprehension and the findings of empirical research. To fulfil this aim, the first chapter includes a review of reading comprehension models, whereas the second chapter concentrates on the findings of empirical reading comprehension research in relation to variation in individual differences and text features which may predict reading comprehension. The third chapter is a continuation of the second chapter, with the focus further narrowed on plausible reader- and text-level predictors of comprehension in the context of health-related texts. In the fourth chapter, I describe the overall research design and rationale for the three studies included in this thesis, including the research questions and the research gap that this project aimed to fill. In the fifth chapter, I examine the readability of a sample of health-related texts (Study 1), in the sixth chapter I investigate perceived comprehension of a sample of health-related texts (Study 2), and in the seventh chapter I consider the comprehension of a sample of health-related texts (Study 3). The eighth chapter constitutes the overall discussion of the thesis, including the theoretical and practical implications of the evidence presented in the preceding chapters, an overall conclusion, and directions for further research.

Chapter 1: Literature Review of Reading Comprehension Models

In this chapter I aim to build the theoretical background required for the investigation of the effects of individual differences and the effects of text features on reading comprehension of health-related texts. First, I provide a brief overview of reading comprehension, including reading comprehension measures and models. Next, I contrast the different models of reading comprehension, with the aim of assimilating the differences into a comprehensive account of comprehension processes. Last, I conclude this chapter with a brief summary, specifying the theoretical framework of comprehension that this thesis follows.

1.1. Reading Comprehension: An Overview

Successful reading comprehension is essential for understanding texts. It is crucial in everyday life as it enables individuals to learn, academically and professionally, as well as to interact with others using social networking sites, emails, and text messages (e.g., Freed, Hamilton, & Long, 2017; Oakhill, Cain, & Elbro, 2014). Reading comprehension is a complex process that happens very quickly and involves many different cognitive processes and abilities, the effects of which interact with the effects of the features of texts read (Kendeou, van den Broek, Helder, & Karlsson 2014; Francis, Kulesz, & Benoit, 2018). Comprehension requires readers to combine their understanding of words and sentences, obtained from text, into a coherent whole (Oakhill et al., 2014). The success of the product of this integration is dependent on readers' ability to construct a mental model (Kendeou et al., 2014), which is a mental representation that is created from the information that a reader has read (Oakhill et al., 2014).

Historically, researchers have tended to focus on one of the following aspects of comprehension: component skills of readers; text features that influence comprehension; and the development of reading comprehension through life stages, which mainly refers to the acquisition of reading by children (Francis et al., 2018). According to Francis et al., these research strands can be classified into three main reading comprehension frameworks: the component skills framework; the text and discourse framework, and the developmental framework. These frameworks approach reading comprehension from different angles and it is rarely stated explicitly how these frameworks connect with each other. The component skills framework elaborates on the component cognitive skills that underlie comprehension (e.g., Gough & Tunmer, 1986). The text and discourse framework focuses on how variation in different text features influences comprehension (e.g., van Dijk & Kintsch, 1983; McNamara & Kintsch, 1996), and the developmental framework is primarily concerned with the developmental changes in reading skill in children and young adults (e.g., Cain, Oakhill, & Bryant, 2004; Garcia & Cain, 2014). However, reading comprehension is a complex mental process which is influenced by developmental factors and is a product of interactions between a reader, text, and the reading process (Francis et al., 2018). These dynamic interactions between developmental changes, individual differences and textual characteristics are often omitted from reading comprehension research, yet it is these interactions that are involved in the construction of a coherent mental representation of the text individuals read.

Reading comprehension is thought to be influenced by general cognitive abilities, lower-level processing, and higher-level processing that is more open to conscious introspection by the comprehender than lower-level processing (Perfetti, 2007; Grabe, 2014). The lower-level processes include fast and automatic word recognition, and lexico-syntactic processing. Lexico-syntactic processing refers to recognising parts of words and their morphology to build a syntactic structure (Grabe, 2014). The higher-level processes consist of comprehension monitoring, inference making, and prior knowledge. Critically, some higher-level processes, such as inference-making, can be reader-initiated if the reader’s standards of coherence are not met (van den Broek & Helder, 2017). Standards of coherence are the criteria that readers have for achieving adequate comprehension and coherence in a specific reading situation, reflecting the desired level of understanding (van den Broek, Bohn-Gettler, Kendeou, Carlson, & White, 2011; van den Broek & Helder, 2017). Cognitive abilities, such as working memory (WM) resources, are thought to be important to comprehension as they are theorised to coordinate the higher-level processes required for comprehension (Kendeou et al., 2014) (I am referring to WM as a resource here and, in Chapter 2, I discuss this in more detail).

1.1.1. Reading Comprehension Measures

As reading comprehension models are theoretical accounts of reading that are built on observations of behaviour, it is important to first provide an overview of reading comprehension measures. To understand written texts, individuals read texts bit by bit, moving their eyes back and forth through the text. As a result of this process, they comprehend bits of text and construct a coherent representation of the situation described in the text they read using these comprehended bits of information and their own background knowledge. Researchers have attempted to capture the comprehension process by observing elements of this behaviour using various performance measures. These measures are typically grouped into two types, on-line and off-line (Kintsch & Rawson, 2007).

On-line measures are used during the reading process, for example, by recording the time spent reading a specific part of the text, such as a sentence or a paragraph. Other on-line measures consist of speeded response tasks that include lexical decision (LD) and word naming. LD involves deciding as quickly as possible whether a string of letters is a word or a non-word, whereas word naming comprises pronouncing displayed words as quickly as possible. Kintsch and Rawson (2007) argued that these on-line measures capture the actual processing of the text when it is happening, and that they can be used for studying the underlying processes of reading comprehension. However, it cannot be assumed that all on-line measures reflect processing performance transparently.

One of the criticisms of on-line measures is that they can potentially be disruptive to the process of reading comprehension, and therefore they may not always offer an accurate insight into reading comprehension processes (Kintsch & Rawson, 2007). However, this concern can be overcome with the use of eye-tracking technology. The analysis of eye movements during reading is a direct method for measuring real-time processing demands during comprehension without interrupting individual’s processing of the text (Raney, Campbell, & Bovee, 2014; Roberts & Siyanova-Chanturia, 2013). Cognitive demands can be studied by observing several aspects of eye movement behaviour, including fixation durations, number of fixations, and number of regressions which refers to the number of returns to previous parts of a text (Raney et al., 2014). In reading comprehension research, the basic assumption of eye-tracking methods is that the increase in cognitive demands imposed by the text is associated with longer processing times or changes in fixation patterns. Slower processing time can be reflected by an increase in the number of fixations or longer fixation durations (Raney et al., 2014). However, eye movements alone do not necessarily reveal whether the increase in cognitive demands imposed by the text leads to successful comprehension, and they also fail to yield insights into readers' thought processes (e.g., Reichle, Reineberg, & Schooler, 2010).

In contrast to on-line measures, it is thought that off-line measures can reveal whether the text was understood or not without interfering with reading processes (Kintsch & Rawson, 2007). Unlike on-line measures, off-line measures are taken after reading has taken place. There are various off-line measures, however the commonality between them is that they frequently involve responding to questions about the text read. These questions can be grouped into categories. One category of questions involves multiple choice questions, these are questions that require participants to select a response from a list of answers that are presented to them. Another category of questions constitutes recall questions, where individuals are asked to give an answer that requires information retrieval from their memory about the text read. Last, there are also short answer questions, open-ended questions which require a short answer from the reader. Typically, these questions target memory for the text read, assess deeper understanding of the passage, or both (Kintsch & Rawson, 2007).

Although off-line measures are thought to be better estimators of the lasting representational outcome of reading comprehension than on-line measures, they do not provide as much information as on-line measures about the ways in which reading processes operate (Kintsch & Rawson, 2007). Off-line measures are also prone to the loss of information caused by readers forgetting what they read and rely on the readers accurately describing the understanding of the read material. The latter is problematic, because readers may not always be able to describe what they read or what factors led to them comprehending the text read. Furthermore, the readers may not realise that they do not understand the text they read. Thus, due to the contrasting strengths and weaknesses of the two types of measurements, as well as concerns relating to their validity and reliability (e.g., Kintsch & Rawson, 2007), it can be argued that reading comprehension is best studied with a combination of off-line and on-line measures to offer a broader picture of comprehension processes.

1.1.2. Reading Comprehension Models

There are several models, within different frameworks, specifying the processes that are thought to be critical to reading comprehension. The Construction-Integration Model is concerned with the steps involved in getting from text to a coherent model of the meaning of the text as reconstructed by the reader (Kintsch, 1988). In contrast, the Simple View of Reading deals with the development of reading comprehension within an individual (Gough & Tunmer, 1986). Specifically, the Simple View of Reading focuses on individual differences and how readers may vary in their ability to recognise words and understand the language being read. Another approach, the Lexical Quality Hypothesis, concentrates on textual processing at the lexical level (Perfetti, 2007). The Lexical Quality Hypothesis proposes that variation in the speed and efficiency in retrieval of mental representations of words influences reading comprehension (Perfetti, 2007). These three models are not the only models of reading comprehension; however, they could be classed as dominant models within the component skills and the text and discourse processing frameworks (Francis et al., 2018). Furthermore, as I mention later, the Simple View of Reading model can be argued to take into account the developmental framework as well. This is because the Simple View of Reading model considers the changing relationship between the effects of some individual differences and reading comprehension across development.

Critically, although the Simple View of Reading model has been predominantly applied to the study of comprehension in children, the Simple View of Reading can be extended to the study of comprehension across the lifespan (cf. Francis et al., 2018). This is because research evidence indicates that the influence of comprehension processes, that underlie successful comprehension, changes in strength not just during childhood, but also during adulthood (e.g., Garcia & Cain, 2014) (discussed in section 1.3). In addition, the Simple View of Reading model has been successfully applied to the study of comprehension of some populations of adult readers, such as those with relatively low levels of literacy (e.g., Braze et al., 2016; Braze, Tabor, Shankweiler, & Mencl, 2007; Sabatini, Sawaki, Shore, & Scarborough, 2010). This is important, and relevant to my research, as my thesis is also concerned with adults who may have relatively low literacy levels. Therefore, in this chapter, I discuss the Construction-Integration model, the Simple View of Reading, and the Lexical Quality Hypothesis.

1.2.The Construction-Integration Model (Kintsch, 1988)

In 1970s, 1980s, and 1990s, comprehension researchers developed the concept of a mental model, also referred to as a situation model (e.g., Johnson-Laird, 1980; Kintsch 1988; Kintsch & van Dijk, 1978). The main notion within the accounts assuming the importance of situation models is that text understanding is reliant on the construction of a mental representation of the situation represented by the text read instead of the construction of a representation of the text itself (Zwaan, 2016). This is because readers do not just understand what the text conveys, they construct the model of the situation represented by the text as they integrate information from the text with information from their background knowledge. Importantly, it can be argued that the most recognised version of the situation models is the Construction-Integration (CI) model (Kintsch, 1988; 1998), which is an extension of the text recall model (Kintsch & van Dijk, 1978).

The CI model describes the types of information represented in comprehension, and the processes involved in it. According to the CI model, reading comprehension involves textual processing at different levels (Kintsch, 1988). These levels consist of surface and text level processes, micro- and macrostructure which form the textbase, and the situation model. First, while operating at the surface level, the reader must process words and phrases contained in the text itself (Kintsch & Rawson, 2007). This processing of words and phrases is thought to rely on perceptual processes, specifically word recognition, and the assignment of words to their roles in sentences and phrases in a process known as parsing. Second, at the text-level, to determine the meaning of the text read, the comprehender has to join individual words' meanings to form propositions, propositions are the meanings of the sentences (Kintsch & Rawson, 2007).

Propositions are thought to be interconnected by the reader, in an active inferential process, in a complex network, forming the microstructure of the text (Kintsch & Rawson, 2007). Propositions can be linked to each other via cause and effect relationships (logical implications), or co-reference. Argument overlap happens when at least two propositions are linked to the same concept by nouns, pronouns, and so on. Readers are theorised to create the microstructure mentally by studying the coherence relations between propositions which they construct based on the meaning of the words in the text and the syntactic relationships between these words. Additionally, to build a logical microstructure, readers are thought to often be required to make inferences. Individuals generate an inference when a specific relation between parts of the text is not explicitly described but is filled by their own knowledge of the world, of the topic of the text, and of the text itself to make sense of the text read (Kintsch & van Dijk, 1978).

Kintsch and Rawson (2007) argued that sections of the text are organised by the reader semantically in specific ways. The microstructure is theorised to be organised into higher order units, referred to as the macrostructure. A key feature of the macrostructure is concerned with the identification of important themes in the text. Some texts contain signalling devices that indicate the themes within them, for example, titles, outlines, summaries, or abstracts. These can improve recall of the information described within the themes of the texts (Lorch, Lorch, & Inman, 1993). However, in the absence of signalling devices, readers will use textual cues such as topic sentences, and surface cues, for example typeface, repetition of concept words, or structural feature of the text, to identify the themes within the texts. In addition, topic identification can also be influenced by relevant prior knowledge of the reader, such as prior knowledge about the representative text structure within the domain of the text read.

The microstructure and the macrostructure are thought to form the subsequent level of text representation, the textbase. The textbase is the product of processing at the surface level, it represents the meaning of the text as it is explicitly given by a network of concepts and propositions derived from the text (Kintsch, 1998). However, the comprehension of the explicit meaning of the text is only sufficient to reproduce the text in recall or other memory tests, but not to develop a deep understanding of it. Deep understanding can be achieved with the construction of a situation model, that is, a mental model of the situation described by the text (Kintsch & Rawson, 2007).

The development of a mental model requires that information provided by the text is integrated with relevant prior knowledge (Kintsch & Rawson, 2007), including relevant memories, beliefs, emotions, and goals (Kintsch, 1998). Without retrieving information from prior knowledge and integrating it with the new information provided by the text, the reader is unlikely to fully understand the text read. Inferences, which are described later, are thought to be critical in constructing the textbase, and in forming a logical situation model. Since texts cannot be fully explicit, there are always gaps for the reader to make inferences about the meaning of the text based on their prior knowledge. These gaps can be local, where the reader has to make inferences between small parts of text, or global, where the theme of the text is not explicit, and the readers have to construct it themselves. It is important to mention that readers' goals can influence the development of the situation model, since reading goals are likely to influence readers' standards of coherence (van den Broek & Helder, 2017).

Reading often involves standards of coherence (van den Broek et al., 2011). These standards can be implicit or explicit and may not involve conscious decisions on the part of the reader. Consequently, the reader may not be aware of the standards they employ, until these standards are violated (van den Broek et al., 2011). High levels of comprehension require the reader to adopt high standards of coherence (van den Broek & Helder, 2017). Standards of coherence are important, because they can influence comprehension through the initiation of passive and reader-initiated processes. Passive processes are associative processes through which information in the text read activates information from memory for the prior text and from comprehenders' background knowledge (van den Broek & Helder, 2017). These processes take place outside of reader's conscious control and can be measured using reading times and eye-tracking measures, such as fixation durations and reinspection (e.g., Yeari, van den Broek, & Oudega, 2015).

Reader-initiated processes do not always occur when reading. Reader-initiated processes require control and WM attentional resources, therefore consuming time and effort (van den Broek & Helder, 2017). However, reader-initiated processes can improve comprehension beyond the level resulting from passive processes alone. Reader-initiated processes range from simple actions, such as re-reading the sentence, to more complex reading strategies such as note-taking, reflecting, comparing with other documents, and generating inferences. These reader-initiated processes can be measured using think-aloud procedures and free recall (e.g., Narvaez, van den Broek, & Ruiz, 1999; van den Broek, Lorch, Linderholm, & Gustafson, 2001). If adequate level of understanding of a text read is not achieved using passive processes alone, the reader is likely to engage in reader-initiated processes, such as inferences, to build coherence (van den Broek & Helder, 2017). If reader's goal is to develop a superficial level understanding of text read, they are likely to engage in fewer reader-initiated processes. On the other hand, if a reader is highly motivated to develop deep understanding of the text read, they are likely to engage in more reader-initiated inference making (van den Broek & Helder, 2017). In addition to goals, standards of coherence can vary as a function of individual and developmental differences, properties of the text and the reading situation. However, reading-initiated strategy use required to attain high standards can be acquired through practice and become subsequently automatised (van den Broek & Helder, 2017).

Inferential processes are important to comprehension as they help readers identify semantic relations in text (van den Broek, Rapp, & Kendeou, 2005). Inferences vary in the cognitive demands imposed on the reader (Kintsch, 1998). This is because they differ along two dimensions. First, inferences can be controlled or automatic (Kintsch & Rawson, 2007). The former are assumed to require more cognitive resources, as in the case of syllogistic reasoning which encompasses integrating information, making inferences, and considering alternative states (Segers & Verhoeven, 2016). A syllogism consists of two premises that are assumed to be true and a conclusion, below is a simple example of a disjunctive syllogism, one characterised by "either…or" statement:

Premise 1: Either pigs will learn to fly, or fossil fuels will run out.

Premise 2: Pigs will not learn to fly.

Conclusion: Therefore, fossil fuels will run out.

Syllogistic reasoning requires individuals to arrive at the right conclusion based on the premises of the syllogism. Compared to controlled inferences such as syllogistic reasoning, it is assumed that automatic inferences, such as bridging inferences, are made effortlessly and quickly. For example, in "Kathy owned a house. The gutters were blocked." the inference being made is that the house has gutters, and this inference is made rapidly by an average reader. A further dimension on which inferences differ is whether they are knowledge- or text-based. Knowledge-based inferences occur when readers' prior knowledge enables them to make an inference. In contrast, text-based inferences require the reader to use the information provided in the text to make an inference (Kintsch & Rawson, 2007).

Inferences in reading comprehension often involve automatic knowledge activation. However, as the themes described by the texts become less familiar, the importance of controlled inferencing increases. This is because readers must retrieve and activate the most relevant prior knowledge they have. However, readers reading unfamiliar text will not have highly relevant prior knowledge or experience and their retrieval process is likely to be more demanding on WM resources than the process of retrieval for readers with prior information. This is because those with prior knowledge of the text read are likely to have relevant prior knowledge available for retrieval and are also likely to be more efficient at retrieving it than those without relevant prior knowledge (Kintsch & Rawson, 2007). Conversely, readers without relevant prior knowledge might need to consciously engage in retrieval of potentially related prior information, while trying to simultaneously inhibit irrelevant information stored in their memory and keeping the relevant information active.

From the perspective of situation models, WM is theorised to play an important role in comprehension, because it is assumed that information processing occurs in the finite capacity of WM (Kintsch & Rawson, 2007). In contrast to more recent theories supposing the importance of WM to comprehension, discussed in more detail in Chapter 2 (section 2.1.1), Kintsch, Patel, and Ericsson (1999) argued that long-term WM (LTWM) could account for the information readers have to maintain in their WM to comprehend the text read. Kintsch et al. (1999) proposed two WMs. Short-term WM (STWM) which is capacity limited and equated by them with content-of-consciousness or the focus-of-attention, and WM which includes a LT component and is restricted to practiced and familiar knowledge domains. The LT component of WM contains everything in readers' LT memory that is connected to the present contents of ST memory through retrieval structures. The retrieval structures enable instant access to information from LT memory that is relevant to the task being carried out, without resource intensive retrieval processes. According to Kintsch et al. (1999), LTWM enables people to perform exceptionally well in their expert domains. For example, LTWM is thought to allow an experienced chess player to determine the next move without having to spend a lot of time thinking about it. However, Kintsch et al's. (1999) conceptualisation of WM is not compatible with the more mainstream theories of WM where WM is closely related to attentional processing and consciousness, but it is not equated to them (e.g., Repovš & Baddeley, 2006) (discussed in Chapter 2, section 2.1.1).

Overall, the CI model describes the complexity of the processes involved in comprehension, as well as the types of information that have to be represented in it. Additionally, it specifies the different processes involved at different levels, such as at the word, sentence, paragraph, and whole passage levels. The model highlights the importance of WM and prior relevant knowledge (Kintsch & Rawson, 2007). However, the CI model tends to focus mainly on the effects of text and features of the texts on comprehension, and to a smaller extent on how the effects of reader attributes influence comprehension. In other words, the CI model mostly focuses on the processing of information in discourse and does not explain developmental and individual differences. The influence of the effects of individual differences in reading comprehension has been the focus of the component skills and developmental framework exemplified by the Simple View of Reading (SVR; Gough & Tunmer, 1986) model which I discuss next.

1.3. The Simple View of Reading (Gough & Tunmer, 1986)

The SVR model explains individual differences in reading comprehension, comprehension of written passages, in terms of differences in two skills. Specifically, the ability to efficiently recognise words and apply knowledge to letter-sound relationships to construct their phonological form (decoding), and all the skills and capacities needed to understand discourse in its oral form (linguistic comprehension) (Gough & Tunmer, 1986). This view is critically different from the CI model's perspective, because the SVR model assumes that paragraph, word, and sentence level skills are components of one of the two skills, whereas the CI model sees them as parts of different levels of representation, such as the surface level or the textbase level (Gough & Tunmer, 1986: Kintsch, 1998). However, although the CI and SVR models are different, they are not necessarily competing theories. This is because the SVR model is concerned with identifying resources and skills necessary to understand a text, but these resources could correspond to multiple processes or levels of representation as mentioned in the account of the CI model.

The modified version of the SVR model (Tunmer & Chapman, 2012), suggests that decoding and linguistic comprehension have a reciprocal influence on each other, rather than being independent of each other as originally proposed by Gough and Tunmer (1986). Tunmer and Chapman (2012) argued that both components are influenced directly, and indirectly, by other variables, such as vocabulary knowledge. Tunmer and Chapman's (2012) factor analysis of data obtained from 122 seven-year-olds shows that vocabulary knowledge and listening comprehension, comprehension of read aloud passages, load highly onto the linguistic comprehension factor. The structural equation model of their data reveals that vocabulary knowledge influenced reading comprehension, not only directly, but also indirectly through decoding. Furthermore, in a separate three-year longitudinal study Tunmer and Chapman (2011) found that vocabulary knowledge, measured at the beginning of the first year of their study, correlated with third-year score on a reading comprehension measure and was indirectly associated with third-year phonological decoding score. Therefore, Tunmer and Chapman (2012) concluded that vocabulary knowledge, linguistic comprehension, and decoding skills are interdependent.

A central feature of the SVR is the developmental assumption that as word reading becomes more fluent and efficient, approaching maximum, the relative proportion of variance in comprehension performance explained by variation in word reading skill will decrease, whereas linguistic comprehension processes will start to play a more influential role (Gough & Tunmer, 1986). This is because, over time, the increasingly diverse and advanced texts, written in English, to which developing readers are exposed make greater demands on higher-level language skills, such as vocabulary knowledge, rather than decoding skills (Vellutino, Tunmer, Jaccard, & Chen, 2007). This has been supported by the results of other developmental studies that focused on the reading comprehension of English texts (e.g., Garcia & Cain, 2014), some of which included English as a second language (ESL) speakers (e.g., Language and Reading Research Consortium (LARRC), 2015). In a meta-analysis of 110 studies drawing on observations from the total number of 42,891 claimed-to-be first language (L1) English readers, ranging in age from five to 53 years, Garcia and Cain (2014) found that the weaker the correlation between decoding and reading comprehension, the stronger the links between linguistic comprehension and reading comprehension became.

Vellutino et al. (2007) also found that the relationship between decoding and reading comprehension was stronger in the younger than in the older group of readers from the United States. Conversely, the relationship between linguistic comprehension and reading comprehension was stronger in the older than in the younger group. Similarly, LARRC (2015) reported, given cross-sectional data analyses of 371 U.S. six to nine-year-olds from different L1 language backgrounds, although mostly L1 English, that the influence of word recognition on reading comprehension diminished, relative to the influence of listening comprehension, over time.

There is also non-developmental evidence to suggest a relatively small role of decoding in adult readers. For example, a large-sample investigation of 737 U.S. 18-year-olds, 33% of whom were second language (L2) English speakers, observed that word reading fluency had a negligibly small effect on reading comprehension (Cromley, Snyder-Hogan, & Luciw-Dubas, 2010). The findings of the developmental and non-developmental studies support the SVR because they suggest that the relationship between decoding, linguistic comprehension, and reading comprehension, changes across lifespan, which is predicted by the SVR (Gough & Tunmer, 1986; Tunmer & Chapman, 2012). A prediction, which can also be derived from the assumptions of the CI model (Kintsch, 1988), is that the demand on cognitive resources imposed by word recognition declines with age due to the increase in skill level. Thus, the older readers can devote greater resources to constructing meaning from text than the younger ones (Garcia & Cain, 2014).

It is important to note that the relationship between decoding and reading comprehension could vary not only for developmental reasons but also for statistical reasons. Decoding skill can be examined using different measures depending on the age of the participants which can make direct comparisons between age groups difficult. In the early stages of reading, decoding is best assessed using measures of non-word reading ability, which focus on the ability to convert text to speech with phonological information (Hoover & Gough, 1990). In contrast, decoding skill of the more skilled readers is generally measured using word reading ability which also often involves assessing decoding speed. However, because skill level of decoding increases as people become more proficient readers with maturation, decoding skill gradually reaches a ceiling level (Plaut, McClelland, Seidenberg, & Patterson, 1996). Thus, the impact of increases in decoding on reading performance tends to diminish with age. This is because after reaching a physiological threshold in decoding performance, different individuals reaction times cluster so closely together that discriminating between their comprehension performance based on the speed of their decoding skill is impossible (see also Garcia and Cain (2014) for an argument that these differences are not a statistical artifact).

Summarising, the SVR model focuses on the effects of individual differences on reading comprehension, but it also acknowledges that the strength of the effects of some individual differences on comprehension varies across lifespan. Thus, although the SVR model is grounded within the component skills framework of reading comprehension research, it also considers the developmental framework. Nevertheless, the SVR model cannot fully account for variation in reading comprehension performance. This is because reading comprehension is not only a product of word recognition and listening comprehension (e.g., Perfetti, Landi, & Oakhill, 2007). Other factors are also involved, for example, variation in the quality of semantic knowledge which refers to the knowledge of the meanings of words and phrases. Individuals can vary in their capacity to access word meaning knowledge efficiently and this is likely to be associated with differences in comprehension (e.g., Bruck, 1990; Nation & Snowling, 1998; 2004).

Critically, when word recognition is slow, more resources are thought to be directed to word-level processes instead of higher-level processes (Perfetti, 1985). This can result in most of the WM resources, including attention, being directed towards recognising words, instead of the higher-level processes, such as inference making, which are needed to build an understanding of the text read. In contrast, when decoding is fast, meaning that it is efficient and automatic, more WM resources can be devoted to high-level comprehension processes. Skilled reading requires efficiency with processing word-level information and only readers who have efficient and automatic decoding can achieve a high-level of comprehension (Perfetti, 2007).

1.4. The Lexical Quality Hypothesis (Perfetti, 2007)

Perfetti (2007) argued that efficient processing is underlined by knowledge about meanings and word form properties, such as phonology and orthography. He referred to efficiency as the rapid low-resource retrieval of the orthographic, phonological and semantic constituents of word's identity. Word identification is theorised to involve selecting appropriate mental representations of words from readers' mental lexicon (lexical selection) and accessing word form properties. This two-stage process is referred to as lexical access (Harm & Seidenberg, 2004). According to Harm and Seidenberg's (2004) computational model of reading, in word recognition for comprehension, lexical access is likely to involve determining meaning directly from orthography, or indirectly where phonology serves as a bridge between orthography and meaning. Thus, orthography and phonology are thought to be critical to successful comprehension.

Lexical quality (LQ) is the degree to which an individual's knowledge of a given word represents the word's form, meaning, and the contexts in which the word is used (Perfetti, 2007). Individuals differ in the LQ of the words they know, and readers' lexicons will include words of varying LQ, from rare words which the readers rarely encountered to known, frequent, words (Perfetti, 2007). Quality refers to the extent to which a mental representation of a word specifies its meaning and form in a way that is flexible and precise.

To be considered high-quality, lexical representations need to be precise, in other words they should relate to a single orthographic representation of one lexical item. Precision is important in comprehension, because it enables the reader to activate the lexical representation corresponding to sensory input, minimising the chance of activating competing lexical items (Andrews & Hersch, 2010). Lexical representations also have to be flexible because some words or their definitions are interconnected and may mean the same thing, for example, “social interactions” and “an exchange between two or more individuals” share the same meaning. Flexibility arises from the binding between the different parts of lexical representations (Andrews, 2015). The implementation of precision and flexibility helps individuals overcome form-meaning complexities encountered in everyday life. For example, both precision and flexibility are needed to comprehend and pronounce some words, such as *lead* in “She will *lead* us home” and “She wants to buy *lead*”.

There are five construct labels that distinguish high from low-quality lexical representations (Perfetti, 2007). These include orthography, phonology, meaning, grammar, and constituent binding. Constituent bindings are connections that establish coherence among the orthographic, phonological, and semantic representations, which together constitute the word’s identity (Perfetti, 2007). In high-quality representations, the features of word identity are more tightly bound with each other than in low-quality representations. Tight connections allow word forms to trigger synchronous and coherent activation of all parts of a word’s identity that are needed for successful comprehension (Andrews, 2015). The more tightly bound the constituents are with each other, the more coherent is the lexical representation of the word read and the less likely it is that the word will be associated with a representation of another similar word.

One of the consequences of the variation in LQ is the different level of meaning integration (Perfetti, 2007). For successful comprehension, words must link with specific mental representations or attractors (Harm & Seidenberg, 1999; Plaut & Shallice, 1993). Perfetti (2007) hypothesised that high LQ is characterised by the presence of word identities which are available for constructing comprehension, creating a connection between the word identification system and the comprehension system. In contrast, low LQ is identified by at-risk comprehension processes operating over word identities (Perfetti, 2007). These processes are at risk because among individuals with low LQ, word knowledge is underlined by lack of orthographic precision and phonological specificity. Thus, for readers with low LQ, context-sensitive word-to-text integration processes that maintain coherence require more working memory (WM) resources as they are slower and less efficient (Perfetti & Stafura, 2014; Yang, Perfetti, & Schmalhofer, 2005). These word-to-text integration processes can be linked to the CI model's concept of the textbase (Kintsch, 1998), as they include paraphrasing, inference making, and pronoun binding. In contrast to comprehenders with low LQ, those with high LQ can execute the meaning integration processes efficiently with minimal WM resource demands (Yang, Perfetti, & Schmalhofer, 2007).

Another consequence of the variation in LQ is synchronicity (Perfetti, 2007). The Lexical Quality Hypothesis (LQH) predicts that, when LQ is high, word identity components will be activated and retrieved in synchrony when reading a word. Synchronous activation means that representations of the word read will be activated more strongly and coherently while inhibiting representations of other similar words, preventing activation of incorrect meanings (Andrews, 2015). In contrast, when the LQ is low, word identities may be activated and retrieved at different times, resulting in a diffused activation across multiple letter and word meanings.

Synchronicity can also be linked to the CI model (Kintsch, 1998), as the construction of the situation model is likely to be dependent on establishing the correct meaning of the situation presented in the text. Asynchronous activation of word identity constituents could lead to activation of incorrect meanings (Perfetti, 2007) and therefore an inappropriate situation model. This is because readers who asynchronously activate components of word identity do not have access to semantic and grammatical information needed for successful word recognition and comprehension (Andrews, 2015). Thus, they must employ additional attentional WM processing resources to support word identification and comprehension. This implies that WM resources might be spent inefficiently if lexical representations are of low quality. The more WM resources are taken up by word recognition, the fewer WM resources are available for higher-level processes and comprehension. Consequently, comprehension is likely to suffer.

Critically, like Gough and Tunmer (1986), Perfetti (2010) claimed that word decoding has a central role in the development of reading comprehension. This is because automatic word decoding is thought to enable readers to devote more mental resources to generating the meaning of a text, and thereby allows readers to acquire new information and knowledge (Perfetti, 1998). The automatization of word decoding, and being able to read with speed, accuracy, and proper expression, is also referred to as the attainment of reading fluency (Perfetti, 1992). In a model, referred to as the DVC model, Perfetti (2010) argued that decoding, the knowledge of the meaning of a word (vocabulary), and reading comprehension, are interconnected (Figure 1.1).

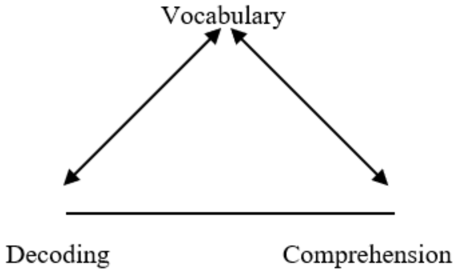


Figure 1.1. The DVC reading skill triangle (Perfetti, 2010).

Perfetti (2010) specified that decoding, vocabulary, and reading comprehension together form the reading skill. Vocabulary is said to include the breadth and depth of word knowledge, whereas the comprehension component is thought to consist of sentence, text, and knowledge-based inferential, as well as other, procedures. Since the three components are interconnected, limitations in one will affect at least one other component, and influence the overall reading skill. Perfetti suggested that decoding influences vocabulary through two processes. First, successful decoding strengthens word-meaning connections. Second, it creates associations between unfamiliar words and familiar contexts. In turn, vocabulary influences decoding, because decoding a known word strengthens the link between its orthography and meaning. Perfetti stated that comprehension is influenced by vocabulary, because it is dependent on the knowledge of the meaning of the words being read. The relationship is thought to be reciprocal, because achieving comprehension from a sentence, which includes an unfamiliar word, can result in the reader learning something new about the meaning of that word (Perfetti, 2010).

Overall, both the DVC model and the LQH assume that word meanings are crucial to word identification and text comprehension (Perfetti, 2010). Furthermore, the LQH can be perceived as a middle ground between the individual differences account of the SVR model, and the comprehension processing account of the CI model (Gough & Tunmer, 1986; Kintsch, 1998; Perfetti, 2007). Regarding the former, it demonstrates how variation in LQ has consequences for text comprehension. It considers the latter, because it recognises the need for a coherent mental model, in order to develop deep understanding of the text being read. The LQH recognises the need for both lower and higher-level processes in reading comprehension, as words and sentences can be seen as the foundation of meaning (Perfetti, 2010). In turn, mistakes at the word and sentence level may limit processing at the higher-level required to build the mental model of the text. In addition, limitations in cognitive abilities, such as WM resources, including capacity and attention-allocation, are also thought to influence processing at the lower and higher-level. This is especially the case among the less skilled readers who rely on inefficient word identification which requires additional WM resources, leaving less resources for WM-resource-demanding higher-level comprehension processes, such as integrating information within and between sentences (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003).

Critically, although the LQH (Perfetti, 2007) can be classed as a complete model of the component skills framework of reading comprehension, it does not explicitly incorporate much of the text and discourse framework and the developmental framework. Thus, it does not account for all the potential factors that influence reading comprehension. From the LQH it can be concluded that accurate and fluent word reading translates into efficient processing of words and sentences; vocabulary knowledge and decoding are important in comprehension because they aid in understanding relations between words and meanings of sentences; and WM is crucial for its role in important comprehension processes such as inference making. In the next section, I briefly discuss the reading comprehension research grounded within the text and discourse framework.

1.5. Text and Discourse Framework

Research grounded within the text and discourse framework has focused on the features of texts and linguistic discourse and how these features influence comprehension, and to a smaller extent on the cognitive, linguistic and motivational characteristics of readers (Francis et al., 2018). Many text and discourse researchers (e.g., van Dijk & Kintsch, 1983), have argued that one of the fundamental properties of discourse is coherence. Coherence is relatively difficult to define, because it is a relatively abstract concept that is closely related to cohesion. Whereas text cohesion is the degree to which the concepts, ideas, and relations within a text are explicit, text coherence can be thought of as the effect of text cohesion on readers’ comprehension (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, 2001; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996; O’Reilly & McNamara, 2007; Ozuru, Dempsey, & McNamara, 2009). According to this definition of coherence, text coherence can be measured using text features, which I discuss later, that are associated with text cohesion. However, the above definition is problematic since highly cohesive texts, which contain cohesive ties between sentences, are not always coherent (e.g., Hamilton & Oakhill, 2014).

Coherence can also be defined as sense relations between sentences or propositions of a text, due to which the text appears to be logically and semantically consistent for the reader. This definition largely corresponds to van Dijk and Kintsch’s (1983) definition of local coherence. However, this definition is also problematic since it is likely to create additional uncertainty in the measurement of text coherence as assessed using indices of linguistic features. This is because it is questionable whether measurements of linguistic features of a text can detect sense relations within that text. Consequently, some text features, which I discuss later, can be thought of as proxies of coherence, but they are likely to include a significant amount of measurement error.

Theoretically, variation in text coherence and cohesion is thought to influence comprehension (e.g., Kintsch, 1988; Kintsch, 1998; McNamara & Kintsch, 1996). Texts that are not coherent may require the reader to establish coherence by generating inferences to fill the gap between sentences using their background knowledge (Hamilton & Oakhill, 2014; van Dijk & Kintsch, 1983). In turn, texts that are not very cohesive are likely to increase the processing demands on WM resources in the construction of the situation model of the text (Kintsch, 1998; Meyer, 2003). As a consequence of the theoretically hypothesised influence of cohesion and coherence on comprehension, there has been a considerable interest in identifying the text features associated with different levels of comprehension.

In the applied research settings, such as hospitals and schools, where there are concerns about text comprehension, many researchers use textual readability measures, such as the Flesch Reading Ease (Flesch, 1948), to assess text readability (e.g., Wang, Miller, Schmitt, & Wen, 2013). However, most of these textual measures of readability are out-dated. They tend to be based on simple indices, such as sentence and word length, that may correlate with text difficulty, but do not account for what is theorised to make a text easier or more difficult to comprehend (Crossley, Greenfield, & McNamara, 2008; McNamara & Magliano, 2009; Kintsch & Vipond, 1979), such as text cohesion and coherence (Kintsch, 1998; van Dijk & Kintsch, 1983) (I discuss the concept of readability, and textual measures of readability, in Chapter 3, section 3.2).

One of the main measures used to assess text coherence is co-reference (Kintsch & Rawson, 2007). Co-reference can be measured using indices of argument overlap and conceptual overlap. High argument overlap indicates that two or more propositions refer to the same concept, whereas high conceptual overlap demonstrates that the propositions share words that are similar in meaning (Graesser, McNamara, & Kulikowich, 2011). The higher the incidence of argument and conceptual overlap, the more closely bound the sense relations between sentences are thought to be, and the easier it should be for the reader to link propositions together and construct the textbase (Kintsch, 1988).

Text cohesion can be manipulated with the use of cohesive ties, such as connectives. Connectives are connecting words that link propositions and clarify relations in the text (Kintsch, 1998). Connectives can be subdivided into causal, *because*, *so*; temporal, *then*, *after*; logical, *therefore*, *if*; additive, *additionally*, *furthermore*; and adversative, *on the contrary*, *however* (Graesser et al., 2011). By helping readers to link propositions, connectives aid the comprehenders in constructing the textbase (Kintsch & Rawson, 2007). Specifically, cohesive ties are thought to increase text cohesion as they prompt readers to generate inferences spontaneously in the right places when reading for understanding (Hamilton & Oakhill, 2014). Consequently, texts that are cohesive and coherent are unlikely to require as many reader-initiated processes to reach adequate levels of understanding as the use of connectives is likely to prompt the reader to generate inferences passively (Hamilton & Oakhill, 2014; van den Broek & Helder, 2017).

Not all texts are highly cohesive and coherent. For example, in an analysis of social studies texts, many texts were found to contain loosely connected statements that were difficult to integrate with previous sections of the text (Beck, McKeown, & Gromoll, 1989). Lack of connectives and low co-reference levels are likely to impede comprehension by forcing the reader to engage in conscious, and effortful, inference-making required to construct a logical situation model of an incohesive and incoherent text (Kintsch, 1988; van den Broek & Helder, 2017). Whether readers engage in this inference-making will depend, amongst other factors, on their standards of coherence and background knowledge (van den Broek & Helder, 2017; McNamara & Kintsch, 1996). Nonetheless, on average, revising relatively incoherent and incohesive texts, by adding connectives and increasing co-reference, has been found to improve comprehension levels of these texts (Beck, McKeown, Sinatra, & Loxterman, 1991). Indeed, past research demonstrates that, for an average reader, improving text cohesion improves comprehension (e.g., Britton & Gülgöz, 1991; Lehman & Schraw, 2002; Linderholm et al., 2000; Vidal-Abarca, Martínez, & Gilabert, 2000). However, research evidence, which I discuss next, suggests that varying text cohesion and coherence is unlikely to have the same effect on all readers.

It is important to acknowledge that individuals vary in their inference-making abilities, for example due to their levels of background knowledge (McNamara & Kintsch, 1996). Therefore, texts that do not cohere are likely to be understood differently by individuals who have different levels of background knowledge (McNamara & Kintsch, 1996). This has been demonstrated in a set of studies motivated by the CI model of comprehension (Kintsch, 1988). McNamara and colleagues (McNamara, 2001; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996) tested comprehension of middle school and university students on a set of original and revised informational texts. The texts used in these studies related to heart disease (McNamara et al., 1996), the Vietnam War (McNamara & Kintsch, 1996), and cell mitosis (McNamara, 2001). To create the revised texts, McNamara and colleagues manipulated cohesion and coherence by increasing the argument overlap between propositions and the incidence of causal connectives. In all three studies, they found evidence for the reverse cohesion effect. The reverse cohesion effect refers to the finding that comprehension of high-background-knowledge readers was higher when reading low-cohesion texts, compared to high-cohesion texts. However, for low-background-knowledge readers, low-cohesion texts were detrimental to understanding, whereas high-cohesion texts were beneficial.

The results of McNamara (2001), McNamara and Kintsch, (1996) and McNamara et al. (1996) support the assumption that inferences are more likely to be generated when the text prompts the reader to engage in inference-making, but only when the comprehender has the relevant background knowledge to make an inference. High-background-knowledge readers might understand more from less coherent and cohesive texts, as they may be more likely to engage in reader-initiated compensatory processing to infer relations between propositions in texts (Kintsch & Rawson, 2007; van den Broek & Helder, 2017). In contrast, low-background-knowledge readers might not engage in such processing, because they do not have sufficient background knowledge to do so. Through reader-initiated processing, high-background-knowledge readers are likely to integrate the information based in the text read with their textbase. As a result of this they are likely to build a more logical situation model (Kintsch, 1998; Kintsch & Rawson, 2007). In turn, exposing high-background-knowledge readers to high-coherence texts may reduce the perceived need for reader-initiated processing of these texts. As such, high-background-knowledge readers might think that their standards of coherence are being met without engaging in reader-initiated processing (van den Broek & Helder, 2017). Consequently, their understanding of the text is likely to be lower as they are likely to generate fewer inferences to build a coherent situation model than they would have done while reading a low-coherence text.

However, not all high-background-knowledge readers are affected equally by the reverse cohesion effect when they read highly cohesive and coherent texts (e.g., O'Reilly & McNamara, 2007; Ozuru, et al, 2009). O'Reilly and McNamara (2007) and Ozuru et al. (2009) examined whether the reverse cohesion effect is dependent on individuals' reading skill. In both studies, university students' comprehension was tested on a set of informational biology texts, where one set of texts was revised to be more cohesive and coherent, through the use of cohesive ties and co-reference, whereas the other set was left relatively incohesive. The students were also tested on their relevant background knowledge, and their reading skill was assessed using the Nelson-Denny reading comprehension ability test (Brown, Fishco, & Hanna, 1993). O'Reilly and McNamara (2007), and Ozuru et al. (2009), found that low-background-knowledge participants better comprehended texts if they were skilled readers, suggesting that reading skill partially compensated for their low levels of background knowledge. In turn high-background-knowledge readers exhibited the reverse cohesion effect, but only if they were less skilled comprehenders. Comprehension of high-reading-skill, high-background-knowledge readers was higher for the cohesive and coherent texts than for the relatively incohesive and incoherent texts.

O'Reilly and McNamara (2007) and Ozuru et al. (2009) explained their findings in terms of different levels of processing employed by the less skilled versus the more skilled readers. Ozuru et al. argued that high-cohesion texts led high-background-knowledge readers to process the texts passively if these readers had low comprehension skill. One possible explanation for this could be that high-cohesion and coherence texts may contain information that high-knowledge readers are familiar with which may give them a false sense of perceived understanding. This false sense of perceived understanding may be more likely to occur amongst the less skilled high-background-knowledge readers, as readers with higher level of reading skill may be more likely to engage in active processes when reading (O’Reilly & McNamara, 2007).

Active processes refer to readers’ use of prior knowledge and reading strategies to build a coherent situation model of the text read, and constant monitoring of their mental representation to check whether it corresponds to the information described in the text (O’Reilly & McNamara, 2007). Active processes can be compared to a closely related concept of reader-initiated processes described by van den Broek & Helder (2017). The processes involved in active processing and reading-initiated processing are similar. The difference between active processing and reader-initiated processing seems to be that the former is the result of being a skilled reader, whereas the latter is the result of having high standards of coherence. Skilled readers are more likely to engage in active processing, but high standards of coherence are not just reader-dependent (O’Reilly & McNamara, 2007). As previously mentioned, standards of coherence are also influenced by comprehenders’ goals and text features (van den Broek & Helder, 2017). Thus, reader-initiated processing can be seen as active processing that is activated by a complex interaction between the reading situation, characteristics of the reader, and text features.

For skilled readers, improving text cohesion and coherence has a beneficial effect on comprehension as they can actively process the text even if it has a high degree of overlap with their knowledge (O’Reilly & McNamara, 2007). One plausible explanation for this could be that the less skilled readers may have relatively low standards of coherence when reading texts that appear familiar, whereas the high-skilled readers may have relatively high standards of coherence regardless of the text they read (van den Broek & Helder, 2017). Consequently, it may be the case that the more skilled readers are less affected by the effects of text features, such as cohesion and coherence, than the less skilled comprehenders. The naturally arising question that follows is: can high reading skill be acquired through an intervention?

There is evidence to suggest that reading strategy training may improve comprehension of low-background-knowledge participants. McNamara (2004) found that self-explanation reading training, that develops a reading strategy whereby readers are required to explain the meaning of information to themselves while reading, was effective in improving comprehension of low-background-knowledge university students on low-cohesion biology texts. The reason as to why the intervention was ineffective for high-background-knowledge readers might be that low-cohesion texts were already stimulating active processing for both high-skill and low-skill readers (O’Reilly & McNamara, 2007). However, it may be the case that strategy use will also benefit comprehension of highly cohesive and coherent texts for low-skill high-background knowledge readers. This is because self-explanation reading training might engage low-skill high-background-knowledge readers active processes when reading texts that give them a false sense of understanding (McNamara, 2004; O’Reilly & McNamara, 2007).

Overall, the studies presented in this section have investigated the effects of variation in some of the properties of texts, such as cohesion and coherence (e.g., Ozuru et al., 2009), that are thought to predict comprehension (e.g., Kintsch, 1998). Critically, research evidence demonstrates that different informational texts are likely to be processed differently by different readers (e.g., McNamara, 2001; McNamara & Kintsch, 1996; McNamara, et al., 1996; O’Reilly & McNamara, 2007; Ozuru, et al, 2009). Thus, we cannot assume that text revisions aimed at improving comprehension, will benefit all individuals equally. It is likely that to optimise understanding, texts have to be revised with their target readers in mind, and text revisions may have to be coupled with reader-aimed interventions (e.g., McNamara, 2004). However, the studies described in this section considered a relatively small proportion of possible individual differences by text features interactions. Next, I discuss research using mixed-effects models that investigated a wider range of interactions between the effects of individual differences and text features on comprehension (I describe mixed-effects models as an analytic approach in Chapter 4, sections 4.2 and 4.5).

1.6. Mixed-Effects Models of Reading

Francis et al. (2018) assert that most comprehension research has historically tended to focus on following one of the reading comprehension frameworks without much integration with the other frameworks. This has constituted a limiting factor in comprehension research, since it has restricted the number of individual differences by text features interactions that researchers could examine. Consequently, many studies have failed to investigate the potential modulations of the effects of individual differences by variation in the effects of plausible text features in predicting comprehension. To improve understanding of reading comprehension processes, some researchers are trying to make the conceptual modelling of these potential modulations, using mixed-effects models, explicit (e.g., Francis et al., 2018; Kulesz, Francis, Barnes, & Fletcher, 2016).

In a study involving mixed-effects models of reading, Kulesz et al. (2016) assessed word reading ability, reading fluency, vocabulary, background knowledge, WM capacity, and comprehension of 1,190 U.S. middle and high school students on 22 passages. Kulesz et al. were interested in how these passages differed in their average word frequency, sentence length, deep and referential cohesion, genre, and Lexile difficulty. Lexile difficulty is a readability measure that considers average word frequency and sentence length of a given text. The text features of these passages were derived using the Coh-Metrix tool (Graesser, McNamara, Louwerse, & Cai, 2004). However, some of these text features do not have a direct link to reading comprehension theories, or the link has not been made explicit. Thus, before describing the findings of Kulesz et al's. study, I briefly discuss the potential reasons for the inclusion of these text features in their investigation.

Typically, high average word frequency values indicate that a text contains a large proportion of relatively frequently occurring words in the English language. In turn, low average word frequency values indicate that the text contains a relatively large proportion of rarely occurring words. Kulesz et al. (2016) hypothesised average word frequency to influence comprehension because in reading comprehension, knowledge of word meanings is thought to be critical (e.g., Perfetti, 2007; 2010). The lexical quality of rare words and words encountered for the first time is likely to be low as the mental representation of the newly encountered words is likely to be inflexible and imprecise (Perfetti, 2007). Consequently, rare words and words that readers have not seen prior to reading are likely to make meaning integration processes slower and less efficient (Perfetti & Stafura, 2014; Yang et al., 2005). This is likely to have a negative influence on textbase formation, and thereby on comprehension (Kintsch, 1998).

Kulesz et al. (2016) were also interested in the average sentence length of their passages. Longer sentences might make meaning-to-text integration processes more WM resource demanding (Perfetti, 2007; Perfetti & Stafura, 2014). However, the reasoning behind this hypothesis is questionable, since short sentences often do not contain connectives. Increasing cohesion and coherence of texts, features associated with comprehension, frequently involves increasing average sentence length (e.g., O'Reilly & McNamara, 2007; Ozuru et al., 2009). Therefore, increasing sentence length may not necessarily have a detrimental impact on comprehension, especially if increasing length improves text coherence and cohesion (Ozuru et al., 2009). Both average word frequency and sentence length indices, of the Coh-Metrix, were used to calculate the Lexile difficulty level (Schnick & Knickelbine, 2007). This textual measure of readability is claimed to provide a score for the overall passage difficulty. However, Lexile level's predictive utility can be questioned as it ignores the effects of cohesion and coherence on comprehension, and the evidence that sentence length might be spuriously related to comprehension (Ozuru et al., 2009) (for further discussion of the limitations associated with textual measures of readability see Chapter 3, section 3.2).

Cohesion and coherence measures are well-grounded within the text and discourse framework and the CI model (e.g., Kintsch & Rawson, 2007). Although not clearly defined, referential cohesion and deep cohesion constructs used in Kulesz et al's. (2016) study are measures of coherence and cohesion respectively. These constructs are calculated using the indices of the Coh-Metrix tool (Graesser et al., 2014). Referential cohesion is calculated using indices of argument and conceptual overlap in adjacent and all sentences. High referential cohesion scores signal high coherence, whereas low scores indicate low coherence. In turn, deep cohesion refers to a component score of the incidence of causal, temporal, and logical connectives (Crossley, Allen, & McNamara, 2011). High deep cohesion scores show that texts are highly cohesive, whereas low scores indicate the opposite. Overall, texts with high referential and deep cohesion scores should be understood better than texts with low referential and deep cohesion scores (Crossley et al., 2011; Kintsch & Rawson, 2007).

Regarding genre, different types of texts, such as expository and narrative, tend to have different structures. Narratives typically have a gradually developing theme. Specifically, they tend to start with an introduction, followed by a series of episodes consisting of a problem, response, action, and outcome, which lead to an overall conclusion

(Dymock, 2007). In contrast, expository texts typically start with a description of the main idea presented in the text, a list of evidence that support the main idea, a problem or question that is considered, detailed information explaining the main idea, and a comparison between the main idea and another idea (Clark, Jones, & Reutzel, 2013). Expository text is also more likely than narrative to include technical vocabulary and convey information about a specific topic that might be unrelated to everyday experiences (Kulesz et al., 2016). The use of technical vocabulary and topic-specificity might make it harder for some readers to create a coherent mental representation of expository compared to narrative texts. This is because lack of relevant background knowledge may impede meaning integration processes, inference making, and textbase formation (Kintsch, 1998; Perfetti & Stafura, 2014). Thus, narratives might be easier to understand than expository texts.

Kulesz et al. (2016) found that among the middle school students high background knowledge and WM capacity were associated with higher probability of correct comprehension responses. On average, all middle school students were likely to understand passages with high referential cohesion or with high average word frequency scores, but those with relatively high WM capacity were more likely to answer comprehension questions correctly regardless of the characteristics of the passages. Students with high vocabulary knowledge were less influenced by variation in referential cohesion and average word frequency of texts than those with low vocabulary knowledge. Those with high vocabulary knowledge also performed better than those with lower vocabulary knowledge regardless of the average word frequency and referential cohesion of the text read.

For high school students, high reading fluency and vocabulary levels were associated with higher comprehension (Kulesz et al., 2016). In addition, background knowledge interacted with referential cohesion whereby passages with low referential cohesion were found to be more difficult to understand for high school students with low background knowledge, but not for their counterparts with high background knowledge. In fact, high school students with high background knowledge had higher level of understanding of low-cohesion passages versus high-cohesion passages. This finding is consistent with the reverse cohesion effect among high-background-knowledge low-reading-skill readers reported by McNamara et al. (1996), suggesting that the effects of text features may affect different kinds of readers differently.

Critically, the interaction effects of reader characteristics by variation in text features reported by Kulesz et al. (2016) were relatively small and overshadowed by the effects of individual differences, mainly vocabulary and background knowledge. Indeed, the proportion of variance in comprehension accounted for by the interaction effects was between 2% and 7% depending on the model considered. Nevertheless, Kulesz et al. argued that text features by individual differences interactions were particularly important in explaining the effects of text features on comprehension, highlighting the importance of mixed-effects models in comprehension research.

A similar approach to that of Kulesz et al. (2016) was taken by Francis et al. (2018). Francis et al. used mixed-effects models with an added component of time. The time component captured the developmental characteristics of readers, such as age, months of instructions, or sessions of interventions, and other time-related variables, such as decoding skill, measured over a two-year period. Francis et al. analysed data obtained from a cohort of 648 struggling and 865 typical U.S. middle school readers. Struggling readers were identified based on their performance on a reading comprehension task administered as part of a larger study from which the Francis et al. data were obtained (Vaughn et al., 2010).

Francis et al. (2018) used a six-month subset of a two-year longitudinal dataset where students from grades 6 to 8 completed an oral reading fluency (ORF) assessment every six weeks (see Vaughn et al., 2010). The ORF scores used by Francis et al. were calculated using the number of words read correctly per minute during the first 60 seconds of reading of each passage included in the assessment. Specifically, during the six months, the students read 35 different passages differing in text type (narrative vs. expository) and text features as assessed using the Coh-Metrix tool (Graesser et al., 2004). Francis et al's. study was exploratory and the selection of texts features was not related to, or justified by, the theoretical accounts of reading comprehension models (e.g., Kintsch, 1998). Francis et al. were primarily interested in the effects of the same text features as those used in Kulesz et al. (2016), except for deep cohesion which they removed from their model due to high correlation with Lexile difficulty. In terms of individual differences, at the beginning of the study the students were tested on their listening comprehension, word-level decoding, and word reading fluency.

The ORF scores used by Francis et al. (2018) were used as a proxy measure for reading comprehension performance. However, ORF can be considered as a relatively bad proxy of reading comprehension. This is because the correlations between reading comprehension measures and ORF can be low (e.g., Burns et al., 2011). Furthermore, the correlation between comprehension measures and ORF tends to diminish with reading experience, and it might reflect developmental differences between poor and good readers rather than causal connection between ORF and reading comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Paris & Hamilton, 2009). Importantly, individuals might be better at comprehending text than reading aloud accurately (Paris & Hamilton, 2009). Thus, care must be taken when generalising the results of Francis et al's. investigation.

Francis et al. (2018) found that ORF performance improved as a function of grade level and decoding skill, whereas as Lexile difficulty increased students, on average, read more slowly. Students also read expository texts more slowly than narrative texts. Regarding potential modulation effects, Francis et al. only looked at the interaction effects of Lexile difficulty and genre by reader type and grade level. The effects of text type and Lexile difficulty differed between good and poor readers, and the effects of Lexile difficulty differed across grades. Like in Kulesz et al's. (2016) study, the effects of individual differences were much greater in Francis et al's. study than the effects of text features and the interaction effects between the two. However, unlike in Kulesz et al's. study, Francis et al. did not find any evidence for the effects of referential cohesion on reading comprehension. The lack of evidence for effects of referential cohesion may be due to Francis et al's. choice of ORF as a proxy measure of comprehension.

Although Francis et al. (2018) and Kulesz et al. (2016) advocated for mixed-effects model analyses, including many text features by individual differences interactions, both studies included a relatively small number of text features and interaction effects. There are many more text features which could potentially modulate the effects of individual differences on comprehension. It is also important to note that both studies sampled a relatively young sub-group of the population, specifically U.S. middle and high school students. There are dangers of generalising from narrow populations, such as young students, since these narrow populations are often not representative of the general population. This is because students, typically, are relatively homogenous in terms of, for example, age and educational status. In contrast, the general population consists of relatively heterogenous members of the society, of varying educational and socioeconomic backgrounds, as well as of varying ages. Consequently, generalising from narrow populations can lead to an incomplete understanding of the studied phenomenon due to over- and underestimations of the strength of effects (Henrich, Heine, & Norenzayan, 2010). Thus, it is questionable whether the findings of Francis et al. and Kulesz et al. can be reliably applied to the adult UK-based population that I am investigating in my study.

Overall, the use of mixed-effects models in comprehension research suggests that written texts could be tailored to different groups of the population to enhance their understanding of texts read. Critically, in the context of this thesis, improving comprehension of health information could potentially increase treatment compliance benefitting both the patients and the medical practitioners in terms of health and time respectively. In addition, in terms of theory development, by incorporating the different comprehension frameworks and explicitly modelling interactions between text characteristics and individual differences, mixed-effects models can potentially explain more variance in reading comprehension than the established older models of reading comprehension such as the SVR (Gough & Tunmer, 1986). Thus, the mixed-effects models of reading are worth emulating and improving upon in this research project.

1.7. Summary

In this chapter, I focused on a limited range of dominant comprehension theories and made a case for bridging the different comprehension frameworks and models using mixed-effects models of reading (Francis et al., 2018; Kulesz et al., 2016). I suggested that this will explain reading comprehension more fully than the older, but currently leading, theories of comprehension, such as the SVR (Gough & Tunmer, 1986), do. Consequently, I use mixed-effects models as a theoretical framework in this thesis (I elaborate on the use of mixed-effects models as an analytic approach in Chapter 4, sections 4.2 and 4.5). In the subsequent chapter (Chapter 2), I discuss empirical reading comprehension research in the context of individual differences, and I attempt to relate the evidence reported by empirical comprehension studies to the reading comprehension theories discussed in this chapter.

Chapter 2: Literature Review of Empirical Individual Differences Reading Comprehension Research

In Chapter 1, I discussed a limited range of dominant reading comprehension models. These models constitute the theoretical framework within which variables such as vocabulary knowledge have the potential to be predictive of comprehension performance (e.g., Perfetti, 2010). However, past studies have tended to examine the influence of a relatively small number of text features and individual differences on comprehension, at a given time, and have frequently used relatively small samples of participants (Freed et al., 2017). This is an important limitation since the development of theories based on relatively few variables can overstate the importance of the contribution of a variable to reading comprehension. Indeed, the research discussed in this chapter shows that some theorised predictors of reading comprehension may be spuriously related to comprehension when tested in more broadly-based participant samples (Freed et al., 2017). Critically, for the purposes of my investigation, the potential moderating impact of individual differences variables on comprehension is especially interesting. Thus, in this chapter, I discuss the findings of empirical reading comprehension research, and I aim to identify the plausible individual difference predictors of reading comprehension.

2.1. Individual Differences in Reading

In Chapter 1, I explained the different reading comprehension models in terms of variation in the typical population. However, in the typical population the readers may experience comprehension difficulties because, amongst other factors, they might have problems with their working memory (WM), decoding, or inferencing. Alternatively, readers might have relatively small vocabularies, low English language proficiency, or struggle with monitoring their comprehension. In this section, I discuss the relation of these variables with comprehension.

2.1.1. Working Memory

Information processing is thought to be essential to successful comprehension, as retrieval of information from prior knowledge and integration of prior knowledge with crucial fragments of the text, such as the macrostructure (see Chapter 1, section 1.2 for a description), is required for the construction of a situation model of the text read (Kintsch & Rawson, 2007). However, the processes that store and manipulate information, such as remembering words within a sentence, retrieving information from the text, parsing of sentences, activating background knowledge, are thought to require resources (e.g., Nation, 2007; Perfetti et al., 2007). Specifically, from the perspective of many reading comprehension models, such as the Construction-Integration (CI) model (Kintsch, 1988; 1998), WM is argued to constitute the limited resource that enables the processing, including storage and manipulation, of information required for successful comprehension (Kintsch & Rawson, 2007; Perfetti et al. 2007).

As the capacity of WM resources is assumed to be finite (e.g., Perfetti et al., 2007), variation in WM capacity is argued to predict comprehension, whereby high WM capacity is thought to be associated with high comprehension (e.g., Kintsch & Rawson, 2007). In addition, variation in WM capacity is also theorised to be associated with variation in engagement of controlled reader-initiated processes, such as inference-making, which are assumed to be important to comprehension (e.g., van den Broek & Helder, 2017) (Chapter 1, section 1.2). However, there are different accounts of WM, and these accounts can be considered as competing theories (e.g., Baddeley & Hitch, 1974; Kintsch et al., 1999). Thus, before discussing the effects of individual variation in WM on comprehension reported by empirical comprehension studies, I briefly discuss the WM account that this thesis follows.

Baddeley and Hitch's (1974) refined WM model (Baddeley, 2000; Repovš & Baddeley, 2006) is claimed to be the most widely used WM model to date (DeKeyser & Koeth, 2011). Baddeley and Hitch originally proposed that their multi-component model of WM contained three components: the central executive (CE) aided by two storage-capacity-limited subsystems, the phonological loop and the visuospatial sketchpad. The CE was assumed to be an attentional control system of limited processing capacity. The phonological loop was dedicated to storing and maintaining verbal information, and the visuospatial sketchpad was envisaged to store and maintain visual and spatial information. In 2000, Baddeley added a new component to the original model, the episodic buffer.

Baddeley (2000) argued that the episodic buffer is a limited-capacity storage system that can integrate information from the phonological loop, the visuospatial sketchpad and long-term memory. The episodic buffer was theorised to be controlled by the CE. Baddeley argued that, through the CE, an individual could access and reflect on, as well as modify, the information stored in the episodic buffer, effectively increasing the likelihood of accurate recall by inhibiting irrelevant information. Thereby, the role of the CE was refined to include dividing attention between concurrent tasks, switching attention between different tasks, and inhibiting distracting material (Repovš & Baddeley, 2006).

Individual differences in CE capacity, specifically inhibition ability, are theorised to be associated with differences in reading comprehension (e.g., Kendeou et al., 2014). This is because to create a coherent situation model, the crucial information must be maintained in active memory, whereas the redundant information must be inhibited (Kendeou et al., 2014; Kintsch & Rawson, 2007). Indeed, there is evidence to suggest that poor comprehenders have difficulty inhibiting irrelevant information and that good and poor comprehenders vary in their inhibition ability (Cain, 2006). As previously mentioned, individual differences in WM capacity are also thought to predict comprehension (e.g., Kintsch & Rawson, 2007), because the construction of a situation model is assumed to happen in the finite capacity of WM (e.g., Baddeley, 2000).

WM models, such as Baddeley's (2000) multi-component model of WM, assume that the capacity of WM components is limited. However, the reason for the capacity limit of different components of WM is not entirely clear (Cowan, 2010). Those who see the capacity limit as a weakness argue that it would be too biologically expensive for the brain to have no WM capacity limit, and that the capacity limit is necessary to avoid too much interference from competing items held simultaneously in WM (e.g., Lisman & Idiart, 1995; Luck & Vogel, 1998; Usher, Haarmann, Cohen, & Horn, 2001). In contrast, those who see the capacity limit as an advantage suggest that the limit is optimal for the concurrently active concepts held in WM to be linked with each other easily, without leading to misinterpretations (e.g., Dirlam, 1972; MacGregor, 1987).

Overall, it is probable that biological economy limits WM capacity, but that the existing limit may be optimal for information processing (Cowan, 2010). Nonetheless, the individual variation in WM limit is thought to have practical implications in reading comprehension (e.g., Perfetti et al., 2007), especially when the text that is being read is too long or complex to permit the use of processing strategies (e.g., Cowan, 2010). For example, during reading the compehender may be required to simultaneously hold in mind the main idea of the text, the argument made in the previous paragraph, and a notion expressed in the current paragraph (Kintsch & Rawson, 2007). If these elements are not integrated into a single chunk the reader cannot continue to read and understand the message of the text (Cowan, 2010).

Critically, research evidence suggests that individual variation in verbal WM capacity, also known as the capacity of the phonological loop (Repovš & Baddeley, 2006), predicts reading comprehension accuracy (e.g., Cain et al., 2004; Liu, Kemper, & Bovaird, 2009). For example, in Cain et al.'s (2004) investigation, verbal WM capacity explained a significant proportion of the variance over and above vocabulary knowledge and reading ability, as measured using a word reading accuracy assessment. This indicates that the capacity of the phonological loop is likely to be related to comprehension, at least amongst children (Cain et al., 2004), and possibly amongst adults (Liu et al., 2009).

However, the problem with research examining the effects of WM on comprehension is that WM tasks are not pure measures of WM components (e.g., Freed et al., 2017; Van Dyke, Johns, & Kukona, 2014). For example, verbal WM capacity measures, such as the reading span task (Daneman & Carpenter, 1980), where participants may be required to read aloud sentences while remembering the last word of each sentence for later recall, tend to measure many processes. These processes may include general reasoning and verbal ability (e.g., Conway et al., 2005; Freed et al., 2017), and these processes tend to overlap across different tasks (e.g., Van Dyke et al., 2014). Due to the overlap of these processes across different tasks, performance on different individual differences tests is likely to be correlated (Freed et al., 2017). The correlations between different measures of individual abilities, attributed to the systematically shared variance, make it difficult to determine whether the effect of a certain reader characteristic, such as WM capacity, is uniquely predictive of comprehension (Van Dyke et al., 2014).

It might be the case that some individual differences predict reading comprehension because they correlate with other individual differences measures that are unique predictors of comprehension. Indeed, research evidence has emerged indicating that variation in WM capacity may not be directly linked to variation in performance in reading comprehension (Freed et al., 2017; Van Dyke et al., 2014). A relatively small-sample study found that, after partialling out the shared variance between verbal and visuo-spatial WM measures and an IQ measure from their models, vocabulary knowledge was the only significant predictor of reading comprehension (Van Dyke et al., 2014). This indicated that verbal and visuo-spatial WM capacity might be a predictor of reading comprehension due to the shared variance between WM capacity and other cognitive capacity measures such as IQ. Critically, in a relatively large-scale investigation, verbal WM capacity and inhibition were found not to have a direct association with comprehension performance when analyses took into account general reasoning and language experience component measures (Freed et al., 2017).

In Freed et al’s. (2017) study, the general reasoning component consisted of three tests measuring numerical problem solving, whereas the language experience component included a measure of vocabulary and background knowledge, and text exposure. The effects of verbal WM capacity on comprehension were only detectable once general reasoning and language experience were removed from Freed et al’s. models, whereas the measure of inhibition did not predict comprehension performance regardless of the model specification. Overall, the findings of Van Dyke et al. (2014) and Freed et al. suggest that verbal and visuo-spatial WM might not be directly associated with reading comprehension. This is because there might be other individual differences, such as general reasoning and language experience, which directly predict successful reading comprehension but also share variance with WM measures. I discuss the role of language in comprehension, but first I discuss the role of decoding in comprehension.

2.1.2. Decoding and Language

In the Simple View of Reading account (SVR; Tunmer & Chapman, 2012), decoding and linguistic comprehension must work well for successful reading comprehension. Therefore, individual differences in decoding and linguistic comprehension should predict variation in reading comprehension performance. Evidence supporting the theorised role of decoding in successful comprehension comes from observations of dyslexics, individuals who have a reading problem that is often assumed to be caused by phonological impairments (Castles & Friedmann, 2014; Harm & Seidenberg, 1999; Ramus et al., 2003; Stanovich, 1988). The phonological impairments can be reduced to poor ability to segment and manipulate speech sounds (phonological awareness), slow retrieval of speech sounds, or relatively weak performance on verbal WM capacity measures (Castles & Friedmann, 2014; Ramus & Szenkovits, 2008; Wagner & Torgesen, 1987).

Indeed, past research evidence indicates that most dyslexic adults and children underperform, compared to non-dyslexics, on tasks measuring phonological awareness, reading speed, and verbal WM capacity (e.g., Ramus et al., 2003; White et al., 2006). However, not all individuals with dyslexia have phonological impairments and struggle with phonological awareness tasks (e.g., Castles & Coltheart, 1996; Friedmann & Rahamim, 2007). In addition, not all dyslexic readers have problems with comprehension (Hulme & Snowling, 2016). Those dyslexic readers that struggle with comprehension, are likely to do so due to co-occurring (alongside) problems with decoding, and language difficulties (Hulme & Snowling, 2016). For example, relatively low vocabulary knowledge may lead to problems with understanding word meanings which is likely to have a detrimental impact on comprehension (e.g., Hulme, Nash, Gooch, Lervåg, & Snowling, 2015).

Among the non-dyslexic population, high levels of phonological awareness and word decoding were found to be necessary (e.g., Engen & Høien, 2002) but not sufficient (e.g., Cromley et al., 2010; LARRC, 2015) for successful reading comprehension. In other words, difficulties in phonological processing were found to be associated with poor comprehension, but good phonological processing was not necessarily associated with high comprehension. However, it is important to note that the measurement of phonological awareness and decoding is problematic. This is because measures of phonology are unlikely to be measuring phonology alone as they are likely to depend on individuals' literacy levels (Castles & Friedman, 2014). The evidence for this dependence comes from Freed et al's. (2017) study which employed structural equation modelling to identify the relations between the predictor variables, such mediation of one variable by another. In Freed et al's. study, decoding was found to have a direct association with a measure of vocabulary, and vocabulary had a direct association with comprehension performance. Individual variation in decoding did not have a direct effect on comprehension in the presence of language experience, suggesting that decoding's effects on comprehension were reliant on decoding's covariation with vocabulary knowledge. Thus, slow decoding is likely to have a negative effect on comprehension in the absence of other measures, as is it likely to be associated with limitations in vocabulary knowledge (Freed et al., 2017).

Consistent with the SVR account (Tunmer & Chapman, 2012), evidence from empirical developmental studies suggests that inefficient decoding is unlikely to be the only source of reading comprehension difficulties (e.g., Nation, Clarke, Marshall, & Durand, 2004). This is because reading comprehension difficulties are observed among some children who have normal-for-age text reading accuracy (e.g., Cain, Oakhill, & Bryant, 2000; Nation & Snowling, 1997). Individuals who read accurately but have below-average reading comprehension are referred to as poor comprehenders (Nation, 2007). However, poor comprehenders’ difficulties are not limited to reading comprehension. Developmental research evidence indicates that low vocabulary, poor grammar, limited inferencing, and weak linguistic comprehension are associated with poor reading comprehension, even for those who do not struggle with decoding accurately (e.g., Cain & Oakhill, 1999; Catts, Adlof, & Weismer, 2006; Elwér, Keenan, Olson, Byrne, & Samuelsson, 2013; Nation, Cocksey, Taylor, & Bishop, 2010; Oakhill, 1984). This view largely conforms to the SVR model (Tunmer & Chapman, 2012) as vocabulary knowledge, linguistic comprehension, and decoding are claimed to be interdependent and necessary for successful comprehension. The joint influence of vocabulary and decoding on comprehension is also supported by the DVC model (Perfetti, 2010) (Chapter 1, section 1.4). This is because knowledge of word meanings might be especially important in mediating the relationship between inference-making and reading comprehension (Cromley et al., 2010; Freed et al., 2017; Silva & Cain, 2015).

Research evidence discussed in Chapter 1 (sections 1.2 and 1.5) indicated that problems with making inferences can be due to readers adopting low standards of coherence (van den Broek et al., 2011; van den Broek & Helder, 2017) or to lack of relevant background knowledge (McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996; O’Reilly & McNamara, 2007; Ozuru et al., 2009). However, even when the relevant background topic is familiar, poor comprehenders were found to draw fewer inferences from text than typical comprehenders (Cain, Oakhill, Barnes, & Bryant, 2001). This suggests that in the absence of direct prompts to make inferences, poor comprehenders’ standards of coherence might be too low to engage in reader-initiated processes, including inference-making (van den Broek & Helder, 2017).

Developmental research using narrative texts has indicated that vocabulary knowledge might be a critical indirect predictor of reading comprehension through its effects on inference-making (Silva & Cain, 2015). After controlling for age and IQ, vocabulary knowledge was found to be the only significant predictor of individuals' inference-making and literal comprehension. Literal comprehension represents the textbase level understanding, in other words, the meaning of the text as it is expressed by the text (Kintsch & Rawson, 2007). In addition, inference-making exerted an independent influence on reading comprehension, measured as understanding of both literal and situation-model-level comprehension questions (Silva & Cain, 2015). Situation-model-level comprehension questions assess understanding of the situation described by the text (Silva & Cain, 2015). This is thought to require inference-making and integration of the information provided by the text with relevant prior knowledge (Kintsch & Rawson, 2007).

One potential explanation for the mediating role of inference-making in the relationship between vocabulary and reading comprehension could be that those with higher vocabulary knowledge may be more likely than those with lower vocabulary knowledge to make inferences (Silva & Cain, 2015). Specifically, greater knowledge about word meanings may allow readers to activate a wider range of associated concepts from their prior knowledge when reading, and to make more accurate predictions about upcoming words, compared to those with lower vocabulary knowledge (e.g., Freed et al., 2017; Kuperman & Van Dyke, 2013; Silva & Cain, 2015).

The ability to make more accurate predictions about upcoming words, coupled with a relatively high activation of associated concepts, may mean that those with high vocabulary knowledge can execute meaning-to-text integration processes more frequently, and efficiently, than readers with low vocabulary knowledge (Perfetti & Stafura, 2014; Silva & Cain, 2015). In turn, evidence indicates that the more efficient meaning-to-text processing, including inference-making, the less WM resources required (Yang et al., 2005; Yang et al., 2007) (Chapter 1, section 1.4). Thus, readers with high vocabulary knowledge may have more WM resources available for the construction of the situation model, and thereby may be more likely to have higher comprehension than those with lower vocabulary knowledge (e.g., Kintsch & Rawson, 2007).

Silva and Cain (2015) also found that knowledge of grammar, knowledge of different grammatical structures, was a direct predictor of comprehension of narrative texts but not a significant predictor of literal comprehension. Knowledge of grammar includes the ability to decode linguistic markers, such as causal and temporal connectives, that indicate relationships between events. This in turn helps in integration of information between clauses and sentences, helping comprehenders to construct the textbase (Kintsch & Rawson, 2007). The findings of Silva and Cain suggest that poor comprehenders might be limited in vocabulary and grammatical knowledge. However, in the comprehension of informational health-related texts, which are not narratives, knowledge of grammar may have limited effects on comprehension. In the case of informational texts, vocabulary knowledge might be a more important predictor of comprehension. This is because informational texts are likely to require more literal comprehension to understand explicitly stated information, instead of inference generation for the purpose of understanding implicitly stated information that is prerequisite to the construction of an adequate representation of the story of the narrative (Graesser, Singer, & Trabasso, 1994).

Knowledge of word meanings is crucial to inference-making and inferences are critical to comprehension (Perfetti, 2010). Thus, it is likely that knowledge of word meanings is an important reader characteristic in predicting individuals’ comprehension of health-related texts. This is because, as mentioned in Chapter 1 (section 1.6), informational expository texts, such as health-related texts, might be more likely than narrative texts to include technical vocabulary (Kulesz et al., 2016). An interesting question arises with regards to how knowledge of word meanings varies for bilinguals or those who speak English as a second language (ESL), and how this variation influences comprehension of informational expository texts. The SVR and the CI models (Tunmer & Chapman, 2012; Kintsch, 1988) are concerned with monolinguals, but a large dimension of variation in vocabulary knowledge and comprehension is whether individuals are monolingual or bilingual (e.g., Brysbaert, Lagrou, Stevens, 2016; Geva & Farnia, 2012).

2.1.3. Language Background

Reading and reading comprehension processes of monolinguals and bilinguals are thought to be different (e.g., Dijkstra & van Heuven, 2002; Geva & Farnia, 2012; Geva & Ryan, 1993; Upton & Lee-Thompson, 2001; van Heuven, Dijkstra, & Grainger, 1998). It has been theorised that bilingual readers might have a shared lexical system for both of their languages, or that there may be some cross-talk or mental translation required to understand a text in their second language (Dijkstra & van Heuven, 2002; Upton & Lee-Thompson, 2001; van Heuven et al., 1998). However, at the simplest level, the difference between monolinguals and bilinguals is argued to reflect differences in vocabulary and differences in lexical quality (Brysbaert et al., 2016; Geva & Farnia, 2012; Grabe, 2014). I elaborate on, and discuss, the influence of second language (L2) vocabulary and L2 proficiency on comprehension, next.

2.1.3.i. Lexical Entrenchment Hypothesis

According to the lexical entrenchment hypothesis, differences between bilinguals and monolinguals can be attributed to the difference in language exposure (Brysbaert et al., 2016; Diependaele, Lemhöfer, & Brysbaert, 2013). Critically, individuals with less language exposure are likely to have smaller vocabularies than those with more language exposure. In turn, those with smaller vocabularies are likely to have lower quality lexical representations and to be less efficient at word recognition and decoding than those with larger vocabulary (Brysbaert et al., 2016; Perfetti, 2007; Perfetti, 2010). Consequently, vocabulary knowledge is thought to be a relatively good measure of language exposure (Brysbaert et al., 2016; Kuperman & Van Dyke, 2013).

Empirical research evidence underlying the lexical entrenchment hypothesis comes from studies examining the word frequency effect, the finding that common words are processed faster than rare words amongst monolingual and bilingual populations (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Brysbaert et al., 2016; Diependaele et al., 2013; Kuperman & Van Dyke, 2013). Language exposure is argued to be the main factor influencing the word frequency effect, suggesting that exposure variation determines processing times for that language, regardless of whether this language is a first language (L1) or an L2. Indeed, Brysbaert et al. (2016) found that the word frequency effect was larger for individuals with smaller vocabulary than for those with larger vocabulary. However, accounting for vocabulary size in their analysis, the frequency effects were not found to differ between bilinguals and monolinguals. Thus, differences in processing between monolinguals and bilinguals may be explained by the lower amount of exposure to an L2 of bilinguals compared to monolinguals (Brysbaert et al., 2016). Consequently, the distinction between monolinguals, bilinguals, and ESL readers may not matter that much, what matters is the amount of exposure to the target language, meaning language proficiency. This is because the higher the proficiency, the higher the vocabulary knowledge (Brysbaert et al., 2016), and the higher the vocabulary knowledge, the higher the lexical quality and the more efficient lexical access is likely to be (Perfetti, 2007; 2010). Therefore, the more proficient L2 readers are likely to be better comprehenders of L2 texts than the less proficient L2 readers.

Brysbaert et al.’s (2016) findings suggest that to optimise understanding of low-proficiency ESL readers, text writers should avoid using relatively rarely occurring words. This is because the less proficient ESL readers are likely to have relatively small vocabulary knowledge and may not know the relatively rarely occurring words. Without word knowledge their comprehension is likely to be negatively affected by texts containing many rarely occurring words that low-proficiency readers may not know (Perfetti, 2010). Thus, it may be the case that, in order to maximise understanding of ESL readers, in particular in the case of low-proficiency ESL readers, texts have to be written without the use of rare words.

If vocabulary knowledge increases through language exposure (Brysbaert et al., 2016), education is likely to play an influential role in the attainment of vocabulary knowledge. This is because individuals are often exposed to new words in their language through formal education. Consequently, the increase in vocabulary knowledge during childhood can be partially attributed to formal education (e.g., LARRC, 2015). The advantages of education for monolinguals, in terms of reading comprehension in their language, are clear. However, some bilinguals finish formal education in a different language to the language that they are using later, acquiring vocabulary knowledge of a different lexicon. A naturally arising question is how the knowledge of words in one language influences reading comprehension in a different language. I discuss this next.

2.1.3.ii. Interdependence Hypothesis and Mental Translation

The interdependence hypothesis posits that attainment of L2 literacy skills is strongly related to the level of the development of L1 literacy skills (Cummins, 1981). Cummins (2000) theorised that learners who have high level of literacy in their L1 will attain higher literacy in their L2 compared to those who are not as literate in their L1. According to Cummins (2000), academic literacy skills are related to common underlying proficiencies across the languages. Thus, the knowledge that has been acquired in L1 can be relied on when reading in L2, resulting in positive language transfer. For example, depending on their English language proficiency, ESL students who developed age-appropriate conceptual knowledge and academic skills in their L1 may be able to use this knowledge in the context of L2 academic learning, such as comprehending an academic text at school (Cummins, 2000).

Extending the interdependence hypothesis, qualitative research evidence indicates that some bilinguals utilise their L1 to monitor their comprehension and accomplish metalinguistic functions that are thought to improve their L2 comprehension (Upton & Lee-Thompson, 2001). These functions include making observations about the text and reading behaviour and adjusting reading behaviour in response to the reading demands imposed by the text (Upton & Lee-Thompson, 2001). In addition, it is argued that activation of both languages has a beneficial effect on reading comprehension, as L2 readers can use cognitive strategies, such as mental translation, to improve their L2 comprehension (Kern, 1994). Mental translation is the psychological reprocessing of read L2 words, phrases, or sentences into their L1 forms (Kern, 1994). Kern (1994) argued that mental translation is indirectly associated with comprehension through semantic processing and consolidation of meaning which are important to successful comprehension. This is because semantic processing and consolidation of meaning are theorised to be required for textbase formation and construction of the situation model (Kintsch & Rawson, 2007).

In a qualitative study, Kern (1994) found that some of his participants experienced problems reading sentences in their L2. Specifically, Kern's participants self-reported that they lost concentration and had to reread sentences to understand them. Kern argued that difficulties experienced by his participants were likely to be caused by lack of automaticity in word recognition. Lack of automaticity is problematic as inefficient word recognition is argued to take-up additional WM resources that would otherwise be used on meaning integration (Perfetti, 2007). Kern (1994) suggested that mental translation could be used to help to overcome his participants' L2 reading comprehension difficulties by reducing the cognitive load imposed by semantic processing and meaning consolidation. This is because

L1 words might be more efficiently processed in WM than their L2 equivalents (Kern, 1994), leaving more WM resources for textbase formation (Kintsch & Rawson, 2007). However, Kern’s study does not allow us to draw causal claims. It is also important to acknowledge that mental L2 to L1 translation might also involve and consume WM resources (e.g., Tokowicz, Michael, & Kroll, 2004). Thus, Kern’s findings have to be considered in the broader context of relatively high level of uncertainty as to whether mental translation could be effective in reducing cognitive load to overcome individuals’ L2 reading comprehension difficulties.

If ESL readers expend more WM resources than monolinguals on word recognition when reading in L2, their available WM resources for meaning-to-text integration processes are likely to be lower. Consequently, relatively incohesive texts with a high proportion of long sentences might have a more detrimental effect on ESL readers’ comprehension than on comprehension of monolingual readers. This is because texts that are not very cohesive are theorised to be particularly taxing on WM resources, as they might require the readers to make more inferences than highly cohesive texts for meaning-to-text integration processes (Kintsch, 1998; Meyer, 2003). Furthermore, the meaning-to-text integration processes are thought to require more WM resources when reading texts containing a high proportion of longer sentences than those containing a large proportion of relatively shorter sentences (Perfetti, 2007; Perfetti & Stafura, 2014). Thus, it might be the case that in order to maximise reading comprehension, texts have to be written differently for monolingual and ESL individuals. For example, low-proficiency ESL readers might benefit more from texts without long sentences and with a high incidence of connectives to increase cohesion levels than high-proficiency ESL readers and monolinguals. It might be the case that such texts will enable low-proficiency ESL readers to recognise words efficiently, thereby allowing them to focus more WM resources on higher-level processes required for comprehension, such as inference-making.

It is important to mention that the proposed effects of text features on the comprehension of ESL readers may be due to differences in word knowledge rather than the direct effect of less efficient expenditure of WM resources on comprehension (Brysbaert et al., 2016; Freed et al., 2017; Perfetti, 2010). This is because word knowledge is associated with the efficiency of word recognition, and the more efficient is word recognition the more WM resources can be directed at meaning-to-text integration processes. WM expenditure of ESL readers of varying proficiency might be limited, but this limit might simply reflect inefficient lexical access and poor lexical quality (Perfetti, 2007). Nevertheless, the research findings presented in this section suggest that, depending on their English language proficiency, ESL readers and English monolinguals reading English texts may utilise slightly different processes when reading for understanding.

Critically, the use of metalinguistic functions or of mental translation (Kern, 1994), comprehension monitoring (Upton & Lee-Thompson, 2001), and the capacity to use L1 conceptual knowledge and academic skills to comprehend L2 texts (Cummins, 2000), may depend on an individual's metamemory and metacognition. Metamemory is knowledge about the contents of the memory, whereas metacognition refers to the processes used to regulate and monitor memory and cognition (Schraw, 2009). In the context of reading for understanding, both metamemory and metacognition can be seen as part of metacomprehension (Schraw, 2009). Metacomprehension refers to individuals' ability to judge their own comprehension (Dunlosky & Lipko, 2007) to evaluate and control their reading comprehension behaviour (Schraw, 2009). Consequently, the term metacomprehension judgement is used in this thesis to refer to a probabilistic judgement of one's level of comprehension from reading.

Mental translation, metacognition, and metacomprehension can be seen as active processes (O'Reilly & McNamara, 2007) related to standards of coherence (van den Broek &

Helder, 2017). This is because individuals must have the right criteria for achieving adequate comprehension to be motivated to engage in active processing involving mental translation, comprehension monitoring, and other metalinguistic and metacognitive activities in order to improve understanding of a given text. Critically, the theoretical accounts of reading comprehension discussed in Chapter 1 (e.g., Gough & Tunmer, 1986; Kintsch, 1998; Pefetti, 2007) (sections 1.2 to 1.6), do not fully account for what processes are involved in metacomprehension, and how metacomprehension can interact with the developing mental representation of the text and influence comprehension (cf. van den Broek & Helder, 2017). This might be an important omission as metacomprehension is thought to contribute to whether individuals engage in specific reading behaviours, such as rereading, that regulate comprehension breakdowns (e.g., Thiede, Griffin, Wiley, & Anderson, 2010). Thus, it may be the case that adequate comprehension theory should incorporate metacomprehension within the wider comprehension system.

Metacomprehension is important in the context of this project not only from the theoretical perspective of its relation to comprehension, but also from a practical perspective. Specifically, some comprehension research, which I discuss in the next chapter, used metacomprehension judgements as proxies for comprehension (e.g., Crossley, Skalicky, Dascalu, McNamara, & Kyle, 2017). However, it is debatable whether metacomprehension can be equated with comprehension, and it is questionable how closely metacomprehension judgements predict comprehension of health-related texts. For the aforementioned reasons, I discuss metacomprehension next.

2.2. Metacomprehension

It has been theorised that metacomprehension judgements are influenced by cues such as interests, mood, ability to summarise the text, background knowledge, and text coherence (Griffin, Jee, & Wiley, 2009; Thiede et al., 2010). Most of these cues can be argued to be similar to predictors of standards of coherence (van den Broek & Helder, 2017). This is because individuals with interest in reading the text, with relevant background knowledge, who are in the right mood when reading, and who find the text coherent are more likely to engage in active reader-initiated processing required for developing a logical situation model, which includes a metacomprehension component (O'Reilly & McNamara, 2007; van den Broek & Helder, 2017). Specifically, reader-initiated processing, such as generating summaries of texts, might promote self-testing of understanding of read information and self-regulation of comprehension breaks (Thiede & Anderson, 2003; Thiede et al., 2010). In turn, this self-regulation might promote comprehension by aiding readers in building relations among concepts contained in a text, as well as in linking these concepts to prior knowledge in their textbase (Doctorow, Wittrock, & Marks, 1978; Wittrock & Alesandrini, 1990; Kintsch & Rawson, 2007).

Research evidence indicates that the accuracy of metacomprehension judgements varies between individuals (e.g., Griffin et al., 2009). This is important as individuals with higher metacomprehension accuracy are likely to be better at identifying texts which they understood poorly than those with lower metacomprehension accuracy. Indeed, those with higher metacomprehension accuracy were found to be better than those with lower metacomprehension accuracy at selecting which texts they needed to reread to improve their understanding (Thiede, Anderson, & Therriault, 2003). Critically, comprehension of those with higher metacomprehension accuracy was higher than comprehension of those with lower metacomprehension accuracy on reread texts. Thus, strategies that increase metacomprehension accuracy might enable greater self-regulation of reading behaviour, potentially improving comprehension (Thiede & Anderson, 2003; Thiede et al., 2003). For example, asking relatively poor readers to construct concept maps, graphic representations of an underlying structure of the meaning of the text, was found to improve both their metacomprehension accuracy and comprehension performance (Thiede et al., 2010).

Although accurate metacomprehension might have a beneficial effect on comprehension, metacomprehension judgements do not necessarily correlate with comprehension performance (e.g., Maki, 1998). A review of 25 studies found that average relative metacomprehension accuracy, the correlation between participants' metacomprehension judgements and their comprehension scores, was only .27 (Maki, 1998). This indicates that individuals' metacomprehension judgements often diverged from their assessed comprehension levels. Further evidence supporting the divergence of metacomprehension judgements from comprehension performance was found in a related but separate study, where individuals often self-reported that they understood specific parts of the text even though on subsequent comprehension questions they provided incorrect answers (Dunlosky, Rawson, & Middleton, 2005). The findings of Maki (1998) and Dunlosky et al. (2005) are important because they offer a partial explanation as to why individuals might not understand some texts. If metacomprehension accuracy is low, individuals may not be motivated to reread the texts that they find difficult to understand and their comprehension might suffer (cf. Thiede et al., 2003). In some contexts, such as in the context of health-related texts, this lack of rereading, which could potentially increase comprehension, might be associated with health problems (e.g., Baker et al., 2002). Therefore, it is crucial to discuss what individual differences could predict metacomprehension accuracy.

There is a general link between education and metacomprehension behaviours like the use of different reading strategies such as rereading (Zabrucky, Moore, & Agler, 2012). Among younger and older adults, university graduates, compared to non-graduates, were found to be more likely to evaluate and regulate their understanding of problematic information in text through selective rereading (Zabrucky et al., 2012). Higher incidence of rereading of problematic information was in turn associated with more recall of information from the text. If recall is a proxy for or reflects comprehension, then higher levels of education would be expected to be associated with strategies helpful to comprehension. This is because education level could indirectly impact comprehension as more educated adults might have higher relative metacomprehension accuracy compared to less educated adults. Therefore, the more educated adults might be more likely to engage in reading strategies to self-regulate potential comprehension breaks, such as rereading a passage that they did not quite understand (Thiede et al., 2010).

Reading strategy use may also vary in association with differences in individuals' language background and English language proficiency. This is important as the population of the UK contains individuals from different language backgrounds (Office for National Statistics, 2016), and ESL readers and English first language readers may utilise slightly different processes when reading, such as the use of mental translation (Kern, 1994). In addition, English language proficiency levels may also differentiate between reading strategy use amongst ESL readers. Hong-Nam and Page (2014) found that English language learners of varying self-rated English language proficiency differed in the frequency of use of metacomprehension strategies when reading, such as comprehension monitoring, managing, and evaluating. Specifically, the more proficient readers were found to use these strategies more frequently than the less proficient readers. Therefore, intermediate and advanced ESL readers could be more accurate in judging their comprehension of health-related information than beginner level ESL readers, due to potentially more frequent self-testing of understanding of read information and greater self-regulation (Thiede et al., 2010).

Another source of variation in relative metacomprehension accuracy, critical to this project, could be variation in relevant background knowledge. This is because National Health Service (NHS) patients are likely to differ in their levels of health-related background knowledge. This variation could be problematic as social psychology research has shown that individuals with relatively low levels of background knowledge tend to overestimate their performance whereas those with relatively high levels of background knowledge tend to underestimate their performance (e.g., Dunning, Johnson, Ehrlinger, & Kruger, 2003; Kruger & Dunning, 1999). Thus, individuals who have relatively weak comprehension of health-related texts might think that they understand these texts relatively well.

Metacomprehension research has found that higher comprehension performance was associated with greater underestimation of comprehension, but individuals with higher relevant background knowledge reported less underestimation, and were therefore more accurate, than those with lower relevant background knowledge (Griffin et al., 2009). This variation in confidence bias could also explain the finding that on average, higher-background-knowledge readers were found to understand more and were better at predicting their mean comprehension performance on a set of texts than low-background-knowledge readers (Griffin et al., 2009). However, it might be the case that high-background-knowledge readers are more likely to be skilled readers who are more likely to engage in active processing (e.g., O'Reilly & McNamara, 2007) to self-regulate their comprehension (Thiede & Anderson, 2003; Thiede et al., 2010).

It is important to acknowledge that the study of the effects of individual differences on metacomprehension accuracy is difficult. This is because, theoretically, the relatively low metacomprehension accuracy levels might be due to the difficulty of translating text comprehension into metacomprehension and the difficulty of translating text metacomprehension into self-rated judgements (Maki, 1998; Maki, Shields, Wheeler, & Zacchilli, 2005). Furthermore, the relative accuracy of comprehension judgements does not seem to be a reliable measure of metacomprehension as participants' self-reported judgements of metacomprehension tend to fluctuate in magnitude when tested on multiple occasions (Maki et al., 2005).

Although the test-retest reliability of the measurement of relative metacomprehension accuracy is an issue, typically individuals who over-estimate their comprehension tend to reliably over-estimate it, whereas those who under-estimate it, tend to repeatedly under-estimate it, albeit to different levels (Kelemen, Frost, & Weaver, 2000). This error in judgement is thought to occur as individuals usually do not adjust their metacomprehension judgments enough to account for differences in their comprehension performance (Maki et al., 2005). However, an alternative explanation for the low relative metacomprehension accuracy could be that readers make their metacomprehension judgements based on the ease of text processing rather than perceived comprehension (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Dunlosky, Baker, Rawson, & Hertzog, 2006).

The ease of processing hypothesis posits that individuals are more likely to judge their comprehension as higher when the text that they read is perceived to be relatively easy to process (Begg et al., 1989; Dunlosky et al., 2006). If the ease of text processing hypothesis is true, then text features associated with easier text processing, such as cohesion and coherence (Graesser et al., 2011; Hamilton & Oakhill, 2014; Kintsch & Rawson, 2007), could predict metacomprehension judgements (Griffin et al., 2009; Thiede et al., 2010). In turn, metacomprehension judgements could be implemented as comprehension performance prediction as cohesion and coherence are theorised to predict comprehension (e.g., Kintsch, 1998; Kintsch & Rawson, 2007; McNamara, 2001; McNamara & Kintsch, 1996; O'Reilly & McNamara, 2007; Ozuru et al., 2009).

Empirical research evidence indicates that individuals judge their comprehension higher as text coherence and cohesion increase and when reading intact texts versus texts with omitted letters (e.g., Rawson & Dunlosky, 2002). Specifically, in a study involving four experiments, reading comprehension performance predictions were found to be largely based on processing ease, as measured using comprehenders' scores on self-reported ease of processing judgement scale and reading times (Rawson & Dunlosky, 2002). However, metacomprehension and ease of texts processing judgements were found to not always correspond to comprehension performance on texts varying in coherence and cohesion levels (Rawson & Dunlosky, 2002).

One explanation for this could be that coherence can act as a signal for engagement in self-regulatory active processing for the goal of understanding text (e.g., van den Broek & Helder, 2017). However, coherence and cohesion levels might not improve comprehension uniformly for all individuals (e.g., Ozuru et al., 2009). Critically, this explanation suggests that self-rated ease of text processing judgements and metacomprehension judgements may be relatively bad proxies for tested comprehension (cf. Crossley et al., 2017). This is because texts that are perceived to be easy to process due to their text features, might not always be well understood by all individuals. This might be because readers thinking that texts are easy to process might be less motivated to engage in the self-regulatory reading strategies required to repair comprehension breaks (e.g., Thiede & Anderson, 2003; Thiede et al., 2010). Thus, high perceived ease of processing may give readers who arrived at a relatively weak understanding of the text, a false representation of their understanding.

Overall, metacomprehension is thought to be an important predictor of comprehension (e.g., Thiede et al., 2010). However, metacomprehension judgements might not always accurately and reliably predict comprehension (e.g., Rawson & Dunlosky, 2002). This is because self-reported judgements of metacomprehension might be based on other factors, such as ease of text processing and these other factors might not always reflect comprehension (e.g., Rawson & Dunlosky, 2002). Indeed, perception of the ease of processing might have a differential impact on the engagement of active processes, including use of reading strategies, depending on individuals reading skill and standards of coherence (Ozuru et al., 2009; van den Broek & Helder, 2017). Thus, it is questionable whether metacomprehension judgements, due to the low levels of reported relative metacomprehension accuracy (e.g., Maki, 1998), predict comprehension.

2.3. Summary

In this chapter, I have shown that some individual differences that are theorised to predict reading comprehension, such as vocabulary knowledge, are likely to be associated with comprehension, but that others, such as verbal WM capacity, might be spuriously related to tested comprehension (e.g., Freed et al., 2017). I also identified and discussed groups which might struggle to comprehend texts written in English such as low-proficiency ESL readers. In addition, I discussed how metacomprehension and reading strategy use might vary depending on educational and language background, and how a complete comprehension theory may have to consider metacomprehension in the comprehension system. In the next chapter, I build on research described here and discuss empirical research findings relating to predictors of reading comprehension of health-related texts.

Chapter 3: Literature Review of the Effects of Individual Differences and Text Features on Reading Comprehension of Health-Related Texts

There is a scarcity of quantitative research that has considered comprehension of health-related texts (e.g., Liu et al., 2009). In this chapter, I draw on this research to discuss a relatively small range of the plausible effects of variation in individual differences, such as health literacy, age, and language background, on comprehension of health-related texts. In addition, I consider the effects of readability formulae and text features, as there is research evidence to suggest that comprehension of health-related texts may be improved by designing texts that are tailored to readers with different profiles (e.g., Ozuru et al., 2009). Last, I identify that researchers that did investigate the comprehension of health-related texts, in terms of interactions between the effects of individual differences and the effects of text features, have considered relatively few textual features (e.g., Liu et al., 2009). Consequently, there is a theoretical and practical need to examine the ways in which such interactions influence comprehension, motivating the studies discussed later in this thesis.

3.1. Individual Differences and Comprehension of Health-Related Information

The theoretical literature discussed in Chapter 1 and the empirical research findings discussed in Chapter 2 revealed some potential predictors of reading comprehension of health-related texts. According to the models of reading comprehension reviewed in Chapter 1 (sections 1.2 to 1.6), the candidate predictors might include vocabulary knowledge, decoding skills, relevant background knowledge, linguistic comprehension, and working memory (WM) resources (e.g., Kintsch & Rawson, 2007; LARRC, 2015; Perfetti, 2007; 2010; Tunmer & Chapman, 2012). In addition, the empirical research evidence indicated that language exposure might be more important than phonological WM capacity, decoding and phonological awareness, to successful comprehension of adults (Brysbaert et al., 2016; Freed et al., 2017; Kuperman & Van Dyke, 2013; Perfetti, 2007; 2010; Van Dyke et al., 2014) (Chapter 2, section 2.1). There is also suggestive evidence that metacomprehension might be an important predictor of comprehension, through its association with reader-initiated active processes, such as rereading of misunderstood information (Chapter 2, section 2.2). Specifically, it might be the case that the better the metacomprehension, the more efficient the activation of these processes to build a logical situation model (O'Reilly & McNamara, 2007; Thiede & Anderson, 2003; Thiede et al., 2010; van den Broek & Helder, 2017).

However, in the context of health-related texts, no complete account of the dimensions of knowledge or skills that influence the comprehension of health-related information exists. This is problematic as the demand for health services in the UK keeps increasing, and is projected to continue to increase, due to the growth and the ageing of the UK's population (Stoye, 2017). As a consequence of this demand, the National Health Service (NHS) faces increasing cost pressures. These demand and cost pressures are exacerbated, amongst other factors, by lack of understanding of health information. For example, ineffective communication about immunisation and lack of comprehension of information about vaccines has been associated with the recent fall in immunisation rates (Hakim et al., 2019; NHS Digital, 2018a). Critically, a decline in immunisation coverage can rapidly increase the demand on health services as it can lead to outbreaks, such as the measles outbreak that affected a quarter of all European countries, including the UK, in 2017 (World Health Organisation, 2018).

Lack of comprehension is also thought to be one of the reasons why people do not attend health checks and national cancer screening programmes which are reliant on high uptake (e.g., Hall et al., 2016; Harte et al., 2018). In addition to comprehension of health-related texts, low health literacy levels were found to be associated with an increased risk of hospital admissions and missed appointments (Baker et al., 1999; Baker et al., 2002; Miller-Matero, Clark, Brescacin, Dubaybo, & Willens, 2016). This is important as besides the potentially fatal consequences for an individual who has not been diagnosed, missed patient appointments alone are estimated to cost the NHS £216 million per annum (NHS England, 2019). Thus, determining how to write health-related texts to improve comprehension regardless of individuals' health literacy levels could be critical. This is because improving comprehension of health-related texts could potentially reduce some of the pressures faced by the NHS by helping to increase compliance. Improving compliance rates could potentially increase the screening uptake and vaccination rates, while reducing the number of missed appointments, thereby saving NHS money in the long term and improving the health outcomes of patients. Given the importance of health literacy to health outcomes (Baker et al., 2002), I discuss health literacy and its relation to reading comprehension of health-related texts, next.

3.1.1. Health Literacy and Health Knowledge

The influence of health literacy on reading comprehension of health-related information is critical in the context of the current study. There are different types of health literacy, but this thesis considers functional health literacy only. Functional health literacy can be defined as the capacity to obtain, understand and use information to make decisions about one's health (U.S. Department of Health and Human Services, 2010). Functional health literacy is a complex construct and it is argued that there are no measures that sufficiently cover the definition of it (e.g., Berkman, Davis, & McCormack, 2010; Chin et al., 2011). However, the typical proxy measures for assessing functional health literacy are the Rapid Estimate of Adult Literacy in Medicine (REALM; Davis et al., 1993) and the Short Test of Functional Health Literacy in Adults (S-TOFHLA; Baker, Williams, Parker, Gazmararian, & Nurss, 1999; Berkman et al., 2011).

The REALM consists of 125 medical terms which the test-takers are asked to read aloud in order of increasingly difficulty (Davis et al., 1993). Consequently, it can be thought of as a measure of reading and pronunciation ability (e.g., Dumenci, Matsuyama, Kuhn, Perera, & Siminoff, 2013). In contrast, the S-TOFHLA includes 36 reading comprehension items, and five numeracy items. These items were designed to measure comprehension of health information, and the understanding of numerical information in the form of health-related materials, respectively (Baker et al., 1999). However, the typical proxy measures of health literacy, such as the REALM and the S-TOFHLA, are limited. This is because according to the process-knowledge model of health literacy (Chin et al., 2011) functional health literacy encompasses processing speed, WM capacity, vocabulary knowledge, and health knowledge. Since not all of these components are measured by the REALM or the S-TOFHLA, some argue that functional health literacy should be assessed using a combination of different measures (e.g., Chin et al., 2011).

The knowledge, processing, and WM capacity components of the process-knowledge model of health literacy are thought to be critical to successful comprehension (e.g., Kintsch, 1998; Perfetti, 2010). Health knowledge might have an important role in comprehension of health information because relevant background knowledge might moderate the activation of reader-initiated processes, such as inference-making, necessary for meaning integration and the construction of a coherent situation model (McNamara & Kintsch, 1996; van den Broek & Helder, 2017). WM is also considered important because the development of a logical situation model is theorised to require readers to integrate the information read in the text with their background knowledge in the finite capacity of WM (Kintsch & Rawson, 2007; Perfetti & Stafura, 2014; Zwaan, 2016). As mentioned in Chapter 2 (section 2.1.1), the limited capacity of WM resources is thought to have practical implications in reading comprehension when the text read is too long or complex to permit the use of processing strategies for the purpose of meaning-to-text integration (Cowan, 2010).

Overall, there is some empirical evidence indicating that health literacy has a beneficial effect on comprehension of health-related texts (e.g., Chin et al., 2015; 2018). In one relatively recent study, Chin et al. (2018) tested 128 older adults' comprehension of health-related texts. The content, organisation and language of the health-related texts were intuitively revised by three medical experts. These revisions did not follow a specific theoretical framework, but Chin et al. did mention simplifying word and sentence structure, which could influence the efficiency of meaning-to-text integration (e.g., Perfetti, 2007; Perfetti & Stafura, 2014; Yang et al., 2005) (Chapter 1, sections 1.4 and 1.6). Critical to the influence of health literacy on health comprehension, Chin et al. found that health and vocabulary knowledge predicted comprehension question and passage summary accuracy of health-related texts. In addition, health knowledge interacted with passage revisions, whereby individuals with higher health knowledge levels benefitted more from passage revisions than did those with lower health knowledge.

However, it is not clear why health knowledge increased the effectiveness of passage revisions and why the revised passages were found to be understood better than the unrevised passages (Chin et al., 2018). This is because Chin et al. (2018) were unable to specify which aspects of text revision were critical to improving comprehension of their passages. Furthermore, the two passage versions did not differ in average readability scores as assessed using word and sentence length indices. This is important because variation in word and sentence length is thought to influence the efficiency of meaning-to-text integration (e.g., Perfetti & Stafura, 2014; Yang et al., 2005) (Chapter 1, section 1.6). Moreover, as I mention later when I discuss readability (section 3.2), several researchers equated readability with comprehensibility (e.g., Flesch, 1948), but there is research evidence to indicate that readability is not the same as comprehensibility (e.g., Chin et al., 2018). Thus, there is a clear rationale for another study to identify which text features have the greatest effect on comprehension of health-related texts, as increasing document's readability alone might be insufficient in increasing its comprehensibility.

In addition to text revisions, the effects of ageing might also play a part in reading comprehension of health-related texts. This is because the process-knowledge model of health literacy assumes that some aspects of health literacy, such as speed of processing, are prone to degenerate due to ageing, but that others, such as vocabulary and health knowledge, remain intact (Chin et al., 2011; 2015). There is some empirical research evidence to suggest that reading comprehension changes with ageing, and that ageing might be negatively associated with reading comprehension and health literacy (e.g., Alberti & Morris, 2017; Hannon & Daneman, 2009; Kobayashi et al., 2015; 2016). However, research in the field of cognitive psychology indicates that ageing is more closely associated with a decline in the speed of processing, and less strongly with changes in reading comprehension and vocabulary knowledge (e.g., Chin et al., 2011; 2015; Davies, Arnell, Birchenough, Grimmond, & Houlson, 2017; Li et al., 2004; Ramscar, Sun, Hendrix, & Baayen, 2017; Verhaeghen, 2003). The discrepancy in findings could occur because the cognitive science research tends to focus on word and sentence-level reading using speed of processing measures. In contrast, empirical comprehension research often concentrates on paragraph or text level comprehension, which is assessed using comprehension questions. Due to the use of different outcome measures by cognitive psychologists and empirical comprehension researchers, it is questionable whether ageing is associated with reading comprehension. However, considering longitudinal studies of health literacy, systematic reviews, and meta-analyses, the link between ageing and reading comprehension appears to be slightly clearer. I discuss this link next.

3.1.2. Age Effects

Ageing might have varying effects on processes in different domains, but in the context of health literacy measures, the decline in the accuracy of responses to health literacy questions over increasing age has been documented in several relatively recent studies (e.g., Kobayashi et al., 2016). Empirical health literacy and comprehension research findings seem to indicate that ageing might be negatively associated with health literacy and comprehension (e.g., Alberti & Morris, 2017). However, the strength of this association is likely to vary depending on the measure used to assess each construct, and the number of factors controlled for in the analysis (Freed et al., 2017; Kobayashi et al., 2015; 2016). Indeed, assessing health literacy using questions that measure the ability to perform minor calculations, as well as comprehension, requires more than health knowledge alone (Chin et al., 2011). The ability to perform minor calculations is likely to be dependent on processing speed, which the cognitive literature associates with ageing (e.g., Chin et al., 2011). Therefore, some research findings indicating that the risk of limited health literacy increases with age (e.g., Alberti & Morris, 2017) could be attributed to changes in the speed of processing rather than in health knowledge (Chin et al., 2011). In addition, assessing health literacy is a complex task because some researchers use the terms health literacy and comprehension interchangeably, and use comprehension questions to assess health literacy, making the distinction between the two concepts difficult (e.g., Kobayashi et al., 2015).

Nonetheless, there is some relatively robust research evidence indicating that ageing is negatively associated with health literacy or reading comprehension of health information (Kobayashi et a., 2015). In a six-year longitudinal study of 5,256 UK adults aged above 52 years, functional health literacy, as measured using four comprehension questions about a fictional aspirin packet, was found to decline with increasing age among the older age groups of participants (Kobayashi et al., 2015). Indeed, 38.2% of adults who were over 80 experienced decline in their health literacy over the six-year period, compared to 14.8% of adults aged 52-54. Critically, the negative association between ageing and health literacy decline persisted while controlling for variables which assessed cognitive abilities sensitive to ageing, such as processing speed (Kobayashi et al., 2015). However, it is important to note that Kobayashi et al. (2015) used comprehension of a medicine label as a measure of health literacy. This measure did not cover all the relevant components of functional health literacy (Chin et al., 2011) (section 3.1.1), as three items of their instrument measured comprehension alone, whereas one item was designed to assess understanding of numerical information. Nevertheless, it is likely that the observed negative association between ageing and health literacy reported by Kobayashi et al. reflects a negative association between ageing and reading comprehension of health-related information.

The association between functional health literacy and ageing is further complicated because standardised health literacy measures vary. A relatively recent ageing and functional health literacy systematic review and meta-analysis found negative associations between ageing and different health literacy measures (Kobayashi et al., 2016). However, some health literacy measures, such as the S-TOFHLA (Baker et al., 1999) were significantly more closely associated with ageing than others, such as the REALM (Davis et al., 1993). The difference in the strength of age effects on health literacy measures might be because measures such as the S-TOFHLA rely more on reading comprehension, reasoning, and numeracy skills than the REALM (Kobayashi et al., 2016). This is because, as mentioned in section 3.1.1, the S-TOFHLA contains questions assessing comprehension of health-related information, and understanding of numerical information, whereas the REALM does not as it is largely a test of decoding (Dumenci et al., 2013; Kobayashi et al., 2016). Critically, numeracy skills were found to be relatively closely linked to cognitive abilities such as WM resources and processing speed which, in turn, are associated with ageing (Chin et al., 2011). Thus, measures such as the S-TOFHLA are likely to be more reliant on processing speed than measures such as the REALM (Chin et al., 2011; Kobayashi et al., 2016).

Overall, the research evidence presented in this section indicates that ageing is likely to be negatively associated with both health literacy (Kobayashi et al., 2016) and comprehension (Kobayashi et al., 2015) of health-related texts. In addition, health literacy measures have been found to be associated with reading comprehension of health-related information performance (e.g., Chin et al., 2018; Davis et al., 1996). Thus, some research evidence from empirical health literacy research tends to converge around the idea that there is a plausible negative effect of ageing on health literacy and comprehension of health-related texts (Alberti & Morris, 2017; Davis et al., 1996; Kobayashi et al., 2015; 2016). However, the association between ageing and health literacy is unlikely to be uniform for all individuals. This is because some individual differences could moderate its strength.

Empirical research evidence indicates that educational attainment could also predict health literacy and moderate the effects of ageing on health literacy and on comprehension of health-related texts (Alberti & Morris, 2017; Kobayashi et al., 2015). Specifically, older adults with higher educational attainment were found to experience less decline in health literacy and comprehension than those with lower educational attainment (Alberti & Morris, 2017; Kobayashi et al., 2015). These findings suggest that less educated older adults may find health-related texts more difficult to understand than more educated older adults. It may be the case that the cognitive abilities, such as processing speed, of educated individuals decline with ageing (e.g., Kobayashi et al., 2016). However, the educated individuals might be better at compensating for this decline through a more effective use of metacomprehension strategies to self-regulate comprehension breakdowns than the less educated individuals (Griffin et al., 2009; Thiede et al., 2010; Zabrucky et al., 2012) (Chapter 2, section 2.2).

Another possibility is that less educated older adults have been less exposed to reading in general across the lifespan. Consequently, the differences could be also based on overall literacy skills and frequency of word reading (cf. Brysbaert et al., 2016). In addition, the less educated older adults may have lower cognitive baseline scores and vocabulary knowledge in the first place, compared to their more educated counterparts (Brysbaert et al., 2016; Kobayashi et al., 2015). Therefore, differences in comprehension could be explained by differences in word knowledge and lexical quality, because the less educated individuals may have fewer WM resources available for meaning-to-text integration processes than the more educated individuals (Brysbaert et al., 2016; Perfetti & Stafura, 2014; Yang et al., 2007). The more educated individuals may experience an age-related decline in cognitive resources, such as WM and processing speed, but their meaning-to-text integration processes are likely to be more efficient due to having relatively high-quality lexical representations (Perfetti & Stafura, 2014; Yang et al., 2005) (Chapter 1, section 1.4). In turn, the relatively efficient meaning-to-text integration processes of highly educated individuals are likely to require fewer WM resources than those of the less educated individuals who have lower quality of lexical representations (Perfetti & Stafura, 2014). Thus, because more educated individuals have larger vocabulary size and richer lexical representations through more exposure to language, they may offset age-related changes in speed of processing by having more efficient meaning-to-text integration processes than the less educated individuals. Critically, if knowledge of word meanings is important in comprehension of health-related texts, then it is important to determine how individuals with less language exposure, such as English as second language (ESL) readers, understand health-related texts written in English (cf. Brysbaert et al., 2016). Next, I discuss the effects of language status on comprehension of health-related texts written in English.

3.1.3. Language Status

Language status, specifically English language proficiency, is likely to influence comprehension of health-related information written in English. This is because ESL readers are likely to have smaller English language vocabularies than monolingual English readers (Brysbaert et al., 2016). In turn, as mentioned in Chapter 2 (section 2.1.3.i), those with relatively limited vocabularies are likely to have lower quality lexical representations and be less efficient at word recognition and decoding than those with larger vocabularies (Brysbaert et al., 2016; Diependaele et al., 2013; Perfetti, 2007; Perfetti, 2010). Thus, the comprehension of ESL readers with relatively low language exposure is likely to be lower compared to those with more language exposure, such as highly proficient ESL readers and monolinguals (Brysbaert et al., 2016). However, to date, few studies have examined the influence of English language proficiency on comprehension of health-related texts directly.

One empirical study found that acculturation, used as a proxy for English language proficiency, predicted scores on the S-TOFHLA (Baker et al., 1999) and REALM (Davis et al., 1996) measures of health literacy among 73 ESL Spanish immigrants living in Canada (Thomson & Hoffman-Goetz, 2010). The acculturation measure assessed, amongst other factors, individuals' language preferences. However, acculturation might not be a good proxy for English language proficiency and the sample of participants used in Thomson and Hoffman-Goetz's (2010) study was relatively small. Nonetheless, using a larger sample of

ESL individuals and an additional measure of English language proficiency, Todd and Hoffman-Goetz (2011) found that the S-TOFHLA (Baker et al., 1999) scores were predicted by age, education, acculturation, and self-rated English language proficiency of 106 ESL Chinese immigrants living in Canada. In other words, the younger, more educated, more accultured, and more proficient participants, scored higher on the S-TOFHLA, than the older, less educated, less accultured, and less proficient, ESL readers.

As alluded to at the beginning of this section, the effects of acculturation and self-reported English language proficiency on the S-TOFHLA (Baker et al., 1999) scores can be explained using the lexical entrenchment hypothesis (Brysbaert et al., 2016) (Chapter 2, section 2.1.3.i). Specifically, individuals with lowers levels of acculturation and self-reported English language proficiency, are likely to have lower levels of exposure to English than those with higher levels of acculturation and proficiency. This lower exposure to English might in turn translate to lower English language vocabulary (Brysbaert et al., 2016), which is theorised to predict comprehension (e.g., Perfetti 2007; 2010; Tunmer & Chapman, 2012). Thus, those with higher levels of self-reported English language proficiency and acculturation, are likely to have higher health literacy than those with lower levels, as assessed using the S-TOFHLA's comprehension questions.

In addition, the finding that Chinese immigrants who were more educated in Chinese had higher health literacy when tested in English (Todd & Hoffman-Goetz, 2011) can be theoretically accounted for by the interdependence hypothesis (Cummins, 1981) (Chapter 2, section 2.1.3.ii). This is because developing literacy in the second language (L2) is thought to be affected by literacy capabilities in the first language (L1) as literacy skills are theorised to be related to common underlying proficiencies across the languages (Cummins, 1981; 2000). Thus, the more educated individuals can rely on conceptual knowledge and academic skills developed in their L1 when reading in their L2 more than their less educated counterparts (cf.

Cummins, 2000), thereby having a greater likelihood of successfully answering the S-TOFHLA questions.

Critically, the reported negative effects of ageing on health literacy suggest that the effects of ageing on the S-TOFHLA scores are likely to be independent of language use (cf. Kobayashi et al., 2016; Todd & Hoffman-Goetz, 2011). Specifically, like the findings of Kobayashi et al. (2016) (section 3.1.2), the findings of Todd and Hoffman-Goetz (2011) can be partly accounted for by the evidence suggesting that the S-TOFHLA (Baker et al., 1999) is relatively highly correlated with cognitive measures, such as speed of processing, which were found to be associated with age-related slowing (e.g., Chin et al., 2011; Davies et al., 2017). An additional explanation could be that ageing results in a decline in health knowledge for monolingual and ESL readers, that is independent of the decline associated with age-related slowing in processing speed (e.g., Kobayashi et al., 2016).

It is important to mention that the empirical investigation of English language proficiency on health literacy is likely to have highlighted some of the reader characteristics which may influence comprehension of health-related information among ESL readers (Todd & Hoffman-Goetz, 2011). This is because, health literacy is a complex construct that is thought to include, but is not restricted to, understanding of health information (e.g., Berkman et al., 2010; Chin et al., 2011; U.S. Department of Health and Human Services, 2010). This is important as the S-TOFHLA claims to measure health literacy (Baker et al., 1999), but it consists almost exclusively of questions assessing comprehension of health information (sections 3.1.1 and 3.1.2). Thus, the effects of ageing, education, acculturation, and self-reported English language proficiency on the S-TOFHLA are likely to be reflective of the influence of ageing, education, acculturation, and self-reported English language proficiency on comprehension of health-related texts.

However, there are methodological issues with relying on the S-TOFHLA for the measurement of reading comprehension of health-related texts. One of these issues is that the S-TOFHLA is a cloze test (Baker et al., 1999). Cloze tests comprise of sentences where a single word has been deleted and a replacement word must be selected by participants (Cain & Oakhill, 2006). The usage of cloze items is an important issue, because performance in cloze tests may not constitute an optimal measure of comprehension. This is because cloze tests might measure recall of information from the text or word recognition rather than situation-model-level comprehension, which includes inference-making (e.g., Cain & Oakhill, 2006; Keenan, Betjemann, & Olson, 2008). Consequently, there is a clear rationale for research to consider the role of English language proficiency on reading comprehension of health-related texts written in English, using more sensitive measures of comprehension than health literacy tests, such as the S-TOFHLA.

In addition to focusing on individual differences, such as English language proficiency, it is important to ascertain how variation in text features influences comprehension of monolingual English, and ESL, readers. This is because many texts are difficult to understand for individuals from different backgrounds, such as low-proficiency ESL readers. Accurately predicting text difficulty for learners from different backgrounds is important for teachers, writers, and publishers, who want to ensure that appropriate English language texts are accessible to readers of varying English langugage proficiency (Crossley et al., 2008). Critically, researchers (e.g., Flesch, 1948; Crossley et al., 2008) have developed readability formulae to calculate texts' reading difficulty levels in an attempt to provide comprehensible texts to individuals from different backgrounds (Beck, McKeown, Sinatra, & Loxterman, 1991). Readability is thought of as an objective measure of comprehension difficulty (Flesch, 1948), that can match a reader to a text most suitable to them (Kintsch & Vipond, 1979). Readability is theorised as being important as it is often equated with comprehension (e.g., Beck et al., 1991), and readability measures have been found to be relatively frequently used to assess comprehensibility of health-related texts (e.g., Wang et al., 2013).

However, equating readability with comprehension is problematic, as readability formulae are often based on simple indices, such as word and sentence length, that were found to correlate with perceived text difficulty, but may not necessarily predict tested comprehension (e.g., Kauchak & Leroy, 2016; Leroy & Kauchak, 2014; Kintsch & Vipond, 1979). In addition, some of these readability measures have been criticised for not being grounded in comprehension theories, as they do not account for what makes a text easier or more difficult to understand (Kintsch & Vipond, 1979). Consequently, it is difficult to determine exactly what readability formulae are measuring. Next, I discuss the relation of readability formulae to comprehension, and the effects of variation in readability formulae estimates, and texts features, on comprehension of health-related texts.

3.2. Readability

Readability research tends to be relatively descriptive and atheoretical, focusing on identifying text features that may predict comprehension at the expense of occasionally ignoring the theoretical accounts of reading (Kintsch & Vipond, 1979). The models of reading comprehension discussed in the previous chapters highlight the importance of cohesion, coherence, and word knowledge in comprehension (e.g., Brysbaert et al., 2016; Kintsch & Rawson, 2007; Perfetti, 2007; 2010). Thus, theoretically, manipulation of text features associated with these concepts is likely to be related to different levels of comprehension (e.g., Hamilton & Oakhill, 2014). However, empirical researchers and text writers often assess the comprehensibility of texts using textual readability measures (Wang et al., 2013). Critically, these measures are not necessarily grounded within theoretical accounts of reading comprehension and were not intended to directly test reading comprehension (Kintsch & Vipond, 1979).

There are many readability formulae, such as the Simple Measure of Gobbledygook (SMOG; McLaughlin, 1969), which are used to assess readability of health-related texts (Badarudeen & Sabharwal, 2010). However, I concentrate on the the Flesch Reading Ease (FRE; Flesch, 1948) due to its widespread use (e.g., Wang et al., 2013), and for the reason that it is based on indices of word and sentence length, which are representative of numerous readability formulae, including the SMOG (e.g., Badarudeen & Sabharwal, 2010; Kintsch & Vipond, 1979; McLaughlin, 1969). The FRE regression formula was constructed to predict the average reading grade level of a child (Flesch, 1948). It is based on data from 363 passages aimed at children of different grade levels (McCall & Crabbs, 1926). The grade level of each passage was normed using the accuracy of children's answers to comprehension questions about each passage in terms of the Thorndike-McCall Reading Scale (Thorndike & McCall, 1921). As aforementioned, the FRE estimates the comprehension difficulty of a given text, using the weighted factors of word and sentence length,

$$FRES = 206.835 - .846 * (word\ length) - 1.015 * (sentence\ length)$$

(Flesch, 1948, p. 225).

The regression weights applied in the FRE formula were based on recomputed statistical coefficients from Lorge's (1939) study of readability (Flesch, 1948). These weights were intended to standardise the FRE scores, but it is not specified exactly how they were derived, and the standardisation has been imperfect. The FRE indicates text readability in terms of a score that is intended to range from 0 (practically unreadable) to 100 (easy for any literate individual). However, negative scores, as well as those above 100, are possible. A score of 100 was originally intended to predict that a nine-year-old child would be able to answer three-quarters of the comprehension questions that could be asked about the passage that is rated (Flesch, 1948). Scores ranging from 30 to 49 are thought to correspond to a university graduate reading level, considered difficult, whereas scores between 80 to 90 are assumed to correspond to texts that are easy-to-read (Flesch, 1948; Patel, Cherla, Sanghvi, Baredes, & Eloy, 2013).

Although the FRE was originally developed for assessing text difficulty of school texts read by L1 English speaking children, its usage quickly became widespread (Flesch, 1948). For example, within Medline and International Pharmaceutical Abstracts databases of articles published between 2005 and 2008, the FRE was found to be used in 69 out of 155 articles that assessed the readability of health-related texts (Wang et al., 2013). The popularity of the FRE can be attributed to the finding that it predicted relatively well the grade level of children, accounting for 70% of the variance in grade level (Flesch, 1948). Another reason is likely to be that the FRE is relatively easy to calculate, offering a relatively easy solution to increasing readability levels of texts (Wang et al., 2013). This makes the FRE practical, as if it is assumed that the link between readability and comprehension is close, increasing readability of texts should benefit text understanding, including patient understanding of health-related texts.

One reason as to why variation in the FRE score may be related to text comprehensibility is that variation in the FRE score could be indicative of the amount of WM resources required for meaning-to-text integration processes. Specifically, longer sentences are thought to require more WM resources for meaning-to-text integration processes than shorter sentences (Perfetti, 2007; Perfetti & Stafura, 2014) (Chapter 1, section 1.6). Thus, it may be the case that the higher the FRE score, the higher the proportion of short sentences in a text, the less WM resources might be required for meaning integration processes, and the easier it might be to comprehend text.

Critically, the link between the FRE and comprehension has been questioned as the FRE is thought to ignore research grounded within the text and discourse framework (e.g., Crossley et al., 2008; van Dijk & Kintsch, 1983) (Chapter 1, sections 1.5 and 1.6). Specifically, texts high in coherence and cohesion are theorised to reduce the need for reader-initiated processes, such as inferences, making it easier for the reader to link propositions together to construct the textbase, thereby improving text comprehension (e.g., Kintsch, 1988; Kintsch, 1998; McNamara & Kintsch, 1996; van den Broek & Helder, 2017). However, short sentences are thought to frequently omit cohesive markers, whereas increasing cohesion and coherence of texts often involves lengthening texts (e.g., Crossley et al., 2008; Hamilton & Oakhill, 2014; O'Reilly & McNamara, 2007; Ozuru et al., 2009). Consequently, increasing the FRE of texts might be detrimental to comprehension if doing so reduces the cohesion and coherence of the manipulated texts. Indeed, there is some empirical research evidence to suggest that increasing the incidence of short words might be associated with lower comprehension and recall of health-related texts (e.g., Friedman & Hoffman-Goetz, 2007). Thus, it is questionable whether the FRE predicts text comprehensibility as it does not consider theoretically important text features such as cohesion and coherence (Carrell, 1987; Liu, Yates, & Rawl, 2013; Ozuru et al., 2009).

In addition to disregarding theoretically important text features, readability formulae based on word and sentence length, such as the FRE (Flesch, 1948) and the SMOG (McLaughlin, 1969), do not aim to account for reader characteristics, such as language background. Indeed, empirical research findings indicate that readability measures reliant on word and sentence length, such as the FRE, are only moderately correlated to tested comprehension of English L2 learners reading English texts (e.g., Brown, 1998; Brown, Janssen, Trace, & Kozhevnikova, 2012). One of the reasons for this might be because word length does not always correlate with word frequency, while word frequency is likely to be an important influence on the comprehension of ESL readers (Crossley et al., 2008). This is because low-proficiency ESL readers' knowledge of English word meanings is likely to be lower than that of English monolinguals (Brysbaert et al., 2016). Consequently, when reading texts, the low-proficiency ESL readers may encounter more words that they do not know than high-proficiency ESL readers or English monolinguals. In turn, the lexical quality of words encountered for the first time is likely to be low making meaning integration processes slower and less efficient (Perfetti, 2007; Perfetti & Stafura, 2014; Yang et al., 2005). Thus, comprehension of low-proficiency ESL readers may be negatively impacted by a high incidence of relatively rare, in the English language, words (cf. Brysbaert et al., 2016).

As an alternative to traditional measures of readability, such as the FRE (Flesch, 1948), Crossley et al. (2008) developed the Coh-Metrix L2 Readability Index (RDL2), a textual measure of readability for ESL readers. The RDL2 is claimed to consider the theoretical models of reading comprehension, such as the Construction-Integration (CI) model (Kintsch, 1998), and previous readability-related research (e.g., Greenfield, 1999) (Crossley et al., 2008). Due to its focus on theoretical accounts of comprehension, I concentrate on and describe the RDL2's formula, but first I discuss the theoretical justification for the text features that it considers.

The first text feature the RDL2 relies on is content word overlap, the proportion of words that carry the meaning of the sentence that are the same between pairs of sentences (Crossley et al., 2008; Graesser et al., 2011). From a theoretical perspective, content word overlap is thought to be important to successful comprehension, as it is theorised to affect text coherence by manipulating the degree of co-reference of a text (Graesser et al., 2011). Specifically, high degree of conceptual overlap is thought to be indicative of close sense relations between sentences, making it easier for the reader to link propositions together and construct the textbase (Kintsch, 1988). However, it is not entirely clear why Crossley et al.

(2008) chose the index of content word overlap over an index of argument overlap for assessing text’s co-reference. This is because, in their preceding publication, Crossley, Dufty, McCarthy, and McNamara (2007) argued that argument overlap is the most robust measure of co-reference. Furthermore, Kintsch and Rawson (2007) equate co-reference with argument overlap. Thus, assessing co-reference using an index of argument overlap would be more theoretically aligned with the more recent thinking underlying the CI model (see Kintsch & Rawson, 2007) than using an index of content word overlap.

The RDL2 also relies on an index of syntax similarity which measures the similarity in syntactic structure between pairs of sentences in a paragraph (Graesser et al., 2011). It is thought that texts with greater between-sentence uniformity of syntactic structures impose lower cognitive demands on the reader, permitting more WM resources to be devoted to meaning integration processes that maintain coherence (Perfetti, 2007; Perfetti et al., 2007; Perfetti & Stafura, 2014). However, it can be argued that a measure of similarity does not necessarily indicate simplicity. For example, if all sentences were syntactically complex, they would have a high similarity score, but they would be relatively hard to process. This is because readers’ meaning integration processes are thought to require more WM resources when reading syntactically complex sentences compared to less syntactically complex sentences (Crossley et al., 2008; 2011; Graesser et al., 2011; Perfetti et al., 2007). Thus, theoretically, an index of syntactic complexity would be more useful than an index of syntactic similarity.

The last index used by the RDL2 is the average of the word count frequency of each word type for all words in a text, estimated using the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). Critically, the use of the CELEX database is controversial as it contains a relatively small number of words and there are other databases, such as the British National Corpus (BNC; BNC Consortium, 2007), that were found to provide better estimates of word frequencies (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Nevertheless, theoretically, the inclusion of a word frequency index in RDL2 is justified, because frequent words tend to be processed more quickly than infrequent words (e.g., Balota et al., 2004; Brysbaert et al., 2016). Thus, infrequent words are likely to make meaning integration processes slower and less efficient (Perfetti & Stafura, 2014; Yang et al., 2005). In turn, inefficiency of meaning integration processes is thought to have a negative influence on textbase formation, thereby comprehension (Kintsch, 1998). Therefore, texts containing a high proportion of relatively low frequency words might be more difficult to understand than those with a lower proportion of such words.

Another theoretically important reason for using word frequency as an index in the RDL2 formula is that the size of the word frequency effect is thought to reflect levels of language exposure (Crossley, Salsbury, & McNamara, 2012). Research evidence has shown that monolingual and bilingual individuals with smaller vocabulary size, a proxy for language exposure, demonstrate larger frequency effects than those with larger vocabulary size, indicative of more exposure to language (e.g., Brysbaert et al., 2016; Yap, Balota, Sibley, & Ratcliff, 2012). Thus, as previously alluded, the detrimental effect of infrequent words on processing and comprehension may be stronger among individuals with lower levels of language exposure, such as ESL readers (Brysbaert et al., 2016; Perfetti & Stafura, 2014).

Overall, the RDL2 estimates the comprehension difficulty of a given text, using the weighted factors of content word overlap, sentence syntax similarity, and average word frequency,

$$\begin{aligned} RDL2 = {} & -45.032 + 52.230 * (content\ word\ overlap) + 61.306 \\ & * (sentence\ syntax\ similarity) + 22.205 * (CELEX\ frequency) \end{aligned}$$

(Crossley et al., 2008).

The factor weights used in the RDL2 formula were obtained from an exploratory regression model (Crossley et al., 2008). Crossley et al's. (2008) regression model used data obtained from an unpublished doctoral dissertation, where 200 Japanese students' comprehension of 31 academic texts was tested using cloze tests (Greenfield, 1999). In Crossley et al's. model, word overlap, sentence syntax similarity, and word frequency were found to account for 86% of the variance in Japanese students' reading comprehension. The relatively large proportion of variance explained in reading comprehension performance by these three variables led Crossley et al. to apply the coefficients from their model to the RDL2 formula. The reasoning behind this was presumably the belief that Crossley et al's. model would predict understanding of other texts for different individuals.

The RDL2 scores range from 0 (lowest readability) to 30 (highest readability). The RDL2 scores have not be standardised, but there is some empirical evidence suggesting that the RDL2 discriminates between texts aimed at different English language proficiency readers relatively well (e.g., Crossley et al., 2011). However, other than using Greenfield's (1999) data, it has not been tested whether the RDL2 predicts tested comprehension scores. This is an important issue, as the cloze tests used by Greenfield may not constitute an optimal measure of comprehension (section 3.1.3), and the texts that Greenfield's Japanese students read were all academic. Thus, it is questionable whether the RDL2 would predict comprehension in other domains, such as health-related texts.

To date, there is no evidence for the effectiveness of the RDL2 in predicting reading comprehension of health-related texts. Consequently, there is a clear rationale for determining whether the RDL2, a formula that is claimed to be grounded in reading comprehension theories (Crossley et al., 2008), is better at predicting comprehension of health-related texts than the older atheoretic formulae that are currently used to assess readability of health texts, such as the FRE (Wang et al., 2013). However, in addition to research focusing on readability formulae utility in predicting comprehension, there has been significant interest in identifying the text features that may predict comprehension. I briefly discuss this next.

3.2.1. Empirical Research on the Influence of Text Features

Empirical research on the influence of text features is related to research investigating readability formulae (e.g., Crossley, Allen, & McNamara, 2012; Crossley et al., 2007; Crossley et al., 2008; Crossley, McCarthy, Louwerse, & McNamara, 2007; Crossley et al., 2012; Crossley, Salsbury, & McNamara, 2011; Crossley et al., 2017; Graesser et al., 2011). However, the exploratory or descriptive nature of the work of many researchers investigating the effects of text features on comprehension makes it difficult to explain why the effects of some text features should be related to comprehension (e.g., Crossley et al., 2012). Furthermore, due to reliance on outcome measures that do not assess comprehension, such as ratings of perceived text ease, it is difficult to ascertain whether some research findings would generalise to comprehension research (e.g., Leroy & Kauchak, 2014).

It is challenging to apply the findings of some empirical research to explain variation in comprehension, as not all text features research focused on the effects of variation in text features on tested comprehension. Instead, using the Coh-Metrix tool (Graesser et al., 2004), some exploratory investigations examined the differences in text feature dimensions between texts that were intuitively simplified for different level of proficiency ESL readers and texts that were not (e.g., Crossley et al., 2011; Crossley et al., 2012; Crossley et al., 2007; Crossley & McNamara, 2008). Here, intuitive simplification refers to the method of simplifying texts through revisions motivated by text writers' personal beliefs and hunches, guided by experiences of language teaching or material writing (Crossley et al., 2012). Typically, the results showed that texts that were written for different audiences varied in several text feature dimensions. Specifically, simplified texts were found to be more cohesive and less lexically and syntactically sophisticated than texts aimed at more proficient ESL readers (e.g., Crossley et al., 2012).

It is important to note that some of the text feature dimensions included in the exploratory studies of intuitive text simplification (e.g., Crossley et al., 2012), did relate to comprehension models. This meant that the effects of variation in some text features on comprehension, could be explained with the aid of existing reading comprehension theories. For example, the incidence of argument overlap relates to textbase construction (Kintsch, 1998), whereas other dimensions, such as the average word frequency, relate to lexical quality (Perfetti, 2007). Therefore, variation in argument overlap and average word frequency is likely to predict comprehension (Chapter 1, sections 1.2, 1.5, and 1.6). However, the focus on some text features in the text comparisons was not theoretically justified by the intuitive text simplification researchers (e.g., Crossley et al., 2012). Specifically, many text feature dimensions that were included in the studies of intuitive simplification, such as the frequency of occurrence of superordinate words (hypernymy), were often not explicitly related to the theoretical accounts of reading comprehension (e.g., Crossley & McNamara, 2008; Crossley et al., 2012). Consequently, it is difficult to explain why variation in some atheoretically selected text features, such as the frequency of hypernyms, could predict comprehension. Nonetheless, I briefly discuss these text features and attempt to relate them to comprehension theories in Chapter 4 (section 4.4).

Critically, Crossley and McNamara (2008) and Crossley et al. (2007; 2011; 2012) did not test whether the simplified texts were easier to understand. This is because an explicit assumption was made that text simplification led to higher text comprehensibility (e.g., Crossley et al., 2011; 2012). However, it is questionable whether the estimates of effects of text features on estimates of text simplicity are sufficiently predictive of actual understanding (e.g., Kauchak & Leroy, 2016; Leroy & Kauchak, 2014). Indeed, there is empirical research evidence to suggest that perceived text difficulty might not always be related to tested understanding. For example, in a study of 239 adults who read 275 words of varying frequency and length, it was found that self-rated judgements of text difficulty were predicted by word length and word frequency (Leroy & Kauchak, 2014). Specifically, the longer and less frequent words were perceived to be more difficult to understand than the shorter and more frequent words. However, only word frequency predicted correct responses to what the words meant on multiple choice questions, whereby the less frequent words were less likely to be understood than the more frequent words.

Some models of reading comprehension theorise that word frequency predicts comprehension, as word frequency can be seen as a measure of word knowledge (e.g., Perfetti, 2007; Brysbaert et al., 2016). However, theoretically, word length is not seen as equally influential in comprehension. Word length has previously been used as a proxy for word frequency (e.g., Flesch, 1948), but word length and word frequency are not always found to correlate (e.g., Crossley et al., 2008). Thus, it might be the case that variation in some atheoretically selected text features predicts perceived understanding or perceived difficulty but not actual understanding (Kauchak & Leroy, 2016; Leroy & Kauchak, 2014; Rawson & Dunlosky, 2002). This is an important point that motivated the inclusion of a relatively large number of text features, as predictors in analyses of measured comprehension, in the third study of this thesis (Chapter 7).

In addition, the findings of individual differences literature indicate that increasing readability using text features alone might be insufficient for increasing comprehension of health-related texts (e.g., Chin et al., 2018; Kulesz et al., 2016). This is because the effects of different text features are likely to have a differential impact on comprehension of health-related texts depending on readers' characteristics. Therefore, there is a need to consider reading comprehension of health-related texts from a perspective of models that account for an interaction of person and text-level factors simultaneously (e.g., Francis et al., 2018). Next, I discuss empirical research findings of investigations that considered the effects of reader attributes and text features on reading comprehension of health-related texts.

3.3. Comprehension of Health-Related Information: Mixed-Effects Models of Reading

There is some research evidence to suggest that the effects of texts features on comprehension of health-related texts vary between individuals. In one study critical to the concerns of this thesis, Liu et al. (2009) tested 124 U.S. older adults on measures of verbal WM, verbal ability, prior health-related knowledge, and health literacy. Liu et al. also asked their participants to read 16 health-texts and answer yes/no comprehension questions about each one of these texts. These health-related texts varied along two dimensions: reading ease, operationalised by calculating the FRE (Flesch, 1948) score for each text using the Coh-Metrix tool (Graesser et al., 2004), and an index of text coherence. The index of text coherence was theoretically motivated as it was created by averaging standardised scores of argument overlap, conceptual overlap, and stem overlap indices of the Coh-Metrix tool (Graesser et al., 2004). These measures are theorised to contribute to text coherence as they assess the degree of argument repetition, conceptual, and semantic similarity within texts, which are thought to influence the sense relations between sentences (e.g., Graesser et al. 2011; Kintsch, 1988). In turn, it is argued that the closer the sense relations between sentences, the easier it is for the reader to link propositions together and construct the textbase (Kintsch, 1988).

Critically, there were two methodological issues associated with Liu et al's. (2009) study, which are important to point out due to their effects on the interpretation of the reported results. Liu et al. claimed that their measure of verbal ability assessed vocabulary knowledge, but this can be questioned. Liu et al. used The American Version of the National Adult Reading Test (AMNART; Grober & Sliwinski, 1991), which is similar to, but has been designed independently of, the National Adult Reading Test (NART; Nelson, 1982). It is important to note that both the NART and the AMNART were designed for the purpose of estimating premorbid intelligence of adults suspected of suffering from intellectual deterioration. In other words, these tests were designed to estimate premorbid verbal intelligence of adults, not adults' vocabulary levels. Specifically, these tests estimate premorbid verbal intelligence by requiring participants to read aloud 50 irregular English words.

In addition, Liu et al. (2009) cite evidence indicating that AMNART correlates with Weschler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) verbal intelligence (Lastine-Sobecks, Jackson, & Paolo, 1998). However, WAIS-R verbal intelligence is calculated based on scores on six subtests, only one of which is a measure of vocabulary. This is important because other research evidence indicates that correlations between only the vocabulary subtest of WAIS-R and AMNART scores are not that high (e.g., Crawford, Parker, & Besson, 1988; Sharpe & O'Carroll, 1991). Specifically, WAIS-R vocabulary subtest scores were found to be more strongly affected by age-related changes than scores on the AMNART. One reason for this difference may be that the ability to read irregular words and performance on WAIS-R vocabulary subtest are qualitatively different.

Since word knowledge is thought to increase through exposure to words, word knowledge can be partial (Christ, 2011). Vocabulary knowledge can be argued to range from recognizing the word's lexical status to knowing the word's meaning in context and to subsequently knowing the word's meaning independent of context (Christ, 2011). In terms of the continuum of vocabulary knowledge, it could be argued that being able to correctly read aloud words does not demonstrate complete knowledge of that word; instead it may indicate that a speaker heard of that word before and is able to repeat it (Dale, 1965). However, it does not measure readers' contextual knowledge or decontextual, generalisation of meaning of a word to different contexts, knowledge of that word (Christ, 2011). Thus, successful pronunciation may demonstrate only partial knowledge that is acquired after initial exposure to a new word and reliance on one or two contexts for determining word meaning. To measure deeper level understanding of a given word, tasks measuring readers' contextual or decontextual knowledge of a word are required. For example, WAIS-R (Wechsler, 1981) vocabulary subtest requires individuals to define 35 words of increasing difficulty, and scoring is influenced by both precision and the depth of an answer; specifically scores for each definition can range from 0 to 2. Consequently, WAIS-R vocabulary performance is likely to be indicative of more complete knowledge of a word than the AMNART.

In contrast to Liu et al's. (2009) use of a measure designed to estimate premorbid verbal intelligence of adults, Freed et al. (2017) used measures that were designed to assess vocabulary knowledge of individuals. Indeed, Freed et al. did not use a test of reading aloud to measure vocabulary knowledge; instead they used vocabulary tests which aimed to measure both contextual and decontextual knowledge of words (Christ, 2011). Contextual word knowledge was assessed using sections of the Nelson-Denny Reading Test, Form F (Brown, Bennett, & Hanna, 1980) and Form G (Brown et al., 1993). These forms require participants to select a word, from a list of five options, to complete a sentence with a missing last word. Decontextual knowledge was measured using Extended Range Vocabulary and Advanced Vocabulary sections of the Ekstrom battery (Ekstrom, French, Harman, & Dermen, 1976). Both sections of the Ekstrom battery require participants to match a word with a synonym. The ability to select a synonym of a word is considered to demonstrate fuller knowledge of that word than successful pronunciation of it as required by the AMNART. This is because to successfully select the synonym of a word, a reader must have the knowledge of the meaning of that word independent of context. Such knowledge typically develops after multiple exposures to that word, across many different contexts (Christ, 2011).

Consequently, the vocabulary measures used by Freed et al. are likely to provide a more complete assessment of individuals' vocabulary knowledge than the vocabulary measure, AMNART, used by Liu et al.

Instead of measuring vocabulary knowledge, it is likely that Liu et al's. (2009) verbal ability measure examined their participants' decoding. Decoding is theorised to be associated with variation in vocabulary knowledge as vocabulary, decoding, and comprehension, are thought to be interrelated (e.g., Freed et al., 2017; Perfetti, 2010; Tunmer & Chapman, 2012) (Chapter 1, section 1.4; Chapter 2, section 2.1.2). Thus, in the absence of a vocabulary measure in Liu et al's. study, decoding might have capture some of the variance associated with variation in vocabulary (Freed et al., 2017). In addition to using a vocabulary test which was a relatively bad proxy of vocabulary knowledge, Liu et al. did not include health literacy as a predictor in their mixed-effects models as most of their participants scored near ceiling on the S-TOFHLA (Baker et al., 1999) and there was little variability between their scores. Consequently, the lack of reported evidence about the influence of health literacy on comprehension in Liu et al's. (2009) investigation should be interpreted with caution. This is because it may be reflective of an insensitive measure of health literacy rather than lack of effects of variation in health literacy on comprehension of health-related texts.

Liu et al's. (2009) mixed-effects models showed that participants' age predicted comprehension of health-related texts, whereby the older adults were less likely to understand health-related texts than the younger adults. This indicates that it may be the case that the ageing has a detrimental effect not only on the speed of processing measures, but also on comprehension of health-related texts (Chin et al., 2011; Kobayashi et al., 2015; 2016) (section 3.1.2). Verbal ability was also found to predict comprehension, whereby individuals with better verbal ability were more likely to understand health-related texts than those with lower verbal ability. The effect of verbal ability suggests that individuals with higher lexical quality of word form properties, especially orthography and phonology (Chapter 1, section 1.4), are more likely to understand texts than those with lower lexical quality of these word form properties (e.g., Perfetti, 2007; 2010; Brysbaert et al., 2016).

In contrast to the findings of Freed et al. (2017) who found that variation in verbal WM had no direct effect on comprehension (Chapter 2, section 2.1.1), Liu et al. (2009) found that older adults' verbal WM predicted comprehension of health-related texts in the presence of other covariates in their mixed-effects model. One explanation for these conflicting findings may be that Liu et al. did not fully account for vocabulary knowledge in their model. This is important, as Freed et al. found that in the presence of vocabulary knowledge, the effects of decoding and verbal WM did not predict comprehension. However, in the absence of vocabulary knowledge, both the effects of decoding and verbal WM did predict comprehension. Another reason for the discrepancy in findings could be attributed to the difference in the ages of participants of the two studies. Freed et al's. participants were aged between 17 to 29, whereas Liu et al's. ranged from 63 to 95 years of age. It might be the case that the decline in processing capacity, including verbal WM resources, related to ageing (Chin et al., 2011), reduces the ability to bind concepts and propositions to create the textbase, leading to limited comprehension of older adults (Kintsch, 1998; Stine-Morrow, Miller, Gagne, & Hertzog, 2008). Therefore, older adults may be more sensitive to varying WM demands, imposed by meaning-to-text integration processes when reading, than younger adults.

The disagreement in findings between Freed et al. (2017) and Liu et al. (2009) could also be explained by the use of different texts to assess comprehension in the two studies. This is because the texts used by Liu et al. and Freed et al. differed in genre, which is thought to place different demands on readers text processing (e.g., Kulesz et al., 2016; McNamara, Graesser, & Louwerse, 2012). Specifically, Liu et al. used 16 health-related texts whereas

Freed et al. used 10 texts that represented a range of genres, including literature, contemporary fiction, biography, and expositions about science and history. In other words, Liu et al. used informational expository texts, while some of the texts used by Freed et al. were narratives. It is plausible that variation in WM resources is more influential in comprehension of health-related texts compared to narratives. This is because, due to the use of more technical vocabulary (Kulesz et al., 2016; McNamara et al., 2012) (Chapter 1, section 1.6), the efficiency of meaning-to-text integration processes may be negatively affected when reading informational texts (Perfetti & Stafura, 2014), thereby requiring more WM resources (Yang et al., 2005; 2007) (Chapter 2, section 2.1.2). Thus, it is possible that individuals with lower levels of WM resources are more disadvantaged than those with higher levels of WM resources when reading health-related texts compared to when reading narratives. Critically, the conflicting effects of WM between Freed et al. and Liu et al. motivated the inclusion of a verbal WM measure in the third study of this thesis.

In addition to the direct effects of age, verbal ability, and verbal WM, Liu et al. (2009) found evidence for the effects of individual differences by text features interactions on comprehension of health-related texts. Individuals with relatively small verbal WM capacity were found to have greater difficulty understanding texts as the proportion of short words and sentences increased. This interaction effect is theoretically interesting, as increasing the FRE of texts is intended to improve comprehension (Flesch, 1948). However, the effects of WM by FRE interaction indicated that increasing the FRE of health-related texts has a detrimental effect on the comprehension of older adults with relatively small verbal WM capacities. This might be because the frequent usage of short sentences may lower text cohesion (Crossley et al., 2008; Hamilton & Oakhill, 2014; O’Reilly & McNamara, 2007; Ozuru et al., 2009). In turn, lower text cohesion may require individuals to engage in reader-initiated processing, including inference-making, to comprehend the text read (van den Broek & Helder, 2017).

Reader-initiated processes, such as inference-making, are likely to depend on WM resources (Kintsch & Rawson, 2007; van den Broek & Helder, 2017; Yang et al., 2007). Thus, comprehension of health-related texts with a relatively large proportion of short words and sentences may be difficult for those with smaller WM capacities.

Verbal ability and ageing were also found to interact with the effects of text features on older adults' comprehension of health-related texts (Liu et al., 2009). Specifically, Liu et al's. (2009) results indicated that when the proportion of short words and sentences in the text is low, increasing text coherence was likely to lead to comprehension problems, and this effect was estimated to be stronger for individuals with lower verbal ability. However, when the proportion of short words and sentences increased, increasing text coherence was likely to be beneficial to reading comprehension of health information amongst all adults, regardless of their verbal ability. This finding can be explained if one considers that the verbal ability measure is likely to be predictive of comprehension due to its association with variation in vocabulary knowledge rather than the direct effects of decoding on comprehension (Freed et al., 2017).

Since decoding has been found to be associated with variation in vocabulary knowledge (e.g., Freed et al., 2017), the effects of verbal ability can be accounted for by considering the theorised importance, and effects of, vocabulary knowledge on comprehension (e.g., Perfetti, 2010; Perfetti & Stafura, 2014) (Chapter 1, section 1.4; Chapter 2, section 2.1.2). As mentioned in Chapter 2 (section 2.1.2), the meaning-to-text integration processes of high-vocabulary individuals, are likely to be more efficient than those with lower vocabulary levels (e.g., Perfetti & Stafura, 2014). In turn, as mentioned in Chapter 1 (section 1.4), efficient meaning-to-text processing is thought to require fewer WM resources than inefficient meaning-to-text processing (Yang et al., 2005; Yang et al., 2007). Therefore, low-vocabulary readers, may have fewer WM resources available for the construction of the situation model than those with higher vocabulary levels.

The variation in available WM resources between high- and low-vocabulary readers, is important in the context of Liu et al's. (2009) findings, as longer sentences might make meaning-to-text integration processes, including inference-making, more WM resource demanding (Cowan, 2010; Perfetti, 2007; Perfetti & Stafura, 2014). Consequently, increasing coherence of health-related texts with low proportion of short words and sentences might impose additional WM-resource demands to understand those texts, as the process of increasing coherence may lengthen the texts further. For example, an increase in argument overlap may lengthen sentences (e.g., O'Reilly & McNamara, 2007; Ozuru et al., 2009). In turn, meaning-to-text integration processes of readers with low vocabulary might be more negatively affected by the relatively long words and sentences than those with higher vocabulary, because low-vocabulary readers are likely to expend more WM resources on meaning-to-text integrations processes (Perfetti, 2007; 2010; Perfetti & Stafura, 2014). However, the differences in the vocabulary knowledge might not matter as much when the proportion of short words and sentences increases as shorter sentences might make meaning-to-text integration processes less cognitively demanding for low-vocabulary readers than longer sentences (Perfetti, 2007; Perfetti & Stafura, 2014). Thus, it might be the case that readers of varying vocabulary levels can benefit from the increase in text coherence, when FRE is high, as they may need to engage in fewer reader-initiated processes, such as inference-making, to construct an integrated textbase (van den Broek & Helder, 2017).

Liu et al. (2009) also found evidence for the effects of a three-way interaction of age, FRE, and text coherence, on reading comprehension of health-related information. They found that when the proportion of short words and sentences is high, increasing text coherence was found to benefit comprehension of all adults, but especially those below the age of 77 years. However, increasing text coherence when the proportion of short words and sentences is low, was found to lead to comprehension problems for adults, but more so for those over the age of 77 years. This could be because when the proportion of long words and sentences is relatively high, increasing text coherence may further lengthen the sentences, thereby increasing the demand for WM resources for meaning-to-text integration processes (Perfetti & Stafura, 2014; Yang et al., 2005). This is likely to be problematic for older adults' meaning-to-text integration processes, as empirical research evidence suggests that ageing is negatively related with measures of processing speed and WM capacity (e.g., Hannon & Daneman, 2009; Kobayashi et al., 2015; Li et al., 2004). Thus, comprehension of older adults of texts with a high proportion of longer sentences and high levels of coherence is likely to be lower than that of younger adults who have more resources for meaning-to-text integration processes. In contrast, when the proportion of short words and sentences is relatively high, increasing text coherence may decrease the WM demands placed on older adults' meaning-to-text integration processes by requiring fewer inferences to bridge relatively short sentences. Thus, their ability to process health-texts at lexical and syntactic levels is likely to improve, resulting in a coherent mental representation of the text read (Liu et al., 2009).

Overall, the limited empirical research findings within the domain of comprehension of health-related texts (e.g., Chin et al., 2018; Liu et al., 2009) demonstrate the need for studying comprehension of health-related information using mixed-effects models of reading (Francis et al., 2018). This is because they show that the effects of individual differences, such as age, and the effects of features of the text, such as word and sentence length, may interact with each other to strengthen or weaken the influence of some text features on comprehension of health-related information (Liu et al., 2009). Liu et al's. (2009) and Chin et al's. (2018) research demonstrates that comprehension of health-related texts is a shared outcome of text features, such as text coherence and cohesion, and reader characteristics, such as age, WM resources, vocabulary, and health literacy. Consequently, there is some evidence to suggest that comprehension of health-related information may be improved by designing texts that are tailored to readers with differences profiles. For example, by writing high-coherence texts with high proportion of short words and sentences to reduce the processing demands of meaning-to-text integration processes for older readers (Liu et al., 2009; Perfetti & Stafura, 2014).

3.4. Summary

In this chapter, I have shown that some individual differences and texts features are likely to influence comprehension of health-related information both in isolation and through interactions with each other. However, given the limited range of individual differences by text features interactions investigated by previous research (e.g., McNamara & Kintsch, 1996; Liu et al., 2009; Chin et al., 2015, 2018), there is a clear theoretical need to ascertain how comprehension of differently written health-related texts varies for different individuals depending on the characteristics of those texts. Critically, given the applied nature of this project it is also necessary to review the guidelines currently provided to health-related information writers, and to investigate the evidence base underpinning those guidelines to establish the study rationale and specify the research questions. In the subsequent chapter (Chapter 4), I draw on the NHS's guidelines (e.g., NHS England, 2018a; 2018b) and the literature reviewed (Chapters 1, 2, and 3) to specify the research gap that this work aims to fill.

Chapter 4: Overall Research Design and Rationale

This chapter draws on the National Health Service's (NHS's) guidelines (e.g., NHS England, 2018a; 2018b) and the literature reviewed in Chapters 1, 2, and 3 to specify the research gap that this work aimed to fill and the research questions that it attempted to answer. It also outlines the rationale for, and introduces, the three studies that constitute this thesis. Last, the chapter ends with a justification for the use of the Bayesian inferential framework in the analyses of the data obtained from the three investigations.

4.1. NHS Guidelines (December, 2018)

In the UK, the NHS is responsible for producing health-related texts that are distributed to its patients. The health-related texts are written by writers who follow the NHS's guidelines for producing these texts (e.g., NHS England, 2018a). The guidelines are important in the production of health information as they are assumed to improve comprehensibility of health-related texts. However, as I discuss next, these guidelines are vague, and it is difficult to consistently produce highly understandable texts based on guidelines which are not specific and whose utility has not been tested.

Although, during my PhD, the NHS's guidelines for writing health-related information kept changing, they did not become any more precise than when I started. At the beginning of my PhD, the NHS's Brand Identity guidelines (NHS, 2015) specified the preference for the use of short sentences, active tense, and "plain language". These guidelines were superseded by the NHS England's (2018a) Information Standard, the NHS Identity Guidelines (NHS England, 2018b), and the Accessible Information Standard (Marsay, 2017a; 2017b). The Information Standard (NHS England, 2018a) requires that writers of health-related texts ask for feedback from the end users when producing health-related texts, peer review their health-related texts, and ensure accessibility of health-related texts for the intended end users. According to internal procedural documents of NHS Trusts, such as Blackpool Teaching Hospitals NHS Foundation Trust, NHS Trusts comply with these requirements by asking a minimum of two members of a patient reader panel for feedback when producing health-related texts (e.g., Burrow & Forrest, 2015).

Patient reader panels consist of members of the public who volunteer to review health-related information prior to it being released to ensure that it is easy to understand for the target end users. Although, patient reader panel members' perceived comprehension of health information is tested, their actual comprehension of these documents is not assessed.

Thus, it is implicitly assumed that perceived comprehension of health-related documents is the same as actual comprehension. It is also assumed that the comprehensibility judgements of reader panel members on what makes texts understandable apply to individuals in the Trust patient population in general. However, we do not know if reader panel members, who are often elderly, highly literate, monolingual English readers, are able to determine what makes health-related texts comprehensible to other individuals with different characteristics, such as education level and health knowledge (cf. Griffin et al., 2009; Zabrucky et al., 2012). Critically, the utility of metacomprehension judgements of health-related texts in predicting comprehension can be questioned as the evidence reported in Chapter 2 (section 2.2) indicates that metacomprehension ratings may be relatively inaccurate proxies for tested comprehension (e.g., Maki, 1998). Consequently, there is a practical need to investigate whether self-reported judgements of comprehension predict actual comprehension of health information.

Regarding the production of health-related texts, the Accessible Information Standard (Marsay, 2017a) recommends writing health-related information in the format referred to as Easy-Read. Easy-Read makes use of “straightforward words” and is intended to increase comprehension levels of health-related texts among individuals with disabilities, impairments, or sensory loss (Marsay, 2017b). In addition, the Information Standard (NHS England, 2018a), which targets the typically developed population, states that each health-related text should be written in “plain language”, free from grammar errors and jargon, with medical terms explained where necessary. However, it is not specified what constitutes “plain language” and “straightforward words”, making the guidelines open to interpretation and thereby difficult to follow. The NHS Identity Guidelines (NHS England, 2018b) contain similar recommendations to those made in the Information Standard. The Identity Guidelines advocate for the use of simple words, and the avoidance of jargon, acronyms, and unnecessary technical language. However, no evidence base is cited to support the effectiveness of these guidelines, and the guidelines are difficult to implement as the definition of simple words, jargon, and technical language can be expected to vary between people.

Interestingly, the NHS Identity Guidelines (NHS England, 2018b) direct readers to guidelines on the Plain English Campaign's (2018) website on how to write clear and concise public information. The Plain English Campaign (2018) is an influential commercial editing and training firm which advocates for clear and concise written communication. It has worked with various UK government departments and private organisations to improve their communication by editing, clarifying, and rewriting documents. The recommendations provided in the Accessible Information Standard (Marsay, 2017a), the Information Standard (NHS England, 2018a), and the NHS Identity Guidelines (NHS England, 2018b) align with the recommendations of the Plain English Campaign (2018). Thus, a natural conclusion based on this is that NHS England supports and advocates the Plain English Campaign's guidelines for its information writers.

The Plain English Campaign's (2018) guidelines on "How to write in plain English" specify the need for: keeping sentences short; preferring the usage of shorter words, active verbs rather than passive verbs; using simple words; avoiding nominalisations. However, no theoretical or empirical research evidence is cited for these recommendations on NHS England's and Plain English Campaign's websites. Thus, it seems that the aforementioned guidelines were not tested by empirical studies and were not based on empirical evidence or reading comprehension theories. This is concerning and warrants an investigation to establish the effectiveness of the guidelines advocated by the NHS England (2018a; 2018b), NHS Trusts (e.g., Burrow & Forrest, 2015), and the Plain English Campaign (2018), in improving comprehensibility of health-related texts.

4.2. Research Gap

According to the review of the guidelines in the previous section, the NHS information writing guidelines (e.g., NHS England, 2018a; 2018b) do not appear to have been tested and have not been written based on empirical or theoretical research evidence. The search for a potential evidence base underlying these writing guidelines is complicated by the lack of explanation for some key recommendations, such as the preference for straightforward words and plain language. Nonetheless, an assumption can be made with regard to what some of these terms may be referring to. For example, plain language and straightforward words are likely to correspond to words that frequently appear in the English language, meaning high frequency words. Word frequency is theorised to affect comprehension, as the lexical quality of high frequency words is likely to be higher than the lexical quality of low frequency words (Perfetti, 2007), especially amongst individuals with lower levels of English literacy (Brysbaert et al., 2016). This is important because in reading comprehension knowledge of word meanings is critical (Perfetti, 2007; 2010), as words that readers have not seen prior to reading are likely to make meaning integration processes slower and less efficient (Perfetti & Stafura, 2014; Yang et al., 2005) (Chapter 1, section 1.6). Thus, there is some evidence to suggest that using straightforward words, assuming this refers to high frequency words, over less-straightforward words, assuming this means low frequency words, could improve comprehensibility of health-related texts.

The recommendation for short words and sentences over long words and sentences (Plain English Campaign, 2018) also relates to some reading comprehension (e.g., Perfetti, 2007; Perfetti & Stafura, 2014) and readability (e.g., Flesch, 1948) research. From the perspective of some reading comprehension models, it is thought that longer sentences might make meaning-to-text integration processes more WM resource demanding (Perfetti, 2007; Perfetti & Stafura, 2014). This theory could possibly account for why variation in the Flesch

Reading Ease (FRE; Flesch, 1948) was initially found to account for a lot of the variance in comprehensibility of school texts. However, relatively recent empirical findings suggest that shortening sentences could make texts less coherent and cohesive (e.g., O'Reilly & McNamara, 2007; Ozuru et al., 2009) (Chapter 1, section 1.6). Critically, cohesion and coherence are theorised as being important to comprehension (e.g., Kintsch, 1988; Kintsch & Rawson, 2007), as texts that are highly cohesive and coherent are predicted to require fewer reader-initiated processes to build a logical situation model (e.g., Hamilton & Oakhill, 2014; van den Broek & Helder, 2017). Thus, there is a potential discrepancy in the expectations placed by the guidelines and some theoretical accounts of comprehension (e.g., Kintsch & Rawson, 2007), as the guidelines used by the NHS (e.g., NHS England, 2018a; 2018b; Plain English Campaign, 2018) ignore text coherence and cohesion.

From the theoretical perspective, it is not clear whether the use of shorter sentences benefits or hinders comprehension as it could make texts less coherent and cohesive (e.g., Ozuru et al., 2009). However, empirical research evidence from mixed-effects models of reading, within the domain of health-related texts, indicates that the effects of word and sentence length on comprehension are likely to be reader dependent (e.g., Liu et al., 2009). In other words, increasing the proportion of short words and sentences in health-related texts is unlikely to have a uniform effect on comprehension of all individuals (Chapter 3, section, 3.3). For example, increasing the proportion of short words and sentences could have a detrimental effect on comprehension of older adults with relatively low WM resources, but it could be beneficial for older adults with relatively high WM resources (Liu et al., 2009). Overall, the different and relatively uncertain, but potentially complementary, predictions of theoretical (e.g., Perfetti & Stafura, 2014) and empirical accounts (e.g., Liu et al., 2009) demonstrate that there is a need for a robust investigation into the effects of sentence and word length on the comprehension of texts by different kinds of readers.

Although some of NHS's recommendations, such as the preference for straightforward words (Marsay, 2017a), can be relatively easily related to reading comprehension theories, others, such as the preference for active voice and the avoidance of nominalisation (e.g., Plain English Campaign, 2018), cannot. Some empirical studies that used or evaluated the Coh-Metrix tool (Graesser et al., 2004), argued that texts with high frequency of passive voice forms are more difficult to process than those with active voice forms (e.g., Dowell, Graesser, & Cai, 2016; Graesser et al., 2011). This may be because passive voice is usually less frequent, and thus it might affect comprehension, especially of the relatively less literate individuals, such as those with relatively low educational attainment (e.g., Street, 2020; Street & Dąbrowska, 2014). However, exploratory investigations which used Coh-Metrix to calculate the incidence of passive voice forms made no attempt to explain why the incidence of passive voice forms should influence comprehension (Crossley et al., 2007; 2008; 2011).

In addition, many exploratory investigations looked at differences between intuitively simplified and not simplified texts aimed at English as Second Language (ESL) readers rather than at the effects of passive voice on tested comprehension (Crossley et al., 2007; 2008; 2011). The same exploratory investigations included the incidence of gerunds in their analyses, a measure of nominalisation of words using the "-ing" form which may relate to the recommendations of the Plain English Campaign (2018). Crossley et al. (2007; 2008; 2011) argued that the more difficult texts were likely to contain a higher incidence of gerunds, but they did not explain why inclusion of gerunds should make texts difficult to understand. Furthermore, since understanding of intuitively simplified texts was not empirically tested (Crossley et al., 2007; 2008; 2011), it may be the case that the incidence of passive voice forms and gerunds is associated with perceived comprehension, but not tested comprehension (Kauchak & Leroy, 2016).

Considering the lack of an evidence base for the effectiveness of some of the text features guidelines used by the NHS to improve comprehensibility of health texts, there is a pressing need for a study which would evaluate the effectiveness of the guidelines in improving comprehension. Critically, to date, there has been a relatively small number of quantitative studies investigating the predictors of reading comprehension of health-related information. Consequently, health organisations and charities in English-speaking countries have no specific guidelines on how to write understandable health-related texts for different groups of individuals, such as ESL readers, within their populations. Previous studies (e.g., Liu et al., 2009; Chin et al., 2015; 2018) have tended to focus on relatively small samples of specific sub-groups of the United States' monolingual population, such as older adults. Accordingly, some of the findings may not generalise to the diverse population of the UK which includes many ESL speakers (Office for National Statistics, 2016). Thus, it is of theoretical and practical interest to consider the effects of individual differences on comprehension of health-related texts, spanning a participant sample including monolingual English, and ESL, readers.

Theoretically, as mentioned in Chapter 1 (section 1.6), the research findings presented in the reviewed literature suggest that reading comprehension is influenced by the effects of individual differences and text features, and that these effects interact with each other. Therefore, it is vital to study comprehension from the perspective of these interactions since understanding of a given text may vary across different kinds of readers (e.g., Francis et al., 2018; Kulesz et al., 2016; McNamara & Kintsch, 1996; Liu et al., 2009; Chin et al., 2015, 2018). Importantly, the studies conducted to date have focused on a relatively small number of individual differences by text features interactions (e.g., Liu et al., 2009). Consequently, there is motivation for an investigation to consider more text and person-level variables to provide a more complete picture of comprehension of health-related texts. Furthermore, it can be argued that even if the NHS were not concerned about the practical utility of their guidelines for presentation of health information, the work of the thesis would be making a substantial theoretical contribution. This is because there has been limited research on the ways that the effects of person attributes and text properties interact to influence comprehension (cf. Francis et al., 2018; Kulesz et al., 2016).

From the methodological perspective, there are problems with the measures of comprehension that that have been used in previous investigations in this area. Experimental measures of reading comprehension used in empirical studies often involved multiple-choice or true/false questions, or weak proxies of reading comprehension, such as the oral reading task (e.g., Francis et al., 2018; Liu et al., 2009). These tasks are argued to be severely limited in their assessment of skills underlying successful comprehension, such as inference making, and are less sensitive in assessing comprehension than open-ended questions (Cain & Oakhill, 2006). Thus, there is a need for verification of past research findings (e.g., Liu et al., 2009) with more robust measures of comprehension, such as the use of open-ended questions to probe understanding at the level of the situation model.

Critically, the standardised measures of other abilities related to comprehension of health-related texts, such as health literacy, are not pure measures of health literacy and tend to measure different aspects of health literacy (Chin et al., 2011; Kobayashi et al., 2015; 2016) (also refer to Chapter 3 for a discussion). In addition, there is some empirical research evidence to indicate that the S-TOFHLA (Baker et al., 1999) measure of health literacy has a tendency to show a prominent ceiling effect amongst individuals with different profiles (e.g., Liu et al., 2009; Morrison, Schapira, Hoffman, & Brousseau, 2014). For example, regardless of differences in age and WM resources (Liu et al., 2009), and regardless of educational attainment (Morrison et al., 2014), individuals were found to score high on the S-TOFHLA and their scores showed very little variability.

The ceiling effect is a major measurement limitation that decreases the likelihood that a measure, such as the S-TOFHLA, adequately assesses the intended construct (Taylor, 2010). This is because the ceiling effect reduces the variability in scores between different participants to the point at which the variance in the S-TOFHLA scores may be unmeasurable. In turn, the inhibited variance limits the sensitivity of models to study the effects of health literacy on comprehension of health-related texts (e.g., Liu et al., 2009). Overall, the S-TOFHLA may not be as sensitive as Baker et al. (1999) intended it to be. Consequently, the effects of health literacy on comprehension of health-related texts should be studied using several measures of health literacy to minimise the potential ceiling effects. Furthermore, there is a practical need for the development of a new measure of health literacy that assesses the construct of health literacy adequately and avoids the ceiling effect. This practical need motivated the development of a new health literacy measure in the third study of this thesis (Chapter 7).

In addition to health literacy measures, there is empirical research evidence to suggest that the estimates of word frequencies provided by the Coh-Metrix tool (Graesser et al., 2004) may also contain a substantial amount of measurement error (e.g., van Heuven et al., 2014). As mentioned in Chapter 3 (section 3.2), the Coh-Metrix tool estimates average word frequency of texts using the CELEX word database (Baayen et al., 1995). However, the CELEX database contains a relatively small number of words. As a result, it is not a very sensitive measure of word frequencies compared to word frequency values estimated using larger databases such as the British National Corpus (BNC; BNC Consortium, 2007; van Heuven et al., 2014). Consequently, the effects of word frequency reported by empirical researchers who used the Coh-Metrix tool (Graesser et al., 2004) may not be as accurate as they would be given a different reference corpus. This is problematic as this tool is relatively frequently used in readability (e.g., Crossley et al., 2008) and comprehension (e.g., Liu et al.,

2009) research. Critically, in health-related texts research, this measurement error could manifest itself as artificially low frequency values given to relatively rare medical words. Thus, it is of theoretical and methodological interest to assess whether the word frequency effects on comprehension of health-related texts using the CELEX are similar to those produced by more recent and larger corpora of words, such as the BNC.

Measurement error is connected to the observation of spurious effects (Type I errors). This is because as measurement error increases, sensitivity, or capacity to detect real effects that are there, as well as precision, or capacity to not detect spurious effects, decreases (Gelman & Carlin, 2014). The Type I error rate can also be affected by the use of different populations, sample sizes, and analytic approaches (Gelman & Carlin, 2014; Freed et al., 2017). Importantly, findings based on small samples are more likely to be biased and not replicable than those based on larger samples of participants (Gelman & Carlin, 2014). This is problematic for some studies that investigated the effects of individual differences and text features on comprehension. For example, some theory-grounded early research on the reverse cohesion effect (Chapter 1, section 1.5) used relatively small samples of participants, such as 56 10 to 15-year-olds (McNamara et al., 1996) and 80 undergraduate students (McNamara, 2001). Thus, the effects reported in some studies of comprehension (e.g., McNamara et al., 1996), are likely to be relatively uncertain.

Critically, the impact of limitations due to measurement or sampling are potentially amplified by the use of some analytic methods, such as Analysis of Variance (ANOVA). This is because such tests are unable to take into account both error variance due to random differences between sampled participants and error variance due to random differences between sampled texts, inflating the Type I error rate by failing thereby to account for critical sources of uncertainty in the data (Gelman, 2015). Unlike more traditional approaches, such as ANOVAs, mixed-effects models allow the researcher to account for random variation between participants and test items, thereby accounting for additional uncertainty in the data and lowering the Type I error rate (Bates, Kliegl, Vasishth & Baayen, 2018; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Thus, based on the analysis approach alone, the effects reported by older studies that did not use mixed-effects models (e.g., O'Reilly & McNamara, 2007) are more likely to be subject to Type I error than the effects found in more recent investigations that did use mixed-effects models (e.g., Francis et al., 2018). This difference in approaches illustrates that there is a need for a robust investigation with findings that replicate in the face of the current replication crisis in social sciences (Gelman, 2015; Gelman & Geurts, 2017). This need motivated the use of Bayesian mixed-effects models as an analytical approach in this thesis (described in section 4.5).

Summarising, there is a practical, theoretical, and methodological need for a new investigation into the effects of individual differences and text features on comprehension of health-related texts. This is because the recommendations of some NHS guidelines related to specific linguistic features that are thought to predict comprehension by empirical researchers, such as preference for the avoidance of passive voice and nominalisations (e.g., Crossley et al., 2007; 2008; 2011; Dowell et al., 2016; Graesser et al., 2011; NHS England, 2018a; Plain English Campaign, 2018), appear not to have been tested in the context of comprehension of health-related texts (section 4.1). Furthermore, the effects of variation in some linguistic features, such as the incidence of passive voice forms and gerunds, are not explicitly accounted for by current comprehension models (Chapter 1, sections 1.2 to 1.5).

In addition, the utility of perceived comprehension judgements of reader panel members in predicting comprehension of health-related texts has not been tested (section 4.1). This is not only important from the practical perspective, but also from the theoretical as the accounts of reading comprehension do not consider how metacomprehension can interact with the developing mental representation of the text and influence comprehension (cf. van den Broek & Helder, 2017) (Chapter 2, Section 2.1.3.ii). Critically, it may be the case that an adequate complete comprehension theory should include the effects of more text features, as well as the effects of metacomprehension (Chapter 2, section 2.2), within the wider comprehension system. This warrants an effort to expand on the current comprehension models by investigating the effects of texts features, alongside the effects of individual differences, speculated to affect comprehension by theoretical accounts of comprehension (e.g., Kintsch & Rawson, 2007; Perfetti, 2007; Tunmer & Chapman, 2012; van den Broek & Helder, 2017) and descriptive research (e.g., Crossley et al., 2007; 2008; 2011).

4.3. Research Aims and Research Questions

The lack of empirical evidence for the guidelines used by health-related information writers, in the context of the questions raised by the analysis of reviewed literature, motivated my research aims and research questions. I aimed to address the methodological shortfalls of previous studies and to identify the factors that predict reading comprehension of health-related information in the adult population. To track the variation in comprehension across individual differences, in language background and in other dimensions, typically developed monolingual speakers of English as well as ESL speakers were tested. I concentrated on Polish ESL speakers as the most common language spoken by the eight million foreign born people in the UK is Polish (Office for National Statistics, 2016). In addition, I sampled from a relatively wide population of individuals of different ages, literacy profiles, and educational levels. This allowed me to examine the impact of the linguistic properties of texts on comprehension of health-related information, and the impact of the effects of the interactions between individual differences and text features on comprehension of health-related texts.

To provide a robust evidence base for the development of guidelines to promote the production of understandable health-related texts, the project aimed to answer the overarching question: How do adults varying on a range of different individual characteristics understand printed health information? My work attempted to answer this overarching question by addressing the following specific research questions:

RQ1. How do reader attributes predict comprehension of written health-related information?

RQ2. How do textual characteristics predict comprehension of written health-related information?

RQ3. How do the effects of reader attributes and textual characteristics interact in predicting the comprehension of health-related information?

These research questions are addressed in Study 3 (Chapter 7). However, considering, by extension, the use of readability formulae in comprehension of health-related information (e.g., Wang et al., 2013), and the usage of perceived comprehension judgments in predicting the comprehension of printed health information (e.g., Burrow & Forrest, 2015), the ancillary focus of the investigation was to address the questions re-stated as follows:

RQ1.a. How do reader attributes predict perceived comprehension of written health-related information?

RQ2.a. How does variation in text readability predict perceived comprehension of written health-related information?

RQ3.a. How do the effects of reader attributes and variation in text readability interact in predicting the perceived comprehension of health-related information?

and

RQ2.b. How do textual characteristics predict the readability of written health-related information?

4.4. Approaches, Methods, and Techniques

The project approach applied empirical methods to testing reading comprehension of printed health-related information among a diverse sample of adults (see Figure 4.1 for a visual representation of the research design). Establishing the readability levels of sampled health-related texts was important, as, based on their readability scores, a small number of texts rated as having high or low readability was selected for inclusion in the subsequent studies. Consequently, the first study involved using the Coh-Metrix tool (Graesser et al., 2004) to analyse the text characteristics of a sample of health-related texts obtained from various NHS Trusts. Specifically, printed health-related texts containing written information with no images or illustrations, tables, or excessive formatting, were used. I focused on readability as assessed using two text readability formulae, namely the FRE (Flesch, 1948) and the Coh-Metrix L2 Readability Index (RDL2; Crossley et al., 2008) (Chapter 3, section 3.2). The FRE was chosen due to its widespread use in the analyses of readability of health-related texts (e.g., Wang et al., 2013), whereas the RDL2 was selected because it is claimed to be theoretically motivated (Crossley et al., 2008), thereby it had the potential to be a relatively good proxy for, or predictor of, comprehension. The textual analysis also considered linguistic features of texts that were empirically or theoretically motivated or were specified in NHS guidelines (see Table 4.1 for an overview of a sample of text features, and justification for their inclusion, in the studies of this thesis). Overall, Study 1 enabled me to investigate the readability levels of sampled health-related texts, and to identify the text features that were related to the readability of health-related texts as measured using text readability formulae (RQ2.b).

The second study involved testing how text readability formulae and individual differences, such as age, health literacy, education, and language background, predict perceived comprehension of health-related texts (RQ1.a; RQ2.a; RQ3.a). Determining the predictors of perceived comprehension was important given the widespread use of patient reader panels used by NHS Trusts to assure the comprehensibility of texts (e.g., Burrow & Forrest, 2015). Theoretically, Study 2 was also important because it has been argued that readability formulae predict perceived comprehension, but not tested comprehension (e.g., Kauchak & Leroy, 2016). Critically, Study 2 in combination with Study 3 permitted to test this theory. Study 2 was also important from the methodological perspective, as the information gathered from the first two studies was used to create the stimuli for the reading comprehension task used in the third study.

The empirical methods of the third study included the testing of attributes that previous work has shown to influence comprehension (RQ1), using a mix of standardized ability and experimental tasks. All participants were tested on: English language vocabulary breadth; phonological WM capacity and processing; perceived understanding; phonological awareness, and health literacy. The participants were also asked about their age and English language proficiency. The data analysis, which I describe in detail next, integrated information on personal attributes, health-related text characteristics, and reading comprehension. This allowed me to account for the way in which comprehension is related to textual characteristics (RQ2), and how those effects are modulated by the impact of the individual differences (RQ3). In addition, this approach enabled me to examine the effectiveness of the current NHS guidelines (e.g., NHS England, 2018a) on how to write, and test understanding of, health-related texts. Critically, the findings may lead to the improvement of these guidelines or to the development of new guidelines.

Figure 4.1. Research design.

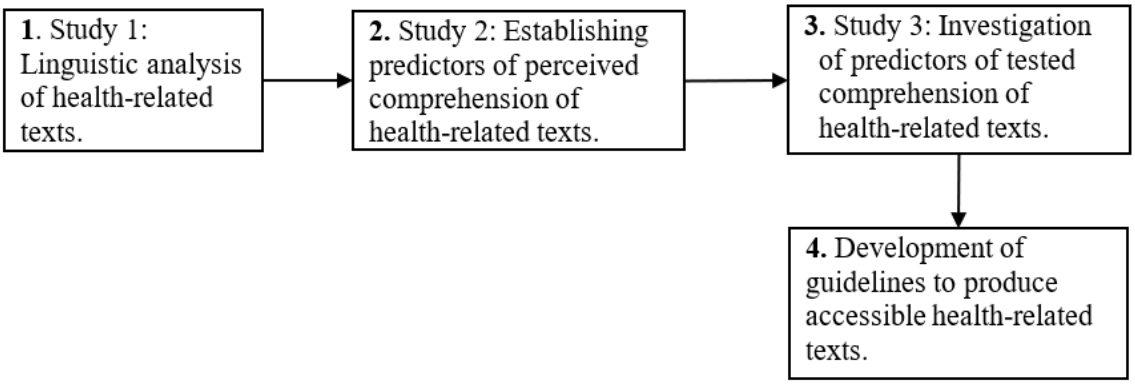

1. Study 1: Linguistic analysis of health-related texts.
2. Study 2: Establishing predictors of perceived comprehension of health-related texts.
3. Study 3: Investigation of predictors of tested comprehension of health-related texts.
4. Development of guidelines to produce accessible health-related texts.

Table 4.1. Text features included in mixed-effects models of reading.

Text Features	**Possible relation theoretical and empirical evidence for plausible effects on reading comprehension**	**Possible relation to NHS guidelines**
Average word frequency	Theorised to affect meaning-to-text integration processes (Perfetti, 2007; Perfetti & Stafura, 2014; Yang et al., 2005). The more frequent words are thought to be easier to process as more readers are hypothesised to have higher lexical quality of such words.	Advocacy for straightforward words and plain language (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018).
Average word length	Evidence indicates that word length has been used as a proxy for word complexity (Flesch, 1948), word frequency (McNamara, Louwerse, Cai, & Graesser, 2013), and lexical sophistication (Crossley et al., 2017).	Advocacy for short words (NHS, 2015; Plain English Campaign, 2018).
The incidence of passive voice forms	Simplified texts were found to contain a smaller proportion of passive voice forms than not simplified texts (Crossley et al., 2007; 2008; 2011).	Preference for active versus passive verbs (Plain English Campaign, 2018).
Text cohesion (connectives, such as, causal, temporal, logical) and text coherence (argument, conceptual, semantic and syntactic overlap)	Variation in text coherence and cohesion is thought to influence comprehension (e.g., Kintsch, 1988; Kintsch, 1998; McNamara & Kintsch, 1996). Empirical research evidence suggests that increasing text cohesion and coherence is associated with an increase in text comprehension among readers with high reading skill level (O'Reilly & McNamara, 2007; Ozuru et al., 2009). Within the domain of health-related texts, the effects of text coherence were associated with reading comprehension depending on reader's profile and the incidence of short words and sentences (Liu et al., 2009; also refer to Chapter 3, section 3.3).	No obvious relation.
The incidence of verbs ending in *ing* (gerunds)	Higher incidence of gerunds was found to be associated with original versus simplified texts aimed at ESL learners (Crossley et al., 2007; 2008).	Preference for avoidance of nominalisations (Plain English Campaign, 2018).

The frequency of occurrence of superordinate words (hypernymy)	Simplified and beginner level texts were found to have a higher incidence of hypernyms than the original and more advanced texts for learners of ESL (Crossley et al., 2008; 2012). High incidence of hypernyms was also found to be positively correlated with ease of processing judgements (Crossley et al., 2017). Varying in the degree of specificity and abstractness, hypernymy can be considered a proxy for word commonality as the more frequent words tend to be hypernyms (Crossley et al., 2012).	Advocacy for straightforward words and plain language (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018).
Flesch Reading Ease (FRE; Flesch, 1948)	Widely used in the assessment of readability of health-related texts (Wang et al., 2013). In the context of health-related texts, the effects of the FRE interacted with the effects of WM, age, and text coherence, to predict comprehension (Liu et al., 2009; see also Chapter 3, section 3.3).	Advocacy for straightforward words and plain language (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018).
Coh-Metrix L2 Readability Index (RDL2; Crossley et al., 2008)	More grounded in reading comprehension theories, such as the CI model (Kintsch, 1998), than the FRE (Crossley et al., 2008; refer also to Chapter 3, section 3.2). The RDL2 scores were found to account for 86% of the variance in Japanese students' reading comprehension (Green, 1999). It is claimed that the RDL2 discriminates between texts aimed at different English language proficiency readers relatively well (e.g., Crossley et al., 2011), but understanding of these texts has not been extensively tested (Crossley et al., 2007; 2008; 2011).	Advocacy for straightforward words and plain language (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018).

4.5. Data Analysis

In this thesis, of primary interest were the effects of interactions between individual differences and text features on reading comprehension of health-related information (RQ3). Text feature effects vary by people (e.g., Francis et al., 2018; Liu et al., 2009) and the presumption of constant text feature effects across individuals with different developmental profiles corresponds to a simplified account of comprehension phenomena (cf. Gelman, 2015). Thus, where possible, the analyses considered the plausible effects of all theoretically and empirically motivated individual differences and text features on reading comprehension, in isolation and in interactions with each other. The investigated variables and interactions were based on research questions, literature reviewed in the literature review chapters (Chapters 1 to 3), the recommendations of the NHS's guidelines relating to linguistic features (e.g., NHS England, 2018a), and the notion that the effects of text features vary by people (e.g., Liu et al., 2009). Critically, however, the analyses were conducted in the Bayesian inferential framework. I justify this choice next.

4.5.1. Inferential Frameworks

The frequentist inferential framework likens probability to the frequency of an event over an infinite number of hypothetical observations (McKee & Miller, 2015), for which the distribution of potential (expected) values is described in terms of a sampling distribution, for example, the normal-shaped sampling distribution of a statistic like the mean for observed measurements for some sample size, over repeated (hypothetical) samples. In other words, the frequentist framework is founded on sampling distributions of invented data (Kruschke & Liddell, 2018a). These sampling distributions are used to compute the *p*-values for null hypothesis significance testing (NHST), and confidence intervals for estimating the uncertainty of the effects. One of the main issues with the frequentist approach is that the sampling distribution, therefore the *p*-values and confidence intervals, are influenced by the sample size and the number of tests or comparisons conducted. This is problematic because with different sample sizes, or different numbers of tests, frequentist models produce different estimates (Kruschke & Liddell, 2018a). Critically, trivially small effects can be found to be "significant" with very large sample sizes (Kruschke & Liddell, 2018a). One way to overcome the limitations of the frequentist inferential framework is to use a different inferential framework, such as the Bayesian inferential framework, which is not based on sampling distributions (Kruschke & Liddell, 2018a). This is one key motivation for the use of the Bayesian inferential framework in this thesis.

The Bayesian inferential framework equates probability with a degree of belief regarding a possible event, like the estimated coefficient value for the effect of a variable (Kruschke & Liddell, 2018b). These beliefs can be based on past studies or other observations and are updated in light of new information. Essentially, for given data, there is a set of considered potential explanations for the observed values. Before observing these data, these potential explanations have some probability of being the best explanation of the data. As we accumulate data, we shift the probability towards the potential explanations that better account for the data, while shifting the probability away from those explanations that do not account well for the data (Kruschke & Liddell, 2018b). This process of shifting probability is so intuitive that it can be illustrated using relatively simple examples. For example, in a criminal case, each suspect accused of committing a crime has some probability of being guilty of committing that crime. However, as the detective gets more information on each suspect, suspicion of who committed the crime is reallocated across the suspects. If the new data eliminates some suspects, for example due to evidence that they were in a different location when the crime occurred, the probability of the remaining suspects having committed the crime increases. Once all suspects, but one, have been eliminated, assuming that the culprit was included in the set of suspects, the remaining suspect is the most likely to have committed the crime and the probability of them being guilty shifts again. This intuitive reallocation of probability across possibilities that are adjusted in light of new data is Bayesian reasoning (Kruschke & Liddell, 2018b).

In Bayesian data analysis, the possibilities, or considered potential explanations, are parameter values in mathematical descriptions (Kruschke & Liddell, 2018b). Before considering new data, we translate our beliefs about the magnitude of the possible effect into a prior distribution of probabilities that is assigned to each parameter value (i.e., potential coefficient) for an effect. Colloquially these prior distributions are simply referred to as priors. There is a probability distribution for every effect and the effect may vary from one value from another, but it is not the case that every value is equally probable. After establishing a prior distribution over parameter values, Bayesian inference re-distributes the probability over the parameter values given new data. This re-distributed probability distribution is referred to as the posterior distribution. The posterior distribution is the end goal of Bayesian inference. It quantifies uncertainty over parameters and encodes the allocation of probabilities throughout all parameter values (Kruschke & Liddell, 2018b). Critically, the posterior distribution can be assessed to determine what range of parameter values is the most plausible given our prior beliefs about the effects and the data.

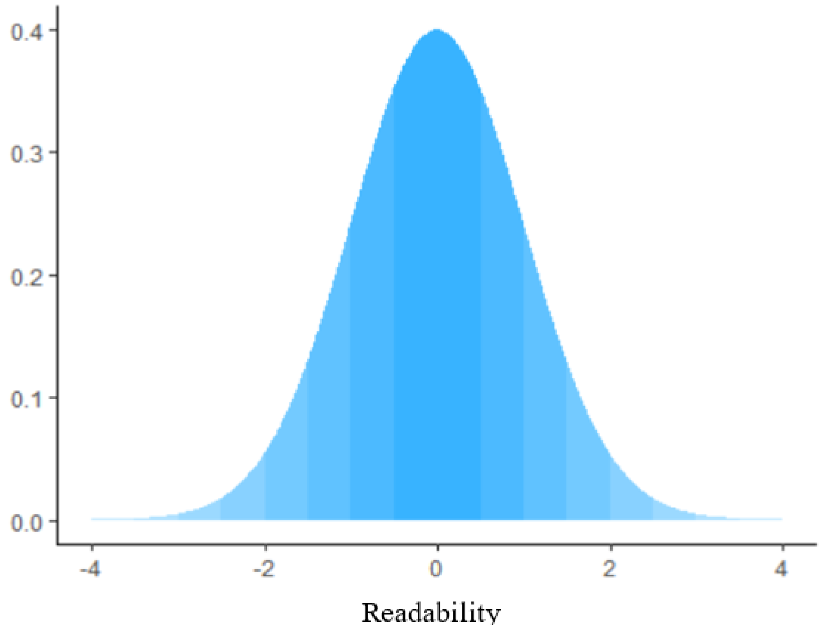
Overall, Bayesian analysis starts with a prior distribution made of beliefs about plausible parameter values, then considers new data, and arrives at a posterior distribution which places higher probability on parameter values that are relatively consistent with the data (Kruschke & Liddell, 2018b). This is different from the frequentist inferential framework because in Bayesian analysis there is no need to generate sampling distributions from null hypotheses (Kruschke & Liddell, 2018b). Thus, in a Bayesian analysis, estimates of uncertainty can be summarised in terms of posterior credible intervals instead of *p*-value-based confidence intervals. Critically, eschewing dependence on sampling distributions (of hypothetical observations, given some specific sampling scheme) means that the Bayesian posterior distribution is robust to variation in sample size, or in the number of tests or comparisons conducted (Kruschke & Liddell, 2018a).

In addition to theoretical reasons for use of the Bayesian inferential framework in this thesis, there are also pragmatic motivations. The first relates to the process of model comparisons. Model comparison refers to comparing different models based on the accuracy of their predictions given the data. In traditional frequentist approaches, model comparison is a relatively difficult procedure, because models are often chosen based on their relative capacity to describe the underlying data. Critically, the more complex frequentist models tend to fit the data better, but the more complex frequentist models are at risk of over-fitting due to being over-parameterised (MacKay, 2003). Over-fitting happens when a more complex model makes worse predictions, on average, than a simpler model. Over-fitting may be reflected in the spurious detection of effects that will not be replicated in future studies. The use of Bayesian posterior distribution militates against the problem of over-fitting by automatically penalizing an increase in model complexity that does not improve the model's predictions (Kruschke & Liddell, 2018b). Thus, in Bayesian analyses, the more probable models tend to be the more parsimonious ones that make more accurate predictions that generalise better to future data (MacKay, 2003).

Critically, over-fitting can still be a problem in Bayesian analysis if all the models that are considered are unreasonably specified (Gelman, Simpson, & Betancourt, 2017). This leads to the second pragmatic reason for the choice of Bayesian analysis in this thesis: some models cannot be fitted using frequentist inference. This is typically observed as a failure by frequentist model-fitting algorithms (e.g., mixed-effects model-fitting algorithms) to converge on a set of estimates for the coefficients of effects. The models will often not converge because prior expectations about potential coefficient estimates assign equal probability for any possible value of each coefficient (Depaoli & van de Schoot, 2017), making it harder, under some circumstances, for the model-fitting algorithms to converge on the specific coefficient estimates that best fit the data (maximise the likelihood of the data) (Eager & Roy, 2017). Frequentist algorithms can be understood to work as a special case of Bayesian analysis incorporating a complete lack of prior knowledge about the plausibility of different potential parameter estimates, in which that lack of knowledge is captured in terms of flat priors. In flat priors, the probability distribution describing the probabilities of different coefficient estimates allocates equal probability to each possible coefficient value. Flat prior beliefs are problematic because it is simply not true that any estimated effect is equally probable. Furthermore, flat prior beliefs tend to permit estimates of effects that are too large, increase the risk of over-fitting, and which often fail to produce good predictions that can generalise to future studies (Gelman et al., 2017).

In contrast to flat priors, the use of regularising priors allows the corresponding model estimates to generalise, guarding against over-fitting. The goal of the use of regularising priors is to provide more stable inferences than would be obtained from frequentist inference or from Bayesian inference with flat priors (Gelman et al., 2017). In practice, this means that models that would not have converged without regularising priors, do converge with regularising priors, and can produce thereby relatively accurate predictions. One type of prior that can serve as a regularising prior is the weakly informative prior because it gives the model enough information to avoid theoretically implausible inferences but is still flexible enough to permit a relatively large amount of variation in the effects of the parameters (Depaoli & van de Schoot, 2017). For example, if one assumes a normal-shaped (Gaussian) prior probability distribution, centred on zero, for the potential values of the coefficient estimate for the effect of readability on comprehension accuracy then we are assuming that the effect of readability could well be zero (the location of the peak of normal curve) but could be somewhere above zero (a positive effect) or below zero (a negative effect) with diminishing probability for larger coefficient values, and very small probability for very large values (see Figure 4.2). It is important to note that Figure 4.2 also shows how weakly-informative regularising priors do not have to bias the effect estimate in either direction a priori, as the probability density of an estimate can be equally split above and below zero.

Figure 4.2. Example Gaussian-shaped prior probability distribution for an effect of readability.



In addition to the potential for more accurate predictions due to the avoidance of over-fitting (Depaoli & van de Schoot, 2017; Gelman et al., 2017; Gelman & Henning, 2017), Bayesian models allow us to assume appropriate (potentially non-normal) probability distributions to model observations (Kruschke & Liddell, 2018b). This is important because most frequentist models assume that observed values of an outcome or dependent variable are normally distributed (or, equivalently, that model residuals are normally distributed). Consequently, the standard practice in frequentist analysis is to make the data appear normally distributed by transforming it or by removing what are perceived to be outlier observations (e.g., Osborne & Overbay, 2004). However, transforming data can bias model's predictions and estimates (e.g., Martin & Williams, 2017), whereas removing data that are not errors of measurement constitutes selective bias and artificially reduces the variance in the data (Kruschke & Liddel, 2018b). Observed outcomes need not be normally distributed in nature so that, for example, reaction time distributions are typically found to be skewed. Critically, simulation studies (e.g., Martin & Williams, 2017) have shown that making an appropriate assumption about the data underlying (that is, generating) the observed outcomes increases the power of the analysis to detect plausible effects, reduces bias in the estimates of the effects, and permits more accurate prediction of future data.

Overall, the pragmatic reasons for choosing Bayesian models in this thesis are based on the potential improvement in the accuracy of predictions that are more likely to generalise compared to frequentist models (Krushke & Liddell, 2018b). However, it is important to acknowledge that regardless of whether a researcher is using frequentist or Bayesian inference, any dataset can be consistent with several models, each of which can lead to a different set of inferences (Gelman & Henning, 2017). Thus, in any analysis, choices made must be explicitly accounted for so that the analyses are reproducible and understandable. Consequently, in my analyses I aimed to explicitly justify my choices by following recent best practice guidelines for statistical science (see Gelman & Henning, 2017). I aimed to be transparent and impartial while acknowledging multiple perspectives to the data analyses and the context dependence of my findings, where relevant. Next, I discuss my plan of analyses.

4.5.2. Plan of Analyses

In the analysis of data obtained from each study, I employed mixed-effects models where that was justified by the clustering of observations in the data. Mixed-effects models have many advantages over traditional methods such as ANOVAs and *t*-tests. As previously mentioned in section 4.2, mixed-effects models allow to account for random variation between individuals and random variation between texts, thereby accounting for additional uncertainty in the data and lowering the Type I error rate (Bates et al., 2018; Matuschek et al., 2017). The mixed-effects models in my analyses were Bayesian. The justification for this is largely pragmatic, specifically the inclusion of prior information in the model often results in better model predictions compared to models with flat priors or frequentist models (Gelman, 2015; Gelman et al., 2017) (refer to section 4.5.1 for the full list of reasons). Critically, the use of Bayesian mixed-effects models in this thesis could have methodological implications in reading comprehension research. This is because in psycholinguistic research, as well as within social sciences research, Bayesian mixed-effects models are relatively rarely used (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). Thus, it is of theoretical and methodological interest to investigate whether theoretical and empirical research findings replicate using Bayesian inference.

4.6. Summary

This chapter outlined the rationale for the three studies conducted in this project. First, I reviewed the guidelines given to health-related information writers on how to write health-related texts, and how to ensure that these texts are understandable using patient reader panels (e.g., Burrow & Forrest, 2015; NHS England, 2018a; NHS England, 2018b). Second, relating to the literature reviewed in Chapters 1, 2, and 3, I identified the research gap that this work aimed to fill and posed the research questions that it attempted to answer. Third, I briefly described the approaches, methods, and techniques of the three studies of this thesis. Last, I justified the use of Bayesian inferential framework in the analyses of the data obtained from the three investigations. In the next chapter, Chapter 5, I discuss the first study of this thesis.

Chapter 5: Linguistic Determinants of Textual Measures of Readability Estimates of Health-Related Information

This chapter describes and discusses the first study included in this thesis (Study 1). Study 1 had a two-fold purpose: (i) to determine the readability of a sample of health-related texts and establish linguistic predictors of readability of the sample as assessed using readability formulae; (ii) and to provide a basis for selecting high-readability and low-readability health-related texts for inclusion in the subsequent studies. This chapter begins with a short literature review, followed by the method section which is followed by the results section. In the results section of this chapter, I describe the model selection process and sensitivity checks in detail, pre-empting detailed repetition in subsequent chapters. The chapter ends with a brief discussion.

5.1. Literature Review

Written health-related information materials are widely used within health care settings. Patients can be given a health-related document before, during, or after a physical appointment with a clinician. Furthermore, patients can access health-related documents on National Health Service (NHS) Trusts' websites, and these documents can provide additional information that might have been omitted during a physical appointment (Patel et al., 2013). Recently, the utility of health-related documents to educate patients in their own time has increased. This is because there is a long-term drive to fully involve individuals in their own healthcare, thereby enabling them to make informed choices about managing their healthcare needs (NHS England, 2014). Critically, it is thought that involving patients in their own healthcare, for example by helping them to develop health knowledge with the aid of health-related texts, can reduce the increasing demand pressures faced by the NHS (e.g., NHS Digital, 2016; 2017; 2018b). However, the readability of health-related texts, or the comprehension level readers must have to understand the written material (Albright et al., 1996; Beck et al., 1991; Flesch 1948), might not be sufficient to fully understand these texts, hindering the development of health knowledge.

Researchers have exposed issues with readability and usability of a wide range of health-related documents such as texts relating to: hormone therapies (Charbonneau, 2013); breast cancer risk assessment tools (Cortez, Milbrandt, Kaphingst, James, & Colditz, 2015); thyroid surgery (Patel et al., 2013); cancer screening information (Liu et al., 2013), and so on. These studies show that many health-related documents require high levels of reading skill to understand their content. This is of great concern as health-related documents are produced with the intention of being easy to understand, and the expectation placed on patients is that they should be able to understand the information presented in health-related documents (Wang et al., 2013). However, relatively recent evidence suggests that approximately 43% of working-age adults in England do not have literacy skills at a level which would allow them to understand and make use of health information (Rowlands et al., 2015). Consequently, a notable proportion of the UK population is unlikely to comprehend all the information presented in health-related documents. Importantly, given the emphasis on readability in guidance to health information producers (e.g., NHS England, 2018a), all health-related texts should be highly readable.

As mentioned in Chapter 4 (section 4.1), regulatory efforts have been made to improve the design of health-related documents within the NHS. These regulatory efforts resulted in various recommendations, such as to avoid the use of passive voice, keep sentences short, use plain language and straightforward words (e.g., NHS England, 2018a; 2018b; Plain English Campaign, 2018; Marsay, 2017a; Marsay, 2017b; see Table 4.1. in Chapter 4). However, most of these recommendations have not been empirically tested. It must be asked, therefore, whether documents produced by the NHS are, in fact, easy to understand or readable?

The readability of health-related documents has been frequently assessed using measures such as readability formulae (Wang et al., 2013), and self-reported perceived comprehension measures, such as comprehension judgements given by individuals (e.g., NHS England, 2018a; Riche, Reid, Robinson, & Kardash, 1991) (this chapter focuses on readability measures, but the next chapter focuses on perceived comprehension measures). Readability testing is intended to provide an indication of the comprehension difficulty level of written text (Flesch, 1948). One of the most trusted readability measures, including in the context of health-related texts (Wang et al., 2013), is the Flesch Reading Ease (FRE; Flesch, 1948) (Chapter 3, section 3.2). The parameters of the FRE formula encapsulate some of the recommendations, specifically to use shorter sentences and shorter words, that the NHS's health-related information producers follow (see Chapter 4, Table 4.1).

Although some of the readability formulae, such as the FRE (Flesch, 1948) are more frequently used than others (Wang et al., 2013), such as the Simple Measure of Gobbledygook (SMOG; McLaughlin, 1969) (Chapter 3, section 3.2), there is no consensus as to which one is best suited for assessing readability of health-related documents (Badarudeen & Sabharwal, 2010). One potential explanation for the lack of consensus may be that the process of assessing readability is complicated as in theory the different formulae should be measuring the same construct, namely readability. However, the different readability formulae use different regression weights to estimate readability given the same or similar text features, such as the FRE and the SMOG, or use different text features to assess readability, for example, the FRE and the Coh-Metrix L2 Readability Index (RDL2; Crossley et al., 2008) (Chapter 3, section 3.2). Presumably, the use of different text features is motivated by the desire to improve the accuracy of readability formulae in measuring the underlying construct of readability (e.g., Crossley et al., 2008). However, the use of different text features in different readability formulae is problematic as both FRE and RDL2 are supposed to measure readability (Crossley et al., 2008; Flesch, 1948). The use of different text features to assess readability has implications for the validity of readability assessment, as it is questionable whether the different formulae assess the same construct.

Within the domain of health information, different readability formulae may produce different readability scores for the same texts (Wang et al., 2013). Given that readability estimates provided by different readability formula can vary, it is difficult to determine which readability formula to trust in predicting comprehensibility. Thus, in addition to measuring readability of health-related texts using readability formulae to sample across readability range, it is critical to consider the effects of theoretically or empirically motivated text feature predictors on readability scores. This is because if we assume that the association between variation in readability scores and variation in comprehension is close (e.g., Beck et al., 1991;

Flesch, 1948), then we can infer benefit to patient understanding should result from writing texts with high incidence of features with positive effects on readability scores. This assumption, however, is open to question (e.g., Kauchak & Leroy, 2016), and this question motivated the subsequent studies in this thesis.

Crossley et al.'s (e.g., 2007; 2008; 2011) (see Chapter 3, section 3.2.1; Chapter 4, section 4.2) studies provide some suggestive evidence as to what variables may have an influence on tested comprehension of English as Second Language (ESL) texts and potentially texts intended for first language (L1) English audiences. However, it has yet to be determined whether the linguistic features mentioned in their studies influence readability of health-related texts. Some suggestive evidence for the potential effectiveness of the RDL2 in predicting comprehension of health-related texts comes from a study involving 52 outpatients who were asked to read three health-related texts (Riche et al., 1991). These patients were split into a metacomprehension group (n = 15) and a cloze test (see section 3.1.3 of Chapter 3 for a description of a cloze test) group (n = 37). Participants in the metacomprehension group were asked to speak their thoughts out loud when they encountered something confusing in the health-related texts they read. In turn, participants in the cloze test group were required to select a replacement word for every deleted word from the test measuring comprehension of health-related texts.

Some individuals from the metacomprehension group reported that technical words, passive voice, gerunds, and rare phraseology increased the perceived difficulty of health-related texts (Riche et al., 1991). Other participants self-reported that they preferred the usage of frequent words over the rarer words, suggesting that the use of frequent words may increase perceived readability of texts (Riche et al., 1991). The reported results of participants in the cloze group lacked detail and were descriptive. Specifically, Riche et al. (1991) reported that for 62% of their participants the sample of health-related texts was too difficult be understood adequately, but it was not specified what constituted an adequate level of understanding, and it was not explained why or what text features contributed to this finding. These critical limitations of previous research (e.g., Riche et al., 1991) highlight the need for a study employing robust data analyses that can make predictions regarding the plausible effects of text features on comprehension or readability levels of health-related texts.

In summary, readability of health-related texts has not been extensively studied. Suggestive evidence (Riche et al., 1991) indicates that some of Crossley et al.'s (2007; 2008; 2011) findings (discussed in Chapters 3 and 4) may apply to health-related texts as the text features that were judged to influence the perceived understanding of health-related texts (Riche et al., 1991), were also identified as potential predictors of comprehension by Crossley et al. Additionally, although some of the guidelines endorsed by the NHS, such as the preference for active over passive voice (e.g., Plain English Campaign, 2018), have not been explicitly related to reading comprehension theories (e.g., Kintsch & Rawson, 2007), they appear to be supported by some empirical evidence (e.g., Crossley et al., 2007; 2008; 2011; Riche et al., 1991; Street, 2020; Street & Dąbrowska, 2014). Overall, more research is needed to investigate the readability of health-related documents used by the NHS, as assessed using common readability formulae, and what text features influence readability of those texts. In addition, it is critical to examine whether the usage of text features recommended by the NHS guidelines, such as preference for active voice (e.g., Plain English Campaign, 2018), has led to a greater prevalence of health information texts with high readability levels. This motivates the research aims of this study.

5.1.2. Research Aims

I aimed to examine what existing readability formulae, such as the FRE and the RDL2, reveal about the readability of health-related texts produced by the NHS, and which linguistic characteristics predict the readability scores of these texts. Another goal of this phase of research was to explore whether the recommendations made in the NHS guidelines (e.g., NHS England, 2018a) improve readability scores of health-related texts.

5.1.3. Research Questions

RQ5.1. What do readability estimates reveal about the readability of health-related texts?

RQ5.2. What linguistic properties of health-related texts contribute to readability scores?

RQ5.3. Do the recommendations of NHS guidelines related to specific linguistic features improve readability levels of health-related texts?

5.1.4. Hypotheses

$H_{5.1}$. If health-related texts are designed to be easy to understand, then the readability scores of both readability formulae should be high.

$H_{5.2}$. If FRE and RDL2 measure the same construct, specifically readability of health-related texts, then they should be influenced by the same or similar predictors.

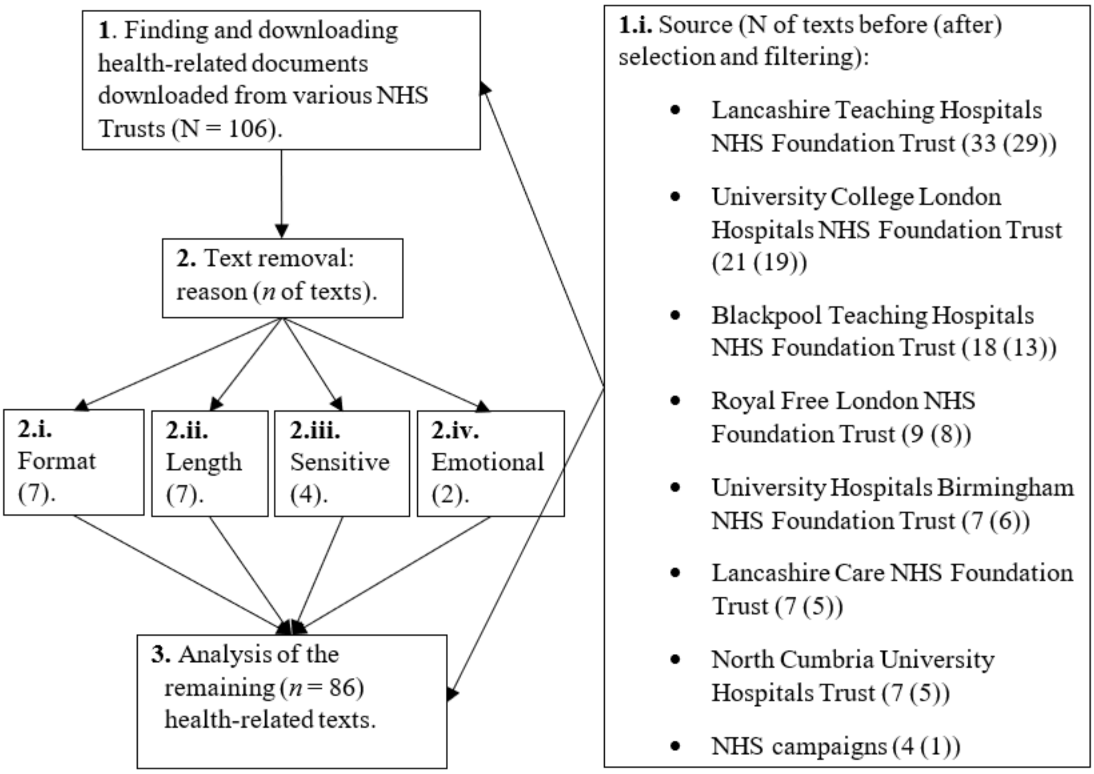
$H_{5.3}$. Following the recommendations of NHS guidelines related to specific linguistic features should improve readability levels of health-related texts.

5.2. Method

5.2.1. Materials and Procedure

I selected an opportunity-sample of 106 health-related documents from the websites of easy-to-access NHS Trusts, such as the Blackpool Teaching Hospitals NHS Foundation Trust (see Figure 5.1). The sample of health-related texts consisted of documents that were available for download at the time of data collection. After the initial selection, the documents were reviewed by two experts (one linguist and one psychologist). This filtering process resulted in the exclusion of 20 leaflets, leaving 86 leaflets for the analysis (Figure 5.1). Amongst the 20 excluded leaflets, seven were excluded because they contained features that would not have been captured by the Coh-Metrix application (Graesser et al., 2004), such as images, tables, or excessive text formatting. Another seven texts were removed because they were too short to be reliably analysed using the Coh-Metrix application; four texts were excluded as they related to sensitive topics, such as sexually transmitted diseases; and two texts were removed as they were judged to potentially evoke strong emotional reactions, specifically texts describing cancer treatments.

Figure 5.1. Health-related documents selection process.


1. Finding and downloading health-related documents downloaded from various NHS Trusts (N = 106).
2. Text removal: reason (*n* of texts).
2.i. Format (7).
2.ii. Length (7).
2.iii. Sensitive (4).
2.iv. Emotional (2).
3. Analysis of the remaining (*n* = 86) health-related texts.
1.i. Source (N of texts before (after) selection and filtering):
• Lancashire Teaching Hospitals NHS Foundation Trust (33 (29))
• University College London Hospitals NHS Foundation Trust (21 (19))
• Blackpool Teaching Hospitals NHS Foundation Trust (18 (13))
• Royal Free London NHS Foundation Trust (9 (8))
• University Hospitals Birmingham NHS Foundation Trust (7 (6))
• Lancashire Care NHS Foundation Trust (7 (5))
• North Cumbria University Hospitals Trust (7 (5))
• NHS campaigns (4 (1))

I used the Coh-Metrix application (Graesser et al., 2004) to estimate the readability of the selected sample of health-related texts. This choice was motivated by the widespread usage of the Coh-Metrix application in estimating readability of texts (see Dowell et al., 2016 for a review of published studies using Coh-Metrix), including health-related texts (e.g., Liu et al., 2009). The Coh-Metrix application is a computational linguistics facility that was developed for calculating cohesion and coherence metrics for written and spoken texts. Specifically, it produces indices of the linguistic and discourse representations of a text and can be used to analyse texts on more than 200 measures of cohesion, language, and readability. It is also important to note that the scores of these measures are often subject to the output of third-party parsers, lexicons, and word frequency databases, such as CELEX (Baayen et al., 1995). It is beyond the scope of this thesis to specify precisely how each measure is computed, but attempts were made where a detailed description of a particular measure was thought to be especially relevant to the present investigation. For example, the formulae underlying the FRE (Flesch, 1948) and the RDL2 (Crossley et al., 2008) scores, which were calculated using the Coh-Metrix application, have been described and discussed in Chapter 3 (section 3.2). In this study, I paid particular attention to the FRE and the RDL2 indices for reasons of widespread use in the assessment of readability of health-related texts (e.g., Liu et al., 2009), and a potential for an improvement in the accuracy of predictions of comprehension (Crossley et al., 2008), respectively (Chapter 3, section 3.2; Chapter 5, section 5.1). As mentioned in section 5.1, the FRE provides an indication of text readability that is based on word and sentence length found in the text (Flesch, 1948). In comparison, the RDL2 considers three linguistic indices: content word overlap, word frequency, and sentence syntax similarity (Crossley et al., 2008).

5.2.2. Variable Selection

In the analyses to be reported, I examined the factors that predicted variation in the (regression formulae based) estimated readability of the sample of health-related texts. In these analyses, the outcome (or dependent) variables were the readability scores generated for each text by the FRE (Flesch, 1948) and RDL2 (Crossley et al., 2008). The choice of selecting these two readability formulae was guided by the finding that the former is one of the most commonly used tools for assessing readability of health-related documents (Wang et al., 2013) while the latter is argued to consider cognitive and psycholinguistic theories of reading comprehension (e.g., Crossley et al., 2008; 2011).

Since I had two outcome variables, I had to create two sets of Bayesian models to answer RQ5.2 and RQ5.3. One set had the FRE estimates as the outcome variable, whereas the other had the RDL2 estimate as the outcome measure. Although both the FRE and RDL2 are supposed to be measuring the same construct, their underlying regression formulae, and the weights associated with these formulae, are different. Therefore, each formula had to be considered independently. By considering each formula independently, each text had only one FRE and RDL2 observation associated with it, warranting the use of linear models to analyse readability estimates.

The text feature predictors, or independent variables in a regression analysis, were chosen based on the literature reviewed in Chapters 1 to 3, and the current practices employed by NHS England Trusts (e.g., NHS England, 2018a; see also Table 4.1 in Chapter 4 for a list of candidate text features evaluated against their relation to NHS's guidelines and theoretical and empirical research evidence). The predictors included in the Bayesian models in this study, alongside a justification for their inclusion, are listed in Table 5.1. It is important to mention that given the relatively small sample of health-related texts considered, the number of predictors had to be reasonable to keep the models parsimonious and avoid over-fitting (see Chapter 4, section 4.5.1, for a description of over-fitting). The definition of reasonable is subjective, therefore it is important to acknowledge that other researchers could have chosen a different set and a different number of predictors, because different analysts can have different perspectives on data analyses (Gelman & Henning, 2017).

Table 5.1. Text features included in Bayesian models of readability.

Text Features	**Justification for inclusion**
Average word frequency	• Part of RDL2 but not FRE regression formula. • Theorised to affect meaning-to-text integration processes (e.g., Perfetti, 2007; Perfetti & Stafura, 2014; Yang et al., 2005); proxy for language exposure (e.g., Brysbaert et al., 2016). • Related to advocacy for straightforward words and plain language (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018). • There is research evidence to suggest that rare words are perceived as more difficult to understand (Riche et al., 1991), and are less likely to be understood (Leroy & Kauchak, 2014), than frequent words. • There is evidence to suggest that simplification of health-related texts based on substitution of rare words with more frequent words is effective at improving comprehension (Leroy, Endicott, Kauchak, Mouradi, & Just, 2013).
Average word length	• Part of FRE but not RDL2 regression formula. • Some argued that it is a proxy for word complexity (Flesch, 1948), frequency (McNamara et al., 2013), and lexical sophistication (Crossley et al., 2017). • Advocacy for short words (NHS, 2015; Plain English Campaign, 2018).
The incidence of passive voice forms	• There is some evidence to suggest that simpler texts contain a smaller proportion of passive voice forms than more difficult texts (Crossley et al., 2007; 2008; 2011). • Preference for active versus passive verbs (Plain English Campaign, 2018). • Suggestive evidence that passive words are perceived as more difficult than active words in health-related texts (e.g., Riche et al., 1991).
Causal connectives; logical connectives (text cohesion)	• Connectives are theorised to aid comprehenders in constructing the textbase (e.g., Kintsch & Rawson, 2007), and cohesive situation model (Dowell et al., 2016). • There is some evidence to suggest that logical connectives, such as *therefore, if* , are relatively good predictors of text difficulty level (e.g., Green, Khalifa, & Weir, 2013). • There is some evidence to suggest that the incidence of causal connectives, such as *because, so* , predicts comprehension (e.g., McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996).
The incidence of verbs ending in *ing* (gerunds)	• Higher incidence of gerunds was found to be associated with original versus simplified texts aimed at ESL learners (Crossley et al., 2007; 2008). • Preference for avoidance of nominalisations (Plain English Campaign, 2018). • There is some suggestive evidence that gerunds increase perceived difficulty of health-related texts (e.g., Riche et al., 1991).

Referential cohesion, Latent Semantic Analysis (LSA), causal cohesion (text coherence)	• The effects of text coherence were found to be associated with reading comprehension of health-related information depending on reader's profile and the incidence of short words and sentences (Liu et al., 2009; also refer to Chapter 3, section 3.3). • Referential cohesion* is a measure of coherence calculated using indices of argument and conceptual overlap in adjacent and all sentences. Argument and conceptual overlap are theorised to predict comprehension (Crossley et al., 2011; Kintsch & Rawson, 2007). • LSA measures sentence semantic, also referred to as conceptual, overlap. The more closely bound are the sense relations between sentences, the easier it should be for the reader to link propositions together and construct the textbase (Kintsch, 1988). • Causal cohesion* is measured by Coh-Metrix (Graesser et al., 2004) by calculating the ratio of causal verbs, for example *make* , to causal particles, such as *as a result* (Crossley et al., 2008). The ratio of causal verbs to causal particles is thought to relate to the text's ability to convey causal content (Crossley et al., 2007). In addition, causal cohesion is theorised to be relevant to the construction of a coherent situation model (Zwaan & Radvansky, 1998; Graesser et al., 2011). • *Both constructs are classed as measures of coherence, since they examine sense relations within the text (see sections 1.5 and 1.6. of Chapter 1).
The frequency of occurrence of superordinate words (hypernymy)	• Simplified and beginner level texts were found to have a higher incidence of hypernyms than the original and more advanced texts for learners of ESL (Crossley et al., 2008; 2012). • High incidence of hypernyms was also found to be positively correlated with ease of processing judgements (Crossley et al., 2017). • Varying in the degree of specificity and abstractness, hypernymy can be considered as a proxy for word commonality as the more frequent words tend to be hypernyms (Crossley et al., 2012). • Advocacy for straightforward words and plain language (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018).

In both sets of readability models, I kept the text feature predictors constant. The reasoning behind this was to find out if RDL2 and FRE readability scores were predicted by similar text features. One would expect RDL2 and FRE scores to be predicted by the indices associated with these regression readability formulae alone, but this may not necessarily be the case. For example, Flesch (1948) argued that a measure of word length was also a measure of word complexity, whereas some other researchers argued that word length has been used as a common proxy for word frequency (e.g., McNamara et al., 2013) and lexical sophistication (e.g., Crossley et al., 2017). Critically, Crossley et al. (2017) did not define what lexical sophistication is, but they did argue that lexical sophistication can be measured using a tool which uses 424 indices for assessing lexical sophistication (see Kyle & Crossley, 2015; Kyle, Crossley, & Berger, 2018). It is therefore reasonable to assume that the readability estimates may be predicted by variation in text features that are not included in their regression formulae. If variation in text features theorised to predict comprehension, such as cohesion and coherence (e.g., Kintsch, 1988;1998; Ozuru et al., 2009), was found to predict readability estimates, the evidence base for assessing text comprehensibility using readability formulae would be strengthened. In addition, more evidence in favour of equating readability to comprehension (e.g., Beck et al., 1991; Flesch 1948) would be generated. I discuss the results of my analyses next.

5.3. Results

First, I discuss the distribution of scores because RQ5.1 asks about the readability of health-related texts, given the estimates of the readability formulae. Second, I briefly discuss the correlations between text feature predictors to examine whether some of these predictors could be used as proxies for other predictors (e.g., Flesch, 1948). Third, I interpret the estimates calculated by the Bayesian models to answer RQ5.2 and RQ5.3. Lastly, I discuss the Bayesian models building process.

5.3.1. Descriptive Statistics

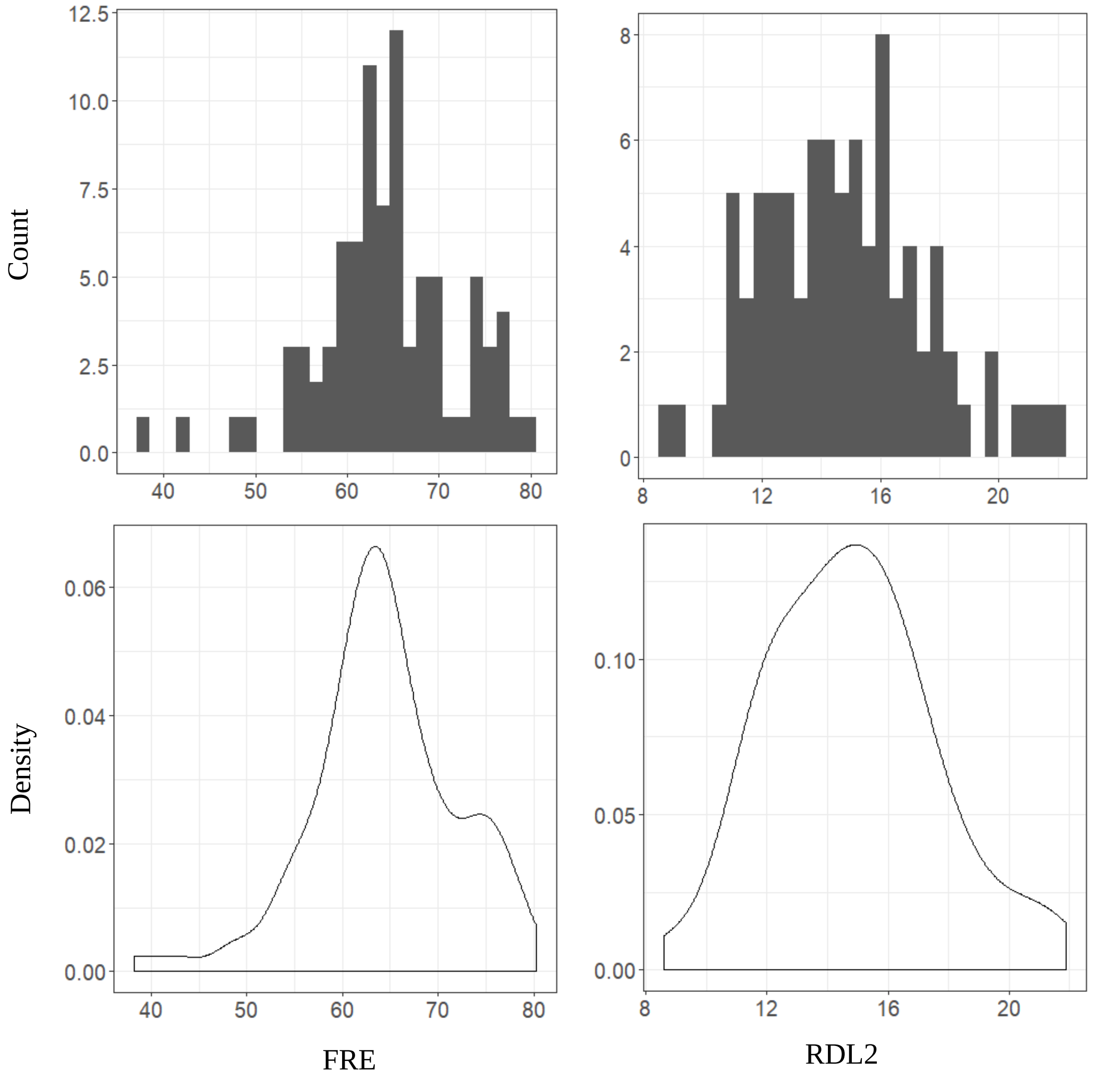
The readability level of sampled health-related texts varied depending on the readability formula used. According to the mean FRE score (Table 5.2), the average leaflet used in this study did not require a reading ability beyond that expected of a 15-year-old (FRE 60-70) (Flesch, 1948; Patel et al., 2013). However, the range of scores in Figure 5.2 and Table 5.2 demonstrates that there was considerable variation in readability levels of health-related documents used in this study. Some texts could be classed as being difficult to read and requiring degree level education (FRE 30-49) to understand, whereas others as being fairly easy to read (FRE 70-79) or easy to read (FRE 80-89) and requiring reading ability of a 12-13-year-old or 11-12-year-old respectively (Patel et al., 2013).

Table 5.2. Readability of a sample of health-related documents.

	FRE (0-100)	RDL2 (0-30)
Mean	64.30	14.80
Lowest	38.17	8.57
Highest	80.30	21.90

Note. The lower the score of each readability formula, the lower the readability level.

Figure 5.2. Histograms and density plots of the FRE and the RDL2 respectively.



In terms of the RDL2 readability scores, on average, health-related texts could be rated as requiring an intermediate level (RDL2 12.90-19.95) of English language proficiency to understand (Crossley et al., 2011). However, the range of scores (Figure 5.2 and Table 5.2) shows that some health-related texts were rated as requiring advanced English language proficiency (RDL2 0-12.90), whereas others as requiring beginner level English language proficiency (RDL2 19.95-30). Overall, the distribution of RDL2 and FRE readability scores (Figure 5.2) demonstrates that not all texts produced by the NHS's health-related information writers, are likely to be rated as highly readable by readability formulae (RQ5.1).

5.3.1.i. Probability Distribution of Outcome Variables

The function and the shape of the distribution of outcome variables is important as it determines the model class that should be used to fit the data. Different model classes have different probability distributions. These probability distributions are chosen for different types of data so that a model can be supposed to adequately approximate the data generating mechanisms. The density plots for RDL2 and FRE outcome variables (Figure 5.2) show that the probability distributions underlying each one of these measures in the given sample of 86 health-related texts can be argued to be approximately normal. Nevertheless, based on the visual inspection of these plots, some may argue that there is some deviation from the "approximate normality" of the RDL2 and FRE distributions, due to slight skew. Typically, researchers would mostly assume normality, whether appropriately or not, especially in frequentist analyses, but this is not something that has to be done in Bayesian analysis. Instead, changing the probability distribution of outcome variables, as well as of the priors, is relatively easy in Bayesian analysis (Kruschke & Liddell, 2018b).

If one assumes that the probability distribution that best describes the outcome variable is skewed, one can assign a distribution that will model this skew better than the normal distribution. One of such plausible distributions is the Skew-Normal distribution. Using the Skew-Normal distribution avoids the need for transformations of non-normal data to make the distribution of the outcome variable appear normal to meet model's assumptions.

This is beneficial, as transforming data prior to analyses can alter inference to such an extent that the transformed outcome variable will meaningfully and erroneously change predictions and estimates (see Martin & Williams, 2017 for a discussion).

Given that it can be argued that at least two probability distributions can describe the data generative process best, I modelled variation in readability using a set of models with normal and Skew-Normal distributions. Fitting a series of models, rather than the more usual one or two models, enables the examination of the robustness or sensitivity of the estimates to different choices of the outcome distribution. The motivation underlying sensitivity analyses is to demonstrate a lack of sensitivity to different perspectives of modelling the data, or to different model specifications. This checks whether model predictions hold under a different set of assumptions. If the inference does not change with different model specification or under a different set of assumptions, the results are thought to be relatively robust. Critically, in addition to choices related to the distribution of the outcome variable, decisions made regarding predictor variables can also influence the inference and form part of sensitivity checks. For example, collinearity and multicollinearity issues can influence the choice of predictors included in the model; therefore, I discuss these concepts and correlations next.

5.3.1.ii. Correlations

Table 5.3 shows the correlations between the different linguistic features and readability formulae estimates. I describe selected correlations that are substantially supported by the data (significant in frequentist terms) next. The RDL2 scores of health-related texts were found to positively correlate with referential cohesion and word frequency. It is important to note that word frequencies were calculated using the CELEX database (Baayen et al., 1995), as the Coh-Metrix tool estimates average word frequency of texts using the CELEX word database (see section 5.2.1 for a more detailed description of Coh-Metrix). These correlations suggest that as texts became more coherent and the incidence of relatively frequent words increased, the readability estimates provided by RDL2 were also likely to increase. These correlations were expected as the RDL2 incorporates an index used to calculate referential cohesion, namely content word overlap, and word frequency in its regression formula. However, it is unexpected that the correlation between syntax similarity and RDL2 estimates was found not to be substantially supported by the data, as syntax similarity is included in RDL2's regression formula.

In addition, verb hypernymy was found to be negatively correlated with the RDL2, possibly due to its correlation with an index of word frequency used in the RDL2 formula (Table 5.3). In the Coh-Metrix tool (Graesser et al., 2004), higher values of hypernymy indicate that average words in the text have higher levels of specificity whereas lower values indicate that average words in the text have lower levels of specificity (McNamara, et al., 2013). Thus, the correlations between verb hypernymy, word frequency, and RDL2 estimates can be interpreted as showing that a high incidence of hypernym verbs was found to be associated with a high incidence of relatively frequent words and high readability values. Hypernym verbs are verbs that denote a class under which sub-categories are subsumed, whereas hyponym verbs are verbs that constitute a sub-category of that class. Thus, hypernyms are broader in meaning than hyponyms, and are possibly more frequently used in the English language than hyponyms (cf. Crossley et al., 2012). For example, the verb *get* is a hypernym of *inherit*, *buy¸* and *find*. This is because *inherit*, *buy*, and *find*, are more specific in meaning than *get*, and can be included within the meaning of *get*.

In comparison to textual features correlated with the RDL2, the FRE estimates were found to be negatively correlated with word length, sentence length, and the incidence of passive voice (I have omitted the correlation between the FRE estimates and word length from Table 5.3, I explain why in the next paragraph). The finding that as the FRE estimates increase, so does the proportion of short words and short sentences in the text is expected as the FRE regression formula is based on indices of word and sentence length. From the perspective of the writing guidelines (e.g., Plain English Campaign, 2018) it is interesting that as the incidence of passive voice increases, the FRE scores decrease. This supports the guidelines of the Plain English Campaign (2018), with regards to the preference for the active versus past tense, because the less frequent passive voice use is associated with higher text readability.

As mentioned previously, I omitted word length from Table 5.3 as it was highly correlated with the FRE (*r* = - .95). I also excluded BNC (BNC Consortium, 2007) average word frequency measure as it was highly correlated with the CELEX (Baayen et al., 1995) average word frequency measure (*r* = .78). I omitted these two variables on the grounds of model parsimony, in the first instance, and collinearity in the second instance. Collinearity occurs when pairs of predictors are so strongly correlated that the model cannot determine which predictors explain the variation in the outcome variable, as the same part of the variance in the outcome variable is being captured by more than one predictor variable (Baayen, 2008). Collinearity is problematic as it leads to relative unreliable and unstable estimates of the coefficients (Dormann et al., 2013).

There are several approaches to diagnosing collinearity, such as by looking at what are perceived to be high correlations, and the measurement of the distortion to standard errors associated with each variable (Dormann et al., 2013). I adopted the commonly used thresholds of correlation coefficient .7, and the square root of the variance inflation factor (VIF) value of 2, to diagnose collinearity throughout this thesis (Dormann et al., 2013). The square root of the VIF indicates by how many times is the standard error for a coefficient as large as it would have been if that predictor were uncorrelated with other variables (Kobacoff, 2011). Given that correlation thresholds for diagnosing collinearity are relatively arbitrary (Dormann et al., 2013), I used the VIF to assess distortion to standard errors when correlations between variables only just exceeded the correlation threshold of .7. I discuss the models and the model-based-predictions next.

Table 5.3. Correlations between linguistic features and readability formulae scores of health-related texts.

	RDL2	FRE	Referential cohesion	Causal connectives	Word frequency	Sentence length	Passive voice	Syntax similarity	LSA	Causal cohesion	Hypernymy noun	Hypernymy verb	Logical connectives
FRE	.18												
Referential cohesion	.29**	.21											
Causal connectives	.21	.03	-.12										
Word frequency	.54***	.04	.38***	.21									
Sentence length	.04	-.22*	.49***	-.01	.48***								
Passive voice	-.02	-.28**	.18	.15	.21*	.33**							
Syntax similarity	.17	.15	-.35***	.14	-.31**	-.62***	-.18						
LSA	-.07	-.17	.47***	-.24*	-.16	.28**	.17	-.14					
Causal cohesion	.20	.01	.05	.51***	.30**	.28**	.19	-.29**	-.14				
Hypernymy noun	-.18	.09	-.03	-.18	-.06	.00	-.01	.14	-.03	-.17			
Hypernymy verb	-.23*	.07	-.05	-.11	-.29**	-.11	-.15	.07	-.04	-.14	-.04		
Logical connectives	.10	.16	-.17	.45***	.23*	.19	-.10	.09	-.14	.16	.02	.00	
Gerunds	-.11	-.19	-.22*	-.19	-.28**	-.15	-.18	.02	-.01	-.30**	.11	.14	.03

Note. Significance values are based on Pearson's *r*. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

5.3.2. Bayesian Models

In Bayesian linear models, I examined the effects of texts features on changes in the FRE and RDL2 readability scores. The predictors included: referential cohesion, the incidence of causal connectives, average word frequency for all words as measured using the CELEX corpus (Baayen et al., 1995), sentence length, the incidence of passive voice forms, syntax similarity, sentence semantic overlap as measured using the LSA, causal cohesion, hypernymy, the incidence of logical connectives, and the incidence of gerunds. Overall, I analysed 86 observations — one observation per text — using Bayesian linear models, fitted with the brm function of the brms package (Bürkner, 2017; 2018) in R (R Core Team, 2019).

5.3.2.i. Prior Distributions

Throughout the studies in this thesis I decided to use weakly-informative regularising priors to improve computational stability by giving the model enough information to avoid inappropriate inferences while allowing for a relatively large amount of variation in the effects of the estimates (Depaoli & van de Shoot, 2017; Gelman & Henning, 2017). Weakly-informative regularising priors permitted to find small to large effects but made it difficult to find relatively implausible effects without massive support from the data.

The prior distributions for the intercept of outcome variables in the FRE and RDL2 models assumed that values closer to the mean are more likely than those further away from the mean, but they were flexible enough to permit values within the possible range of both readability formulae. Given that the range of possible FRE scores is from 0 to 100, the intercept for the FRE model was assigned a normal prior distribution, with a mean of 50 and standard deviation (*SD*) of 50. The mean of 50 was chosen because without looking at the data, it was more plausible to assume that the mean readability score would be closer to 50 than to, for example, zero or 100. Since RDL2 scores range from 0 to 30, the intercept for the RDL2 model was given a normal prior distribution with a mean of 15 and *SD* of 15. The prior distributions for all the linguistic predictors of readability formulae were normal, with a mean of zero and *SD* of 10. The mean was at zero, because the effect could be close to null. The predictors were scaled by two *SDs* because this allows for generic comparisons with other predictors and is thought to guard against understating the effects of predictor variables on the outcome (Gelman, 2008).

Overall, I did a series of analyses which I discuss in the next section. First, I report the final analysis then I report the checks and sensitivity analyses I ran, in turn, to see if estimates differed due to different model specifications. I start with the model predicting linguistic features that have plausible effects on the FRE readability estimates of health-related texts, then I look at RDL. I compare the estimates from the models in the discussion section.

5.3.2.ii. FRE Models

I built a series of models to answer RQ5.2 and RQ5.3. I present the summary of the final model showing the plausible effects of the potential predictors of FRE scores of health-related texts in Table 5.4. Figure 5.3 shows spaghetti plots of the probable effects of six predictors of the FRE. Each one of the lines of spaghetti plots represents one possible prediction for the effect of each predictor. The most probable estimate, or the best guess at what the effect is overall, is indicated by the black line of each plot. Table 5.4 presents 95% credible intervals (CIs), Bayesian counterparts to frequentist confidence intervals. CIs are different from the frequentist confidence intervals, as we are not looking at significance, but considering the relative plausibility of estimates, hence, meaningfulness.

It is important to note that the FRE scores must decrease with sentence length, given the parameterisation of the FRE regression formula. Plausible effects of any other predictors are problematic, because in theory, nothing else should predict changes in the FRE scores. In practice, Table 5.4 shows that there was an effect of referential cohesion on FRE scores such that, for each unit increase in cohesion, FRE scores were predicted to increase between 5.16

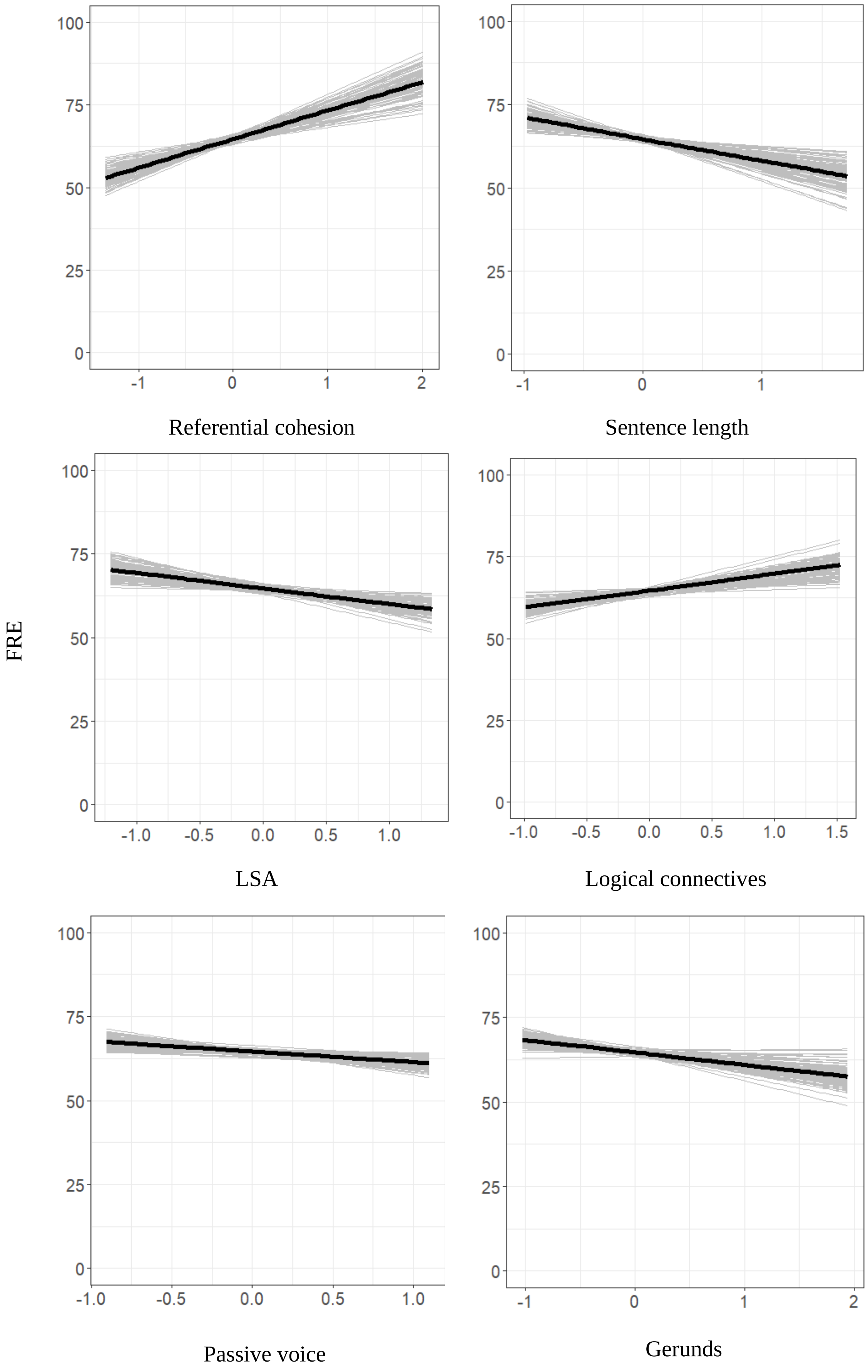
and 12.34 points (on average the increase in the FRE was predicted to be 8.75). Similarly, for each unit increase in the incidence of logical connectives, FRE readability was predicted to increase on average by 4.91 (95% CIs [1.64, 8.13]). Thus, the higher the referential cohesion and the higher the incidence of logical connectives, the higher the readability of health-related texts was predicted to be.

There are four text features which were predicted to have a negative plausible effect on readability of health-related texts as judged using the FRE readability formula. First, for each unit increase in sentence length, FRE scores were predicted to decrease by an average of 6.28 (95% CIs [-10.31, -2.22]). Second, for each unit increase in the incidence of passive voice, FRE scores were predicted to decrease by an average of 3.12 (95% CIs [-5.89, -.33]). Third, for each unit increase in sentence semantic overlap, as measured using the Latent Semantic Analysis (see Table 5.1, section 5.2.2), on average FRE scores were predicted to decrease by 4.34 (95% CIs [-7.53, -1.11]). Lastly, for each unit increase in the incidence of gerunds in the text, FRE scores were predicted to decrease by an average of 3.53 (95% CIs [-6.27, -.82]). Overall, the longer the sentences, the higher the incidence of passive voice in the text, the greater the sentence semantic overlap, and the higher the incidence of gerunds, the lower the FRE readability of health-related texts was predicted to be.

Table 5.4. Summary of the final model (FRE 2.2).

Coefficients	Estimate	Est.Error	L-95%	U-95%	Probable (sign)
Intercept	64.60	0.62	63.38	65.82	
Referential cohesion	8.75	1.81	5.16	12.34	(+)
Causal connectives	-0.98	1.81	-4.49	2.57	
Word frequency	-2.00	1.79	-5.50	1.55	
Sentence length	-6.28	2.03	-10.31	-2.22	(-)
Passive voice	-3.12	1.41	-5.89	-0.33	(-)
Syntax similarity	-0.72	1.74	-4.13	2.74	
LSA	-4.34	1.65	-7.53	-1.11	(-)
Causal cohesion	1.32	1.64	-1.90	4.55	
Hypernymy noun	1.94	1.31	-0.66	4.51	
Hypernymy verb	0.82	1.34	-1.82	3.45	
Logical connectives	4.91	1.66	1.64	8.13	(+)
Gerunds	-3.53	1.38	-6.27	-0.82	(-)
Error term (sigma)	5.67	0.48	4.83	6.72	

Figure 5.3. Probable effects of predictors of FRE.



5.3.2.iii. RDL2 Models

The analysis procedure for models with the RDL2 as the outcome variable was identical to that followed in the previous section with the FRE models. I present the summary of the final RDL2 model showing the plausible effects of the potential predictors of RDL2 scores of health-related texts in Table 5.5. Figure 5.4 shows the effects of plausible predictors of RDL2 visually. In the case of RDL2 scores, given the parameterisation of the RDL2 regression formula, nothing but referential cohesion, average word frequency, and syntax similarity should predict changes in the RDL2 scores. Indeed, as referential cohesion, average word frequency, and syntax similarity increased, the RDL2 scores were predicted to increase by an average of 2.32 (95% CIs [1.16, 3.46]), 3.59 (95% CIs [2.42, 4.75]), and 2.40 (95% CIs [1.24, 3.53]) respectively. Therefore, the greater the referential cohesion, average word frequency, and syntax similarity, the higher the RDL2 readability of health-related texts was predicted to be.

There were two predictors which were found to have a plausible negative effect on the RDL2 scores. Specifically, as the incidence of passive voice and noun specificity (noun hypernymy) increased, the RDL2 scores were predicted to decrease by an average of 1.06 (95% CIs [-1.99, -.13]) and 1.03 (95% CIs [-1.88, -.18]) respectively. Thus, the higher the incidence of passive voice and the higher the average noun specificity, the lower the readability of health-related texts as judged using RDL2 was predicted to be. Overall, I found evidence to suggest that only referential cohesion and the incidence of passive voice are likely to predict RDL2 and FRE readability scores of health-related texts (see Table 5.6). I discuss this in section 5.4, but I first describe the model selection process and sensitivity checks of the analyses.

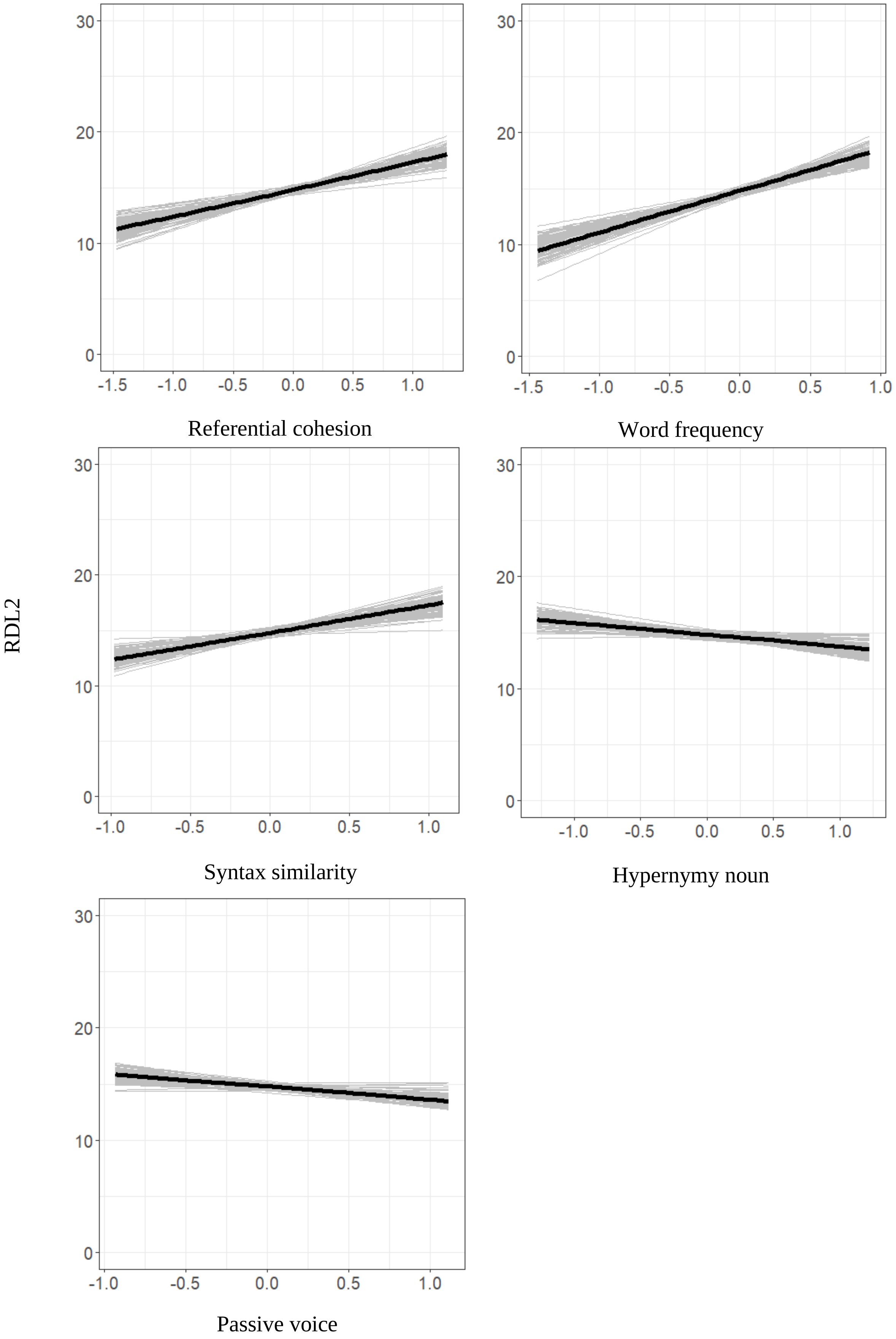
Table 5.5. Summary of the final model (RDL2 2.2).

Coefficients	Estimate	Est.Error	L-95%	U-95%	Probable (sign)
Intercept	14.81	0.20	14.42	15.20	
Referential cohesion	2.32	0.59	1.16	3.46	(+)
Causal connectives	-0.16	0.59	-1.30	1.00	
Word frequency	3.59	0.59	2.42	4.75	(+)
Sentence length	-0.88	0.66	-2.19	0.41	
Passive voice	-1.06	0.47	-1.99	-0.13	(-)
Syntax similarity	2.40	0.58	1.24	3.53	(+)
LSA	0.08	0.55	-1.01	1.18	
Causal cohesion	0.85	0.55	-0.22	1.93	
Hypernymy noun	-1.03	0.43	-1.88	-0.18	(-)
Hypernymy verb	-0.15	0.45	-1.04	0.74	
Logical connectives	0.16	0.54	-0.91	1.22	
Gerunds	0.68	0.46	-0.23	1.59	
Error term (sigma)	1.85	0.16	1.57	2.19	

Table 5.6. Comparison of the final FRE and RDL2 models.

	FRE				RDL2			
Coefficients	Estimate	L-95%	U-95%	Probable	Estimate	L-95%	U-95%	Probable
Intercept	64.60	63.38	65.82		14.81	14.42	15.20	
Referential cohesion	8.75	5.16	12.34	(+)	2.32	1.16	3.46	(+)
Causal connectives	-0.98	-4.49	2.57		-0.16	-1.30	1.00	
Word frequency	-2.00	-5.50	1.55		3.59	2.42	4.75	(+)
Sentence length	-6.28	-10.31	-2.22	(-)	-0.88	-2.19	0.41	
Passive voice	-3.12	-5.89	-0.33	(-)	-1.06	-1.99	-0.13	(-)
Syntax similarity	-0.72	-4.13	2.74		2.40	1.24	3.53	(+)
LSA	-4.34	-7.53	-1.11	(-)	0.08	-1.01	1.18	
Causal cohesion	1.32	-1.90	4.55		0.85	-0.22	1.93	
Hypernymy noun	1.94	-0.66	4.51		-1.03	-1.88	-0.18	(-)
Hypernymy verb	0.82	-1.82	3.45		-0.15	-1.04	0.74	
Logical connectives	4.91	1.64	8.13	(+)	0.16	-0.91	1.22	
Gerunds	-3.53	-6.27	-0.82	(-)	0.68	-0.23	1.59	
Error term (sigma)	5.67	4.83	6.72		1.85	1.57	2.19	

Figure 5.4. Probable effects of predictors of RDL2.



5.3.2.iv. Sensitivity Analyses

The models discussed above were chosen based on their superior predictive performance relative to a series of alternative models that I built. I now describe the steps taken that led to these final models being chosen over a series of alternative models. First, two sets of models with FRE and RDL2 as the outcome variable were fitted using the normal and Skew-Normal distribution (see Table 5.7 and Table 5.8 in Appendix A). As mentioned before, this allowed me to check whether the results were sensitive to the specification of the probability distribution for the outcome variables. Checks demonstrated that the estimates were not sensitive to the choice of distribution. I also checked for influence of individual observations on inference. This was motivated by the desire to have estimates that are grounded in the data in the sample, but not excessively determined by individual observations. If estimates are excessively influenced by individual observations, they may not be replicated in future studies.

I found that the estimates were affected by a particularly influential observation, an observation which a model was failing to adequately predict. Using Pareto-Smoothed-Importance-Sampling (PSIS) algorithm to compute Leave-One-Out Cross-Validation (LOO-CV), a relatively new procedure for, amongst other things, diagnosing influential observations (Vehtari, Gelman, & Gabry, 2017a; 2017b), I found that for some observations, for example observation 47 in the FRE set of models (Table 5.7, Appendix A), the Pareto $\hat{k}$ estimate was higher than the preferred threshold of .7. The Pareto $\hat{k}$ diagnostic measure reveals problems due to posterior distribution's sensitivity to observations. Vehtari et al. (2017a; 2017b) observed that acceptable sampling and convergence rates are achieved below the .7 Pareto $\hat{k}$ values, but above this threshold the performance of the Markov Chain Monte Carlo (MCMC) algorithm, an algorithm that is used to generate the samples from the posterior distribution in Bayesian models, provide unreliable estimates. I briefly explain how

MCMC works before returning to the discussion of the models (a full explanation of the mechanism underlying MCMC is beyond the scope of this thesis, but interested readers can refer to van Ravenzwaaij, Cassey, & Brown, 2018 for an accessible introduction to MCMC).

MCMC combines two concepts: Markov Chain and Monte Carlo (van Ravenzwaaij et al., 2018). Monte Carlo is the practice of estimating the posterior distribution by drawing random samples from it. Markov Chain refers to generating random samples using a special sequential process. Overall, MCMC constructs a Markov Chain to do a Monte Carlo approximation of the posterior distribution, without knowing all the distribution’s properties, by randomly sampling values from the distribution (van Ravenzwaaij et al., 2018). For example, let us assume that we are interested in finding out the mean exam score on a class test. Let us also assume that the exam board specified that the exam scores are normally distributed, and that this distribution has a standard deviation of 15. Let us also assume that one marked exam paper has been leaked and we find out that one student scored 100 on the test. To find the mean exam score, the MCMC algorithm would draw samples from this normal distribution of plausible scores to estimate the mean given a single observation of 100. The sampling process would start with an initial guess, the first sample, for what a plausible value might be given the score of 100 and standard deviation of 15, for example 110. From this first sample the next value would be sampled after adding some uncertainty to the value from the preceding sample. For example, if the first sample resulted in a score of 110, the second sample could be 108. If the second sampled value would be deemed as a plausible exam score given what we know about the posterior distribution, the second sample would be retained in the MCMC chain and another score would be sampled after adding some uncertainty to the score of 108. However, if the second guess would be implausible it would be discarded, and the second sample would just be a copy of the first sample. Once a sufficient number of values would be sampled, the mean score and the whole posterior distribution would be approximated.

The MCMC approach may seem counter-intuitive, but frequently in modelling calculating the mean number of samples is easier than calculating the mean from the distribution using equations (van Ravenzwaaij et al., 2018). This is because we often have partial information about any given distribution. For example, in the models described in Chapters 6 and 7, many model parameters were not known in advance. In such instances, MCMC is necessary to estimate these parameters. However, it is important to mention that using MCMC to sample from the posterior distribution in the context of mixed-effects models is not straightforward. One reason for this is because the efficiency of the sampling process is affected by correlations between model parameters (van Ravenzwaaij et al., 2018).

In practice, high correlations between model parameters increase the likelihood of generating implausible proposal values, meaning that the number of plausible independent samples can be relatively low if a lot of samples are discarded. As mentioned before, to approximate a posterior distribution reliably, a sufficient number of samples is required (what constitutes sufficient is subjective and there is no agreed number that can be generalized to different analyses). One way of increasing the number of samples, is to run multiple Markov Chains (van Ravenzwaaij et al., 2018). Using multiple chains instead of one, starts with multiple initial guesses of plausible values and generates one Markov Chain of samples from each initial guess. Consequently, the total number of plausible samples drawn is higher when using multiple chains versus a single chain to approximate a posterior. This motivated the use of multiple MCMC chains in the analyses employed in this thesis. I now return to describing the model selection procedure.

To test for sensitivity of estimates, I ran the initial set of candidate FRE and RDL2 models without influential observations. I found that for both RDL2 and FRE models, the effect of the incidence of passive voice on readability estimates was found to be sensitive to the presence of influential observations. Specifically, the chosen RDL2 and FRE models without influential observations predicted a probable effect of the incidence of passive voice on readability, whereas the models with the influential observations did not (see Tables 5.7 and 5.8 in Appendix A). However, as I describe below, from the set of considered models, the models without influential observations were most likely to provide the most accurate representation of the estimated effects.

The chosen FRE and RDL2 models, models which I believe to be most representative of reality, were chosen based on their LOO Information-Criterion (LOOIC). LOOIC is an estimate which is used to compare models in terms of their estimated out-of-sample predictive accuracy (see Tables 5.7 and 5.8 in Appendix A for LOOIC values of considered models). The closer to negative infinity the LOOIC estimate of a given model, the better that model is at predicting what might be happening in the real-world compared to a different model with a higher LOOIC estimate (Martin & Williams, 2017). Tables 5.7 and 5.8 (in Appendix A) show that FRE and RDL2 models 2.1 had the lowest LOOIC estimate of the considered models, suggesting that these models performed better than the other models at out-of-sample predictions, approximated the real-world better than the rest (Martin & Williams, 2017).

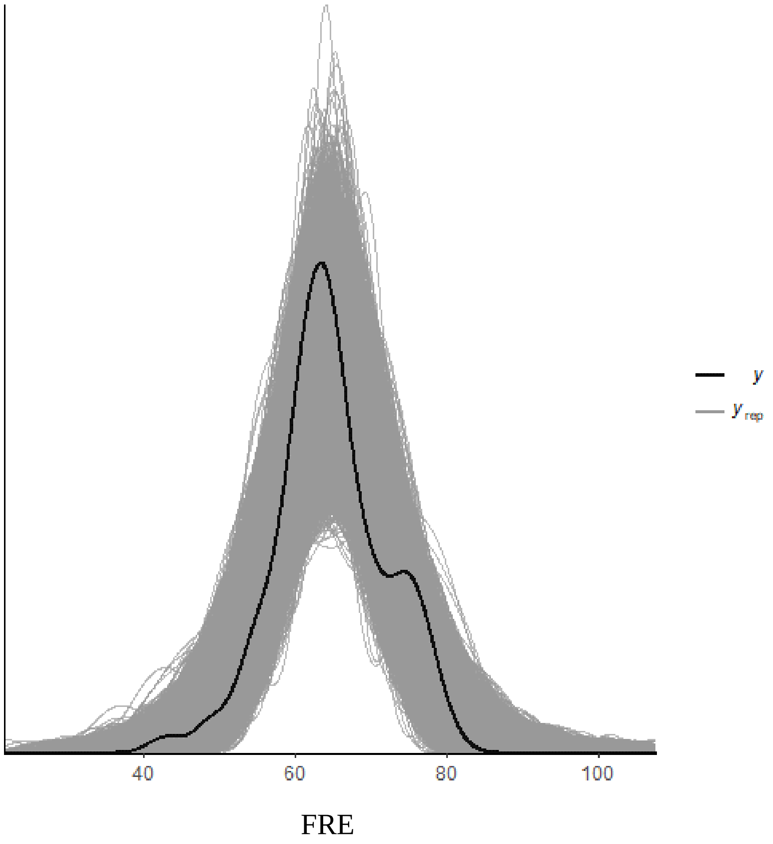
Referring to Tables 5.7 and 5.8 (in Appendix A), there are additional three columns which require further explanation. The “Chains” column refers to the number of MCMC chains that were run during the analysis for each model. All the models in these tables were running 6 MCMC chains, and in all but two cases the number of iterations was 4000 per chain. The “Highest $\hat{R}$” column indicates convergence of the MCMC chains. $\hat{R}$ is a diagnostic measure of Bayesian chains convergence; if the chains have converged to a common distribution then the $\hat{R}$ statistic will be 1.00. However, if the chains have not converged to a common distribution, the $\hat{R}$ will be greater than 1.00 (Gelman et al., 2013). In Tables 5.7 and 5.8 (in Appendix A) we see that $\hat{R}$ values of all candidate models are 1.00, indicating convergence.

Examining FRE and RDL2 models, I also considered LOO-adjusted predicted R^2 instead of traditional R^2 (see Tables 5.7 and 5.8 in Appendix A). This is because traditional R^2 can result in over-fitting, as it is based on the sample underlying the model. Consequently, traditional R^2 increases with the addition of each new parameter to the model, even if the new parameter does not improve the model's predictions. In turn the predicted R^2 indicates how well a model predicts responses for new data. In addition, predicted R^2 guards from over-fitting as parameters that do not improve the model's predictions will lower the predicted R^2. The LOO-adjusted R^2 is a variant of predicted R^2 designed specifically for Bayesian regression models (Gelman, Goodrich, Gabri, & Vehtari, 2018). LOO-R^2 can be defined as an estimate of the proportion of variance explained for new data (Gelman et al., 2018), which is LOO-adjusted for over-fitting. The FRE chosen model, model 2.1 in Table 5.7 (Appendix A), explained 26% of the variance for new data (LOO-R^2 = .26), whereas the RDL2 chosen model, model 2.1 in Table 5.8 (Appendix A), explained 45% of the variance for new data (LOO-R^2 = .45).

As an additional check in the sensitivity analyses, I doubled the number of iterations for each MCMC chain of each chosen model to check for the presence of local convergence. Local convergence occurs when convergence appears to be obtained with a relatively small number of iterations, but when the MCMC chains are running for longer the convergence shifts to another location (Depaoli & van de Shoot, 2017). In practical terms, the presence of local convergence can influence model's estimates. After doubling the number of iterations in the chosen FRE and RDL2 models, I found that the estimates did not change. Thus, I did not find any evidence for local convergence, suggesting that the estimates of the chosen models are relatively stable over the number of iterations.

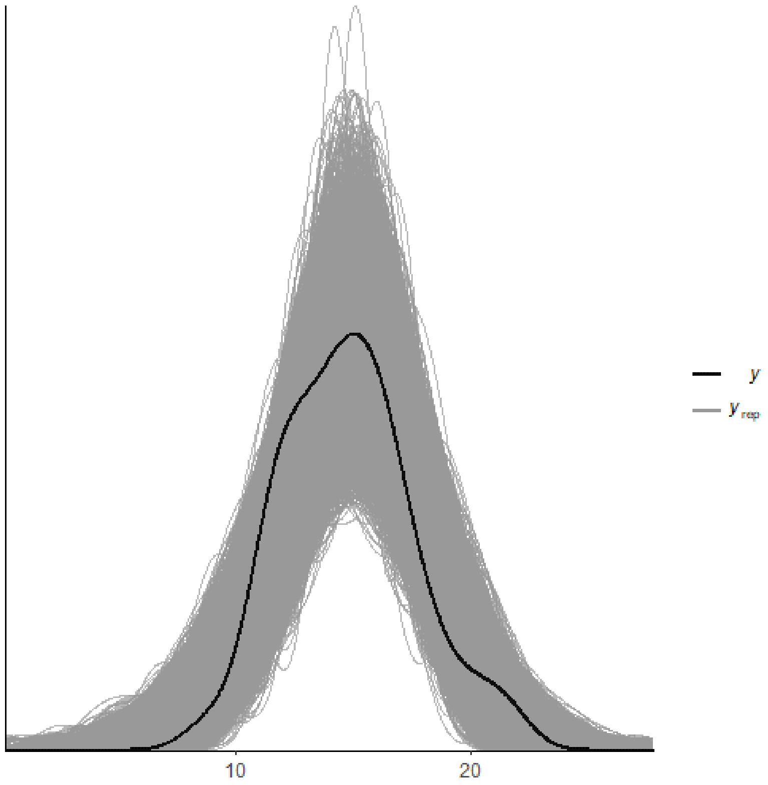
Last, to assess the predictive performance of the FRE and RDL2 models, I used a posterior predictive check (PPC). PPC generates model-implied datasets and shows the degree to which replicate model-implied datasets are similar to the observed data (Gelman, 2003). If a model approximates the data generating process relatively well, the model-implied replicate datasets will closely resemble the observed data (Martin & Williams, 2017). The PPCs (Figures 5.5 and 5.6) show that the final FRE and RDL2 models (models 2.2 in Tables 5.7 and 5.8 in Appendix A) had relatively strong predictive performance as the model-implied replicate datasets closely resembled the observed data.

Figure 5.5. PPC of the FRE model.



Note. Replicate model-implied datasets are plotted in grey and labelled y_{rep}, the observed data is plotted as a black line labelled y.

Figure 5.6. PPC of the RDL2 model.



RDL2

Note. Replicate model-implied datasets are plotted in grey and labelled y_{rep}, the observed data is plotted as a black line labelled y.

5.4. Discussion

In this study, I wanted to compare the distributions of readability scores generated by two readability formulae. I also aimed to examine whether readability scores of different readability formulae would be predicted by the effects of similar or different text features. To answer my research questions, I generated readability scores for a sample of health-related texts. My analyses showed some similarities but, critically, some differences in effects of text features on different readability scores. The differences are problematic as both readability formulae used in this study claim to measure readability (Flesch, 1948; Crossley et al., 2008). I discuss the implications of my findings in the following.

I asked, “What do readability estimates reveal about the readability of health-related texts?” (RQ5.1). I found that there was considerable variation in the estimated readability level of health-related texts, depending on whether estimates were derived using the FRE or the RDL2 formula (section 5.3.1). The distribution of FRE scores suggests that 23.26% of the health-related texts could be classed as being relatively difficult to read, 55.81% could be thought of as being within the average range of text difficulty, and 20.93% could be classed as being relatively easy to read (Flesch, 1948). In comparison with the FRE, the distribution of RDL2 scores indicates that many health-related texts are likely to be difficult to understand for some individuals, such as low-proficiency ESL speakers. This is because 29.07% of the health-related documents used in this study were estimated to require advanced level of English language proficiency to understand, 63.95% were estimated to require an intermediate level, and only 6.98% were estimated to require beginner level English language proficiency to understand (Crossley et al., 2011). Critically, both readability formulae have shown that not all health-related texts had high readability scores (contrary to $H_{5.1}$). This variation in readability suggests that health-related information writers do not or cannot always adhere to NHS guidelines favouring some text features, like a preference for short words or sentences (e.g., Plain English Campaign, 2018), which underlie some readability formulae, such as the FRE (Chapter 3, section 3.2).

In my study I also investigated what linguistic properties of health-related texts contribute to readability scores (RQ5.2). The analysis revealed that, contrary to $H_{5.2}$, readability estimates were predicted by different sets of predictors for different readability formulae, and that the predictors include but are not the same as the variables that specify the readability formulae. The misalignment in terms of predictors of readability may be potentially explained by the fact that the two readability formulae are based on different textual features (refer to Chapter 3, section 3.2). However, it is problematic that the estimates of RDL2 and FRE formulae were found to be predicted by the plausible effects of text features that are not included in the specification of these readability formulae. This is because, in theory, text features that are not part of readability formulae should not predict estimates of these formulae (Section 5.2.2).

One potential explanation for the plausible effects of some text features not included in readability formulae on the estimates of readability of these formulae may be that texts that differ in one text feature dimension, such as word frequency, are also likely to vary in another text feature dimension, such as hypernymy (e.g., Crossley et al., 2012). This variation is illustrated by the correlations reported between some text features in this study (Table 5.3), and the use of some text features as proxies for other text features in the literature (see section 5.2.2 for examples). Thus, instead of variation in a single text feature, what may matter is textual variation on an underlying latent dimension that might be manifested overtly in variation some text features, such as word frequency, but also to a lesser extent in other measured variables, such as hypernymy (cf. Crossley et al., 2012). This may explain why some variables that are not included in the readability formulae, such as hypernymy, predicted variation in readability. Specifically, some variables, such as hypernymy, might be a weaker proxy than variables such as, word frequency, for the latent construct which is theorised to matter to comprehension, such as language exposure (e.g., Brysbaert et al., 2016). (I discuss the effects of plausible text feature predictors on readability estimates next).

Another construct which is thought to matter to comprehension is text coherence (e.g., van Dijk & Kintsch, 1983) (Chapter 1, sections 1.5 and 1.6). Indeed, research evidence suggests that text coherence, as measured using referential cohesion index of Coh-Metrix (Graesser et al., 2004), could be a plausible predictor of comprehension in general (e.g., Kulesz et al., 2016), and the comprehension of health-related texts (e.g., Liu et al., 2009). Therefore, it is reassuring to see that referential cohesion predicted variation in readability scores generated by both readability formulae. Furthermore, the finding that referential cohesion predicted readability estimates flags it as a potentially important predictor of both the estimated readability and the actual (tested) comprehension of health-related texts. One more potentially important predictor of readability and comprehension is the incidence of passive voice forms. Indeed, the plausible effects of the incidence of passive voice on both readability formulae estimates are supported by some research findings (Crossley et al., 2007; Riche et al., 1991), and some of the guidelines adopted by the NHS (e.g., Plain English Campaign, 2018).

The remaining effects of other plausible predictors discovered in this study varied depending on the readability formula used. For example, the greater incidence of logical connectives was associated with higher estimates of the FRE, but not RDL2, readability. This is problematic, because there is evidence to suggest that logical connectives are a relatively good indicator of difficulty level of texts aimed at ESL learners (e.g., Crossley et al., 2012; Green et al., 2013). Furthermore, text features which influence cohesion, such as the incidence of connectives, are thought to influence the formation of the textbase (Kintsch & Rawson, 2007) and of a situation model (Dowell et al., 2016). However, it might be the case that logical connectives play a less important role in readability of health-related texts than other texts such as narratives. This is because inference-making is necessary for the purpose of understanding implicitly stated information in narrative texts (Graesser et al., 1994), but informational texts might require more literal comprehension to understand explicitly stated information (Silva & Cain, 2015). Thus, it might be the case that in comprehension of health-related texts logical connectives are less important than in comprehension of narrative texts as all the information in health-related texts should be explicit.

In addition to differences in the plausible effects of the incidence of logical connectives on readability scores between the different readability formulae, there were other differences which present interesting contrasts. One such difference, was the word frequency effect. Specifically, higher average word frequency estimates were predicted to be associated with higher RDL2 scores, but not with higher FRE scores (Table 5.6). This is interesting because research evidence indicates that frequent words are processed more quickly than infrequent words (e.g., Brysbaert et al., 2016; Diependaele et al., 2013; Kuperman & Van Dyke, 2013), are perceived (Riche et al., 1991) and understood better than less frequent words (Leroy & Kauchak, 2014), and that simpler texts tend to have a higher proportion of frequent words than more complex texts (Leroy et al., 2013). Consequently, due to the lack of association of the FRE estimates with variation in word frequency, the validity of the FRE as an effective measure of readability of health-related texts can be questioned.

The construct validity of the FRE scores as measure of readability for health-related texts is further undermined by the finding that higher semantic overlap was associated with lower FRE readability. Specifically, the models supported an effect of sentence semantic overlap in the FRE data, such that greater overlap was associated with lower readability, but there was not substantial evidence for a similar influence on readability in the RDL2 data. These findings are problematic for both readability formulae, but more so for the FRE, as they are inconsistent with evidence suggesting that higher sentence semantic overlap is found in simpler texts compared to more difficult texts (e.g., Crossley et al., 2007). Another difficult to explain finding in relation to the FRE is that increasing syntax similarity was associated with higher readability in the RDL2, but not in the FRE. This is problematic as, assuming that syntactic similarity refers to equally simple syntax used throughout the text, rather than equally difficult syntax (Dowell et al., 2016), the effects of high syntax similarity should predict high readability. This is because texts that have sentences with similar syntactic structures are thought to lower the cognitive demands imposed on the reader when reading

(Crossley et al., 2011), meaning that the reader can concentrate on building a logical situation model.

My research also sought to answer the question whether the recommendations of NHS's guidelines relating to specific linguistic features are associated with higher readability levels of health-related texts (RQ5.3). There is some evidence to suggest that following existing NHS writing guidelines might result in a higher readability of health-related texts ($H_{5.3}$). First, the usage of passive voice forms was predicted to decrease readability in both the FRE and the RDL2 data. This is in line with the Plain English Campaign (2018) guidelines, which are endorsed by NHS England (2018a; 2018b), where writers are advised to use active verbs over passive verbs. Second, the variation in the readability scores derived using the RDL2 formula was predicted by average word frequency, noun hypernymy, and syntax similarity. These linguistic features could be argued to be proxy measures of NHS's recommendations to use straightforward words (Marsay 2017a; 2017b), plain language (NHS England, 2018a), and simple words (NHS England, 2018b). Thus, the plausible effects of word frequency, hypernymy, and syntax similarity on readability may be consistent with NHS's recommendations to favour the use of simple words and plain language (NHS England 2018a; 2018b).

However, contrary to $H_{5.3}$, there is also some evidence to suggest that following existing recommendations of NHS guidelines relating to specific linguistic features will not improve the readability of health-related texts as measured using different formulae. For example, the Plain English Campaign (2018) guidelines embody the view that longer words and sentences (consistent with the parameterisation of the FRE readability formula), and the use of constructions like gerunds, will make texts harder to understand. In this study, the FRE readability scores were found to be negatively predicted by sentence length and the incidence of gerunds, suggesting that the Plain English Campaign guidelines might be effective in improving FRE readability levels of health-related texts. However, there was no substantial evidence to suggest that sentence length predicts RDL2 readability of health-related texts. This indicates that some guidelines to which NHS information writers may be directed to, such as the need to avoid long words (Plain English Campaign, 2018), might only predict readability of health-related texts as assessed using a particular readability formula. This is problematic as without testing actual comprehension, we do not know which readability formula predicts comprehension or could be used as a proxy of tested comprehension. Consequently, further research is needed to examine whether the NHS's recommendations (e.g., NHS England, 2018a) are influential when it comes to predicting (or promoting) the comprehension of health information texts.

5.4.1. Limitations

Although the sample of texts analysed in this investigation was not small compared to studies looking at comprehension of health-related texts (e.g., Liu et al., 2009), it was relatively small compared to some corpus studies (e.g., Crossley et al., 2011). One reason for having a relatively small number of texts is that during the sampling period for this study, not all NHS Trusts published their health-related texts online and the sample and variety of the health-related documents that were published by NHS Trusts were limited. Consequently, the number of health-related texts used in this study might not have been sufficient to enable robust or precise estimates of all potential effects of linguistic features on readability scores. Nevertheless, it was probably sufficient to detect the effects of theoretical interest, such as those of text coherence. To improve on this research, future investigations should aim to reduce the sensitivity of the readability model estimates to influential observations by using a larger sample of texts.

5.4.2. Implications

As the readability scores derived using the two different formulae appear to be influenced by different sets of predictors, it can be concluded the readability formulae reflect different aspects of linguistic basis of readability. This is expected to the extent that each formula relies on a different set of linguistic features and calculations. But it is important that the evidence shows that variation in readability estimates is influenced by contrasting linguistic features when scores are derived from different formulae. More strikingly still, the correlation of the scores generated by different readability formulae was very low (*r* = .18): so low that the existence of a correlation between readability scores was not substantially supported by the data. From a purely statistical perspective, in the context of the predictors used in this study, the RDL2 model can be argued to be better than the FRE model, because it accounted for more variance in readability. Theoretically, the predictions made by the RDL2 models also seem to be more grounded in reading comprehension theories (e.g., Kintsch & Rawson, 2007; Perfetti, 2007). However, perhaps the key message here is that the readability scores derived from two widely used readability formulae present divergent estimates of the readability of a sample of health-related texts. This is problematic as it indicates that RDL2 and FRE measure a different construct when it comes to the comprehension of health-related information. The practical implication is that writers who use different readability formulae will produce health-related texts of differing readability.

Overall, the analyses performed in this study examined the factors that influenced the estimated readability of health-related texts, as assessed using two readability formulae, not the measured comprehension of these texts by people. The estimates may be claimed to predict actual comprehension (RDL2: Crossley et al., 2008) (FRE: Flesch, 1948), but we do not know, yet, how close the association between variation in readability scores and variation in comprehension is. This is largely because these formulas were validated against one dataset, and it is assumed that they apply to new data, but they may not (Chapter 3, section 3.2). If the association between readability and tested comprehension were found to be strong then the estimates of the influence of linguistic features to promote or diminish readability scores would enable us to predict variation in comprehension of health-related texts. However, that association is open to question because the formulae appear not to measure the same construct, and it is not known whether variation in estimated readability predicts the perceived or tested comprehension of health information (e.g., Kauchak & Leroy, 2016; Leroy & Kauchak, 2014) (Chapter 3, section 3.2.1). I attempted to address this question in the following two studies of this thesis (Chapters 6 and 7).

Chapter 6: Examining the Relation between the Perceived Comprehension of Health-Related Information and Textual Measures of Readability

This chapter concentrates on judgements of comprehension of health-related texts (perceived comprehension or metacomprehension) and reports the second study included in this thesis. Study 2 aimed to investigate how variation in reader attributes and text readability scores, specifically the Flesch Reading Ease (FRE; Flesch, 1948) and the Coh-Metrix L2 Readability Index (RDL2; Crossley et al., 2008), predicts perceived comprehension of health-related information. A second aim of the study was to examine whether the effects of individual differences interact with the effects of variation in readability in predicting perceived comprehension of health-related texts. To justify these aims, this chapter starts with a short literature review that builds on the metacomprehension research discussed in Chapter 2 (section 2.2). The literature review is followed by a method and results sections, based on which the research questions are answered in the discussion section. The chapter ends with a discussion of the implications of the findings for the National Health Service (NHS).

6.1. Literature Review

The guidelines adopted by the NHS health-information writers encourage the production of texts with certain features, such as short words and sentences (e.g., Plain English Campaign, 2018). However, it is not known whether those features relate to estimates of perceived comprehension (metacomprehension) or to actual, that is, to directly tested comprehension. Some of the recommendations that NHS health-information producers follow, such as the preference for short words and sentences, are captured by text-feature-based estimates of readability, for example the FRE (Flesch, 1948). However, it is not known whether variation in estimated readability predicts the perceived or tested comprehension of health-related information (e.g., Kauchak & Leroy, 2016; Leroy & Kauchak, 2014). Critically, the NHS is reliant on perceived comprehension measures rather than tested comprehension of health-related texts, as it uses the evaluations of reader panel members to ensure that health-related documents are easy to understand (NHS England, 2018a; see also section 4.1 of Chapter 4). Thus, it is important to investigate the factors that influence both perceived and actual comprehension of health-related information.

One potential concern associated with using patient reader panel members to evaluate the comprehensibility of health-related texts is that it is not clear what metacomprehension judgements of health-related texts are based on. Furthermore, we do not know whether the potential reader characteristics and text features that predict metacomprehension would also predict tested comprehension. If similar effects predict both metacomprehension judgements and actual comprehension, then metacomprehension judgements might be relatively good proxies of tested comprehension. In addition, we do not know whether health-related texts would be equally comprehensible (in perception or in actuality) for individuals who differ from patient reader panel members, for example, in age, educational background, and health literacy. If the effects of variation in reader characteristics do not modulate metacomprehension judgements, then the use of reader panel members might provide a sufficient evaluation of the comprehensibility of health-related texts for the wider population.

Metacomprehension judgements are theorised to be important, in the context of comprehending health-related texts, as they are thought to affect comprehension by contributing to whether individuals engage in specific reading behaviours that regulate comprehension breakdowns (Thiede et al., 2010) (Chapter 2, section 2.2). These self-regulatory reading behaviours typically include strategies aimed at improving understanding of the read material, such as rereading texts which were not understood at the desired level of understanding in the first instance (Thiede et al., 2003; 2010). Thus, metacomprehension judgements could influence NHS patients' motivation to reread health-related texts that they did not understand the first time they read them, thereby affecting comprehension of health-related texts. However, as discussed in Chapter 2 (section 2.2), engaging in such strategies is likely to be dependent on readers' standards of coherence (van den Broek & Helder, 2017), and on other reader characteristics such as background knowledge (O'Reilly & McNamara, 2007).

Due to their relatively frequent exposure to health-related documents (Burrow & Forrest, 2015), it is reasonable to assume that reader panel members' background health knowledge is likely to be higher compared to that of a typical NHS patient. Health knowledge is a component of functional health literacy (Chin et al., 2011), consequently reader panel members are also likely to be more health literate than typical NHS patients. This matters as variation in relevant background knowledge is thought to predict metacomprehension judgements about the texts that are read (Chapter 2, section 2.2). Specifically, there is evidence to suggest that metacomprehension judgements of individuals with higher levels of relevant background knowledge are more accurate at predicting their comprehension than metacomprehension judgements of those with lower levels of relevant background knowledge (Griffin et al., 2009).

One of the reasons for the probable effects of background knowledge on accuracy of metacomprehension could be that high-background-knowledge readers may be more likely to be skilled readers who in turn are more likely to engage in active processing (e.g., O’Reilly & McNamara, 2007) to self-regulate their comprehension (Thiede & Anderson, 2003; Thiede et al., 2010). Thus, reader panel members may understand health-related texts better and be more aware of their own level of understanding, than other individuals. However, it is questionable whether we can generalise reader panel members’ evaluations of comprehensibility to the wider population. This is because the potentially high self-awareness of the panel members is only likely to be valuable if the metacomprehension judgements of individuals with different health literacy levels are shaped by the same set of textual feature effects in the same ways. Critically, this may not be the case because, in addition to health literacy, individuals also vary on other dimensions that may interact with the effects of text features.

As mentioned in Chapter 2 (section 2.2), it is thought that more educated adults might have higher relative metacomprehension accuracy compared to less educated adults (Zabrucky et al., 2012). This is because there is evidence to suggest that the more educated adults are more likely to evaluate and regulate their understanding of problematic information they read in the text, compared to less educated adults (Zabrucky et al., 2012). Specifically, educated adults’ more frequent detection of inconsistencies across pairs of sentences (Zabrucky et al., 2012) is likely to improve the accuracy of their metacomprehension judgements as it is likely to lead to engagement in active processes (O’Reilly & McNamara, 2007), such as rereading, that regulate understanding. In turn, engagement in active processes is also likely to be helpful to comprehension (O’Reilly & McNamara, 2007).

Critically, if the less educated adults do not engage in active processes to the same extent as the more educated adults (Zabrucky et al., 2012), the metacomprehension accuracy of the less educated adults is likely to be lower than those of the more educated adults. This is a problem if education level predicts comprehension, as the less educated adults may not understand some texts well, but, due to lower engagement of self-regulatory active processes, they may think that their understanding of such texts is better than it is. Therefore, it may be the case that reader panel members' evaluations of texts could be different to those of other individuals from different educational backgrounds.

In addition to health literacy and education, there may be differences in the accuracy of metacomprehension judgements between readers of different ages. Indeed, some suggestive evidence indicates that older readers may be over-estimating their metacomprehension more than younger readers due to potential age-related changes in comprehension monitoring (e.g., Dunlosky et al., 2006; Miles & Stine-Morrow, 2004). However, this evidence is weak, as the findings implying differences in metacomprehension between older and younger adults do not always replicate (e.g., Dunlosky et al., 2006; Lin, Zabrucky, & Moore, 2002; Olin & Zelinski, 1997). Furthermore, differences in metacomprehension ratings between younger and older adults could be reflective of differences in text comprehension due to potential higher levels of relevant background knowledge of older adults (Griffin et al., 2009; O'Reilly & McNamara, 2007). Thus, it is unclear if there is a direct link between age differences and variation in metacomprehension of health-related texts. Examining this potential link is important as reader panel members tend to be older individuals. Critically, the potential effects of variation in age on metacomprehension judgements cannot be studied independently. This is because people of different ages can be of different education background and have different levels of background knowledge. Therefore, what is required is a study where the effects of health literacy, education, and ageing on metacomprehension are measured together.

A study of metacomprehension of health-related texts in the United Kingdom (UK) would be incomplete without a measure of English language proficiency, as a significant minority of the UK's population are foreign born (Office for National Statistics, 2016). Consequently, not all NHS patients are of the same language background, and it cannot be assumed that those from different language backgrounds have the same English language proficiency. This is important because as discussed in Chapter 2 (sections 2.1.3 and 2.2), reading processes and reading strategy use may vary depending on individuals' language background and English language proficiency (e.g., Brysbaert et al., 2016; Hong-Nam & Page, 2014; Kern, 1994). For example, there is evidence to suggest that the more proficient English as Second Language (ESL) readers more frequently monitor, manage, and evaluate their comprehension than the less proficient ESL readers (Hong-Nam & Page, 2014). Thus, advanced ESL readers may be more accurate in terms of their metacomprehension judgements than beginner ESL readers, due to potentially more frequent self-testing of understanding (Thiede et al., 2010). Critically, it is unclear whether monolingual reader panel members' evaluations of health-related texts would match those of ESL readers, since in addition to varying in English vocabulary knowledge (Brysbaert et al., 2016), ESL readers can also vary in education, age, and health literacy levels.

Overall, we may have an idea how some reader characteristics may predict metacomprehension of health-related texts in isolation of other variables, but it is unclear whether these effects are robust in the presence of other potentially relevant individual differences predictors. Therefore, it is important to examine the effects of English language proficiency, alongside measures of health literacy, educational background, and age, on metacomprehension judgements. However, in addition to individual differences, it is also important to consider how variation in the features of health-related texts may predict metacomprehension judgements of different readers. This is because health-related texts vary in their text features, and concomitantly, in their readability levels (e.g., Liu et al., 2009), but we do not know whether the effects of individual differences on metacomprehension are different for texts that vary in their readability levels. Thus, there is a need to investigate how variation in the readability of health-related texts may affect metacomprehension judgements, and how these judgements may generalise from reader panel members to others across different texts.

As mentioned in Chapter 4 (section 4.4) and 5 (section 5.1), readability formulae, such as the FRE (Flesch, 1948), are relatively frequently used to assess comprehensibility of health-related texts (Wang et al., 2013). Furthermore, from the practical perspective, the FRE is a relatively important measure of readability of health-related texts. This is because the indices underlying the FRE, word and sentence length, overlap with some of the guidelines for writing health-related texts that the NHS endorses, such as preference for straightforward words, plain language, and short words and sentences (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018) (see also Table 4.1 in Chapter 4, section 4.4). However, we do not know whether readability formulae, such as the FRE, predict tested comprehension, perceived comprehension, or perceived ease of text processing (e.g., Kauchak & Leroy, 2016; Leroy & Kauchak, 2014; Rawson & Dunlosky, 2002). Thus, it is of theoretical interest to investigate the effects of variation in readability on both comprehension performance (investigated in Chapter 7) and metacomprehension judgements (investigated here).

Although at present we do not know how variation in the text features underlying readability formulae predicts metacomprehension judgements of reader panel members, there is some evidence to suggest that variation in text coherence may predict metacomprehension judgements. Specifically, as discussed in Chapter 2 (section 2.2), individuals were found to judge their perceived comprehension at a higher level, and self-reported exerting less effort on understanding texts, when text coherence was higher (Crossley et al., 2017; Rawson & Dunlosky, 2002). Critically, since text coherence is thought to predict comprehension (e.g., Kintsch, 1998; Kintsch & Rawson, 2007), it may be the case that metacomprehension judgements are a relatively good proxy of comprehension and could be implemented as comprehension performance predictions. However, variation in coherence does not predict comprehension of all individuals in the same way (see Chapter 1, section 1.5), as the effects of coherence vary due to differences in reader characteristics, such as background knowledge (McNamara, 2001; McNamara & Kintsch, 1996; O'Reilly & McNamara, 2007; Ozuru et al., 2009). Consequently, there is a need for a study that examines the effects of a coherence-based readability formula on metacomprehension judgements in interaction with reader characteristics. This motivated the use of the RDL2 (Crossley et al., 2008) as the second text readability measure employed in this study (for a description of RDL2 refer to Chapter 3, section 3.2).

In addition, as argued in Chapter 2 (sections 2.1 and 2.2), it may be the case that metacomprehension judgements are influenced by readers' standards of coherence. Standards of coherence are thought to vary between readers due to individual differences (e.g., Ozuru et al., 2009), and are theorised to be affected by readers' representation of the text read (e.g., van den Broek & Helder, 2017). Readers' representation of the text read is assumed to influence the amount of text processing, including reader-initiated active processing, that readers perceive to be required to understand the text read (van den Broek & Helder, 2017). Critically, readers' representation of the text is thought to be affected by properties of the text (Chapter 1, section 1.2). Thus, if increasing the FRE scores is associated with a reduction in perceived text processing (Crossley et al., 2017), it may be the case that readers are less likely to engage in active processing, such as inference making, to comprehend texts high in FRE. However, texts high in FRE may not be any more coherent than texts low in FRE, as decreasing sentence length may reduce text coherence (e.g., Crossley et al., 2008) (see also Chapter 3, section 3.2). Therefore, high FRE score texts may still require the engagement of active processes to understand these texts, but the high FRE scores may falsely signal to readers that these processes are not necessary (Chapter 2, section 2.2). Consequently, readers' metacomprehension accuracy may be negatively impacted by the representation that texts high in FRE potentially evoke (van den Broek & Helder, 2017).

Importantly, it may be that following some of the guidelines endorsed by the NHS, such as keeping sentences and words short, could give the producers and the readers of health-related texts a false representation of the effort required to understand these texts (Chapter 2, section 2.2). Consequently, it is plausible that some of the guidelines could have a detrimental effect on metacomprehension of health-related texts. Critically, this potential problem is also likely to be exacerbated by the possibility that the effects of text features are likely to have different effects on readers of different backgrounds (Chapter 1, section 1.5). Therefore, it is vital to examine whether variation in readability scores predicts metacomprehension and ease of processing judgements, and how these judgements are affected by readers' backgrounds. This is investigated in this chapter. However, given that there is evidence to suggest that readability scores may predict metacomprehension, but not tested comprehension (Kauchak & Leroy, 2016; Leroy & Kauchak, 2014), it is also important to investigate whether metacomprehension judgements predict actual comprehension of health-related texts. This is examined in the next chapter.

6.1.1. Research Aims

In this study, I aimed to investigate whether health-related texts produced by the NHS are perceived to be understandable, as assessed using judgements of self-rated perceived comprehension and text processing, and whether variation in readers' backgrounds predicts these judgements. I also wanted to examine whether variation in the estimated readability of health-related texts predicts self-rated judgements of comprehension, and whether it predicts these judgements differently for different kinds of readers.

6.1.2. Research Questions

RQ6.1. What is the self-rated understanding of health-related texts used in this study?

RQ6.2. What individual differences and textual measures of readability predict self-rated comprehension of health-related texts?

RQ6.3. Do the effects of textual measures of readability interact with the effects of individual differences to predict self-rated judgements of perceived comprehension?

6.1.3. Hypotheses

$H_{6.1}$. Given that health-related texts are designed to be easy to understand, the self-rated judgements of perceived comprehension should be high.

$H_{6.2}$. Textual measures of readability and individual differences, specifically health literacy, age, education, and English language proficiency, should predict self-rated judgements of perceived comprehension of health-related texts.

$H_{6.3}$. The effects of textual measures of readability are likely to interact with the effects of individual differences to predict self-rated judgements of perceived comprehension.

6.2. Method

6.2.1. Participants

Participants were recruited through an open call on social media platforms and using the Qualtrics application for collecting and analysing data, available through Qualtrics.com. Qualtrics is a well-established crowdsourcing service in which participants anonymously complete surveys or short tests in return for small fees. Qualtrics allowed for efficient data collection with minimal time commitment and costs, and it provided access to a relatively heterogenous sample of participants. English speaking individuals living in the UK were offered £5.00 to participate. Those who agreed to participate were briefed on the nature and the purpose of the study and were then given an opportunity to provide informed consent to participate or to withdraw from the study (see pages 1-4 in Appendix B).

In total, the sample comprised 129 participants (N_{female} = 78, N_{male} = 51) aged 16 to 84 years (M_{age} = 42.59, *SD* = 14.56). Most of the participants (*N* = 69) were native English speakers, 60 were ESL speakers. The ESL speakers were of 28 different first language backgrounds. Most ESL speakers (*N* = 39) self-reported having an advanced English language proficiency, whereas 19 reported being intermediate speakers of English, and two thought that they were at beginner level. In addition to language background, the participants also differed in terms of their educational background. Specifically, 85 self-reported being university educated, 18 reported having completed further education, and 26 finished secondary school only (Table 6.1).

Table 6.1. Participants by English proficiency, education, and age.

Proficiency	Education	Number	Mean age (*SD*)
Native	Higher Education	44	45.32 (13.63)
	Further Education	8	48.25 (13.88)
	Secondary School	17	53.12 (15.11)
Advanced	Higher Education	31	36.61 (10.83)
	Further Education	4	33.50 (7.75)
	Secondary School	4	44.75 (18.20)
Intermediate	Higher Education	10	32.50 (11.94)
	Further Education	5	39 (12.19)
	Secondary School	4	43.25 (9.33)
Beginner	Higher Education	-	-
	Further Education	1	36 (-)
	Secondary School	1	34 (-)

Note. Proficiency refers to self-reported English proficiency. Number refers to the number of participants within a particular group.

6.2.2. Materials and Procedure

6.2.2.i. Study Overview

The Qualtrics survey collected information from participants about their age, gender, first language (L1), self-rated English language proficiency, and education level (see pages 4-5 in Appendix B). Next, the survey asked participants to complete a functional health literacy assessment which aimed to measure their levels of relevant background knowledge (see pages 6-10 in Appendix B). Last, each participant was required to read four of eight health-related texts, and to judge these texts in terms of their understanding, ease of understanding, and effort required to understand these texts (see pages 10-20 in Appendix B; the judgement scales, alongside a justification for their design, are discussed later).

At the beginning of the study, each person was assigned to one of two study groups (blocks). This was done for practical reasons as each study session had to be completed within 30 minutes. Participants allocated to Block 1 were presented with a set of four health-related texts (see pages 10-15 in Appendix B), whereas those in Block 2 were shown a different set of four health-related texts (see pages 17-20 in Appendix B). Participants were randomly assigned to blocks and the order in which the health-related texts appeared was counterbalanced. Ethical approval for the study was granted by Lancaster University's Research Ethics Committee in June 2016. Data collection took place between July and August 2016.

UK-S-TOFHLA. To measure participants' relevant background (health) knowledge, a component of health literacy (Chin et al., 2011), I used a UK adapted version of the Short Test of Functional Health Literacy in Adults (S-TOFHLA; Baker et al., 1999). The original S-TOFHLA is a cloze item test consisting of 36 prose passage items. Each item has four available answer options and the participants must choose one of these options to fill-in the missing spaces. Critically, in previous research, the test was found to have a reliability of .97

(Cronbach's α) and to be highly correlated (r = .80) with scores on the Rapid Estimate of Adult Literacy in Medicine (REALM; Davis et al., 1993), a well-established measure of health literacy (Baker et al., 1999).

The S-TOFHLA was modified for the British population by von Wagner, Knight, Steptoe, and Wardle (2007). Like the original S-TOFHLA, the UK-S-TOFHLA (von Wagner et al., 2007) consists of a series of cloze texts where certain words are removed from a sentence and participants must replace the missing words with one of four available options. However, the UK-S-TOFHLA has slightly fewer items than the original S-TOFHLA as some items which were specific to the U.S. healthcare system, such as those mentioning medical insurance, have been removed. In total, there were 30 missing spaces that participants had to fill in the UK-S-TOFHLA used in this study (see pages 6-10 in Appendix B).

Health-Related Texts. I chose eight health-related texts from the sample of 86 health-related texts analysed in Study 1 (Chapter 5, section 5.2.1). These health-related texts were selected based on their FRE (Flesch, 1948) and RDL2 (Crossley et al., 2008) scores (Table 6.2). Specifically, texts with either the highest or lowest FRE and RDL2 scores, corresponding to texts that were high or low in readability levels, were chosen. As mentioned in Chapter 5 (section 5.2.1), the chosen texts did not discuss potentially emotive topics, such as cancer treatments, which could influence comprehension judgements.

Table 6.2. Readability scores per text.

	Text (Block)							
Readability measure	1 (1)	2 (1)	3 (2)	4 (2)	5 (1)	6 (1)	7 (2)	8 (2)
FRE	44.56	51.45	45.61	69.53	70.27	75.67	70.23	74.33
RDL2	9.31	9.47	10.89	26.68	20.82	15.32	14.6	18.27

Notes. FRE = Flesch Reading Ease; RDL2 = Coh-Metrix L2 Readability Index. The scores for FRE range from 0 to 100, where the higher the score the higher the readability of the text. RDL2 scores range from 0 to 30, where the higher the score the easier the text.

Judgement Scales. As discussed in Chapter 2 (section 2.2), comprehension judgements may be based on ease of text processing (e.g., Crossley et al., 2017; Rawson & Dunlosky, 2002).

Thus, three judgements scales were devised to test metacomprehension: perceived effort exerted to understand each health-related text scale, ranging from 1 (no effort at all) to 9 (a lot of effort); perceived understanding of each health-related text scale, ranging from 1 (not well understood at all) to 9 (extremely well understood); and perceived ease of understanding scale ranging from 1 (impossible to understand) to 9 (extremely easy to understand). Nine-point judgement scales were chosen because such scales have been found to outperform alternative scales with fewer categories in terms of criterion validity, internal consistency, test-retest reliability, and discriminating power (Preston & Colman, 2000).

Pilot. I piloted the study with 12 participants. I excluded the UK-S-TOFHLA from this process on the basis that its reliability and validity has been claimed to have been established (von Wagner et al., 2007). During the pilot, I found that participants were confused by reverse coding of the judgement scales, thus I decided not to vary the scales between texts.

6.2.2.ii. Variable Selection

The study had a repeated measures design, with each participant nested within a block of four health-related texts, rating four health-related texts on three judgement scales. In addition, all participants were asked to complete the UK-S-TOFHLA (von Wagner et al., 2007) and answer background questions. In the analyses, the dependent or outcome variables were: (1.) perceived effort exerted to understand each health-related text; and (2.) perceived understanding of each health-related text. I excluded the perceived easiness of understanding judgement scale from the primary analysis because it was highly correlated (r = .87) with perceived understanding. In addition, some participants verbally reported in the pilot study that the two scales were indistinguishable, suggesting that understanding judgements were likely to be based on the perceived easiness of understanding or vice versa.

The primary analysis consisted of two separate sets of models, one set for each outcome variable. In both sets of models, the predictors were based on the literature reviewed at the beginning of this chapter (section 6.1). Specifically, the FRE (Flesch, 1948), and the RDL2 (Crossley et al., 2008), were used to assess the readability levels of health-related texts. These two readability formulae were chosen as the former mapped onto some recommendations, such as word length, for the production of health information (e.g., NHS England, 2018a; Plain English Campaign, 2018), whereas the latter was theoretically promising due to its consideration of text coherence (e.g., Crossley et al., 2017; Kintsch, 1998; Kintsch & Rawson, 2007; Rawson & Dunlosky, 2002).

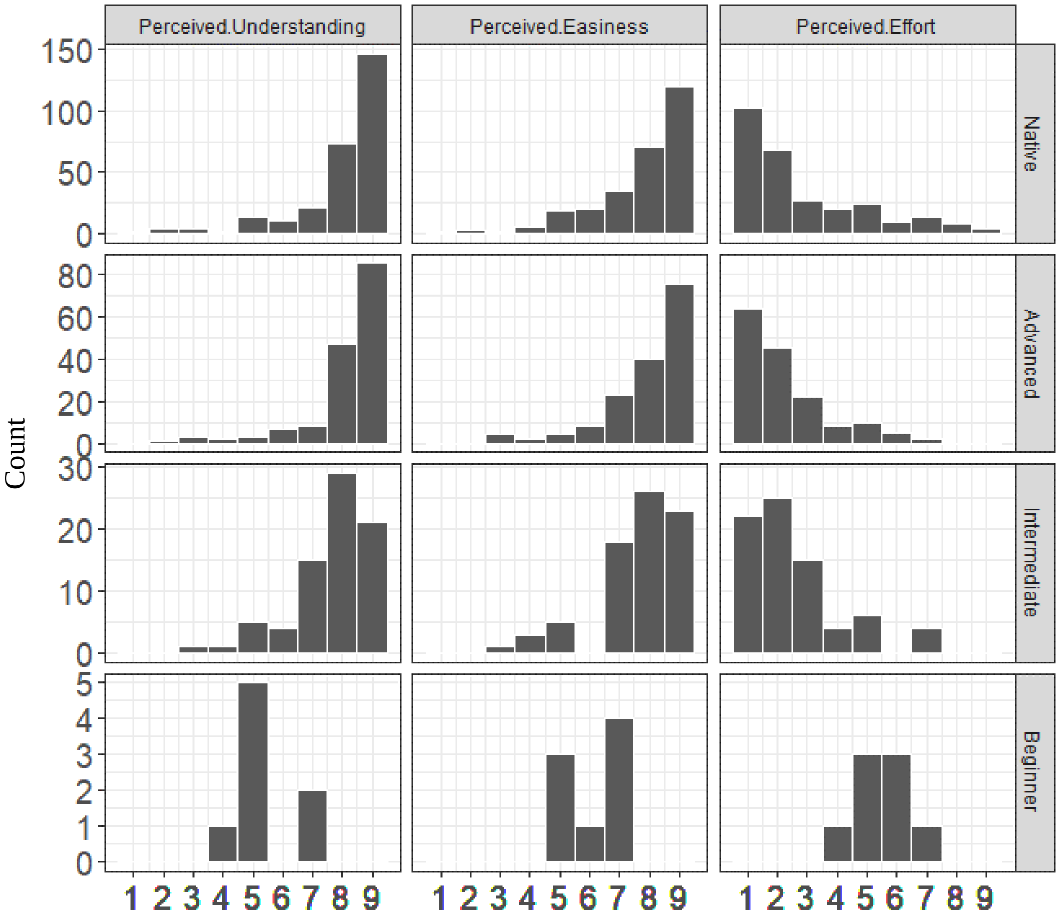
In terms of measuring the effects of variation in individual differences, age, self-reported English language proficiency, health literacy, and education level were chosen as they were considered plausible predictors of metacomprehension (e.g., Dunlosky et al., 2006; Griffin et al., 2009; Hong-Nam & Page, 2014; Kobayashi et al., 2015; 2016; Zabrucky et al., 2012). Critically, the effects of variation in readability and individual differences were allowed to interact, as there is evidence to suggest that the effects of variation in readability levels may vary for readers depending on their backgrounds (e.g., Liu et al., 2009; McNamara, 2001; McNamara & Kintsch, 1996; O'Reilly & McNamara, 2007; Ozuru et al., 2009).

Following the recommendations of Gelman (2008), all the predictors were standardised, meaning they were scaled by two standard deviations and centred to have a mean of zero. As mentioned in Chapter 5 (section 5.3.2.i), this allowed simple comparisons among predictors. In the case of the categorical variables, self-reported English language proficiency and education level, prior to standardising, the variable values were first converted to numeric codes. Standardising categorical predictors meant that the models incorporated the assumption that the effects of variation in education and English language proficiency were linear. This assumption of linearity enabled the models to yield estimates of the effects while reducing the risk that the models would fail to converge. The risk of nonconvergence was associated with the uneven sampling of participants from different strata of education or language proficiency.

Standardising categorical variables transforms the resulting coefficients of these variables so that a unit change in the transformed predictor is comparable to a change in the category of the untransformed categorical predictor (Gelman, 2008). Considering the distribution of judgements across self-reported proficiency levels (see Figure 6.1), native English and advanced English language speakers were assumed to have the highest level of English language proficiency, and were therefore treated as one group, followed by intermediate and beginner English language readers, respectively. The levels of the other categorical predictor, participants' education level, ranged from secondary school to further education to higher education. Consequently, both categorical predictors had three levels and their coefficients can be interpreted in a similar way to their untransformed categorical coefficients. This is because, for example, a change from intermediate to advanced level proficiency, corresponds approximately to a change in two standard deviations from the mean.

Overall, I fitted two sets of Bayesian ordinal mixed-effects models (these are discussed in the results section). One set had the perceived understanding as the outcome variable, whereas the other had the perceived effort as the outcome measure. The predictors were the same in both sets of models and were kept constant during sensitivity analyses (section 6.3.2.iv). I discuss my analyses next.

Figure 6.1. Distributions of ratings per judgement scale and self-reported English language proficiency.



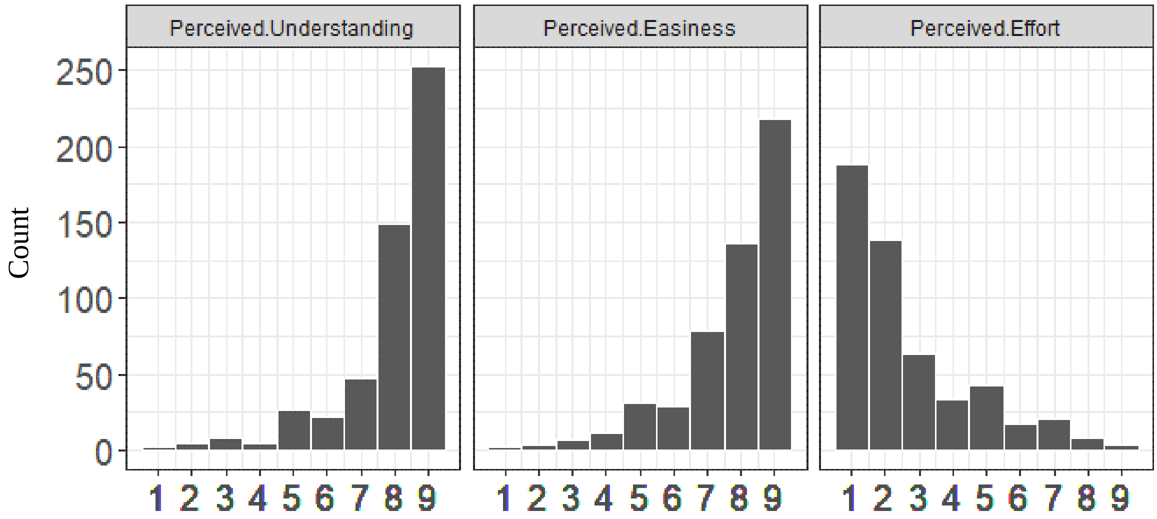
6.3. Results

In this section, first, I discuss the distribution of scores. This is because the distributions play the critical role of determining the model classes used in the primary analyses (section 6.3.2). In addition, visualising the distribution of ratings allows to examine the levels of self-rated understanding of health-related texts among the participants of this study (RQ6.1). Second, I briefly describe the correlations between the different judgement scales because it is theoretically interesting whether perceived effort and comprehension judgement scales correlate (e.g., Rawson & Dunlosky, 2002). Third, I interpret the estimates calculated by the Bayesian models to answer RQ6.2 and RQ6.3. Last, I discuss the Bayesian models building process and the sensitivity analyses.

6.3.1. Descriptive Statistics

In the following description, I discuss the distribution of observed judgements in terms of medians, as the mean is much more sensitive to skew than the median (e.g., Maronna, Martin, & Yohai, 2006). The distributions of judgement ratings were skewed towards ceiling for perceived understanding ($Mdn_{\text{perceived understanding}}$ = 8) and perceived easiness of understanding ($Mdn_{\text{perceived ease}}$ = 8). In contrast, the distribution of judgement ratings for perceived effort was skewed towards floor ($Mdn_{\text{perceived effort}}$ = 2) (Figure 6.2). Ratings on these judgement scales ranged from 1 to 9, where higher ratings corresponded either to better perceived understanding and ease of understanding, or to greater perceived effort. Overall, most participants perceived their understanding of health-related texts to be, on average, relatively high. Concomitantly, many participants judged their effort required to understand these texts as, on average, relatively low.

Figure 6.2. Distributions of ratings per judgement scale.



There was some variation in judgements between different texts, but this variation was on average relatively small (Table 6.3). Text 6 was perceived as easiest to understand and requiring the least effort, whereas text 3 was perceived as least easy to understand and requiring the most effort. Overall, the judgements suggested that, on average, participants thought their understanding of all texts was relatively high. Critically, the distributions of ratings were similar across understanding, effort, and easiness scales between participants from different educational backgrounds (Figure 6.3). However, there was slightly more variation in judgements on the three scales for the less educated individuals, compared to university educated participants. Specifically, a higher proportion of the less educated individuals perceived some texts to be more difficult to understand, and requiring more effort to understand, compared to university educated participants. This suggests that less educated individuals may be more likely to struggle to understand health-related texts than higher educated individuals. In addition, there were some differences in perceived understanding between participants of varying English language proficiency levels, indicating that beginner proficiency level ESL speakers may struggle understanding health-related texts more than intermediate and advanced ESL speakers (refer to Figure 6.1 in section 6.2.2.ii).

Table 6.3. Mean judgements per text.

	Understanding		Effort		Easiness	
Text	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	7.76	1.62	3.01	1.89	7.49	1.60
2	8.03	1.41	2.52	1.87	7.93	1.41
3	7.71	1.62	3.45	2.25	7.39	1.54
4	8.06	1.45	2.55	2.14	7.90	1.49
5	8.09	1.43	2.37	1.71	7.94	1.50
6	8.22	1.30	1.99	1.24	8.03	1.37
7	7.74	1.68	2.82	1.88	7.45	1.70
8	8.03	1.57	2.29	1.78	7.94	1.63

Note. For Understanding and Easiness, judgements range from 1 (not understood/difficult to understand) to 9 (extremely well understood/easy to understand); for Effort, the judgements range from 1 (no effort at all to understand) to 9 (a lot of effort).

Regarding the functional health literacy scores, the UK-S-TOFHLA (von Wagner et al., 2007) scores ranged from 10 to 30, but most of the participants scored at or near ceiling (Figure 6.4; $Mdn_{\text{UK-S-TOFHLA}} = 29$). One explanation for the high UK-S-TOFHLA scores may be the relatively high proportion of university-educated participants (see Table 6.1 in section

6.2.1). An alternative explanation may be that the UK-S-TOFHLA may have relatively low discriminant validity, meaning that it does not discriminate between participants of different health literacy levels that well.

Figure 6.3. Distributions of ratings per judgement scale and self-reported education level.

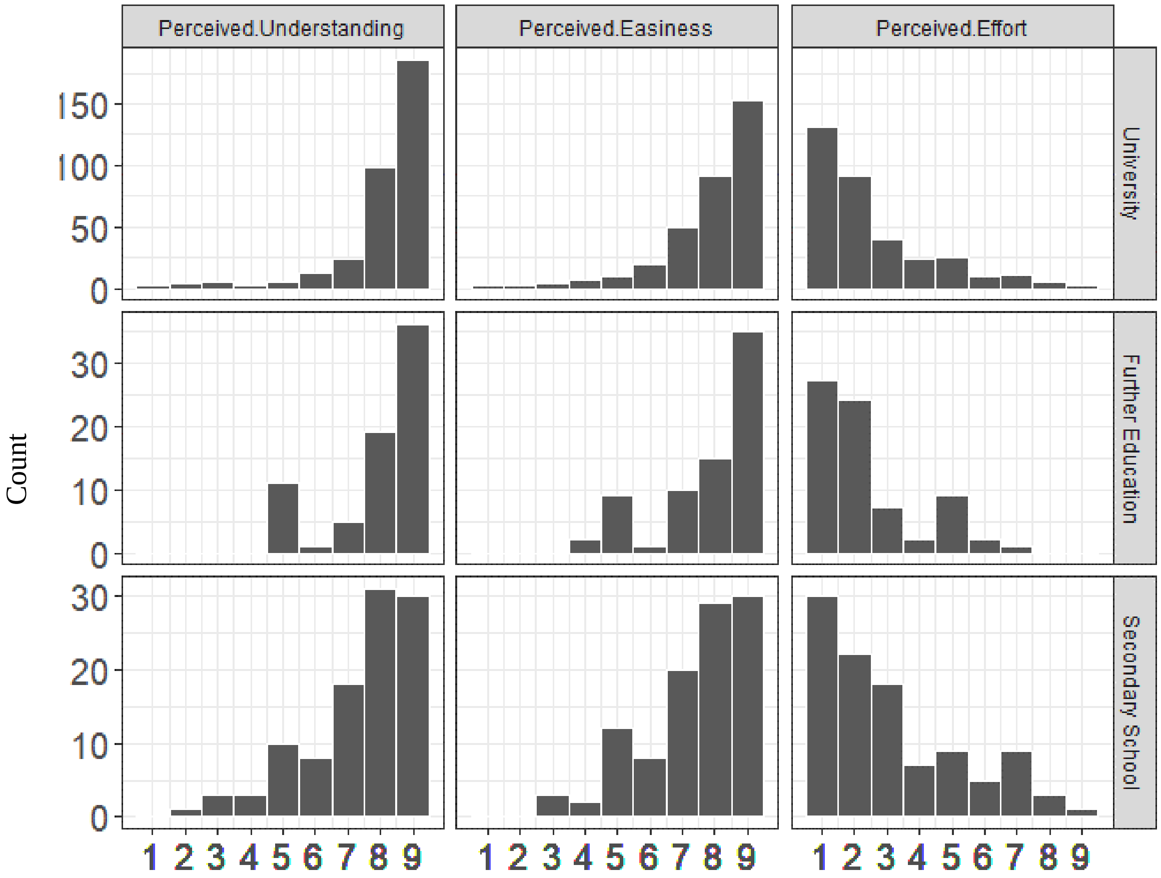
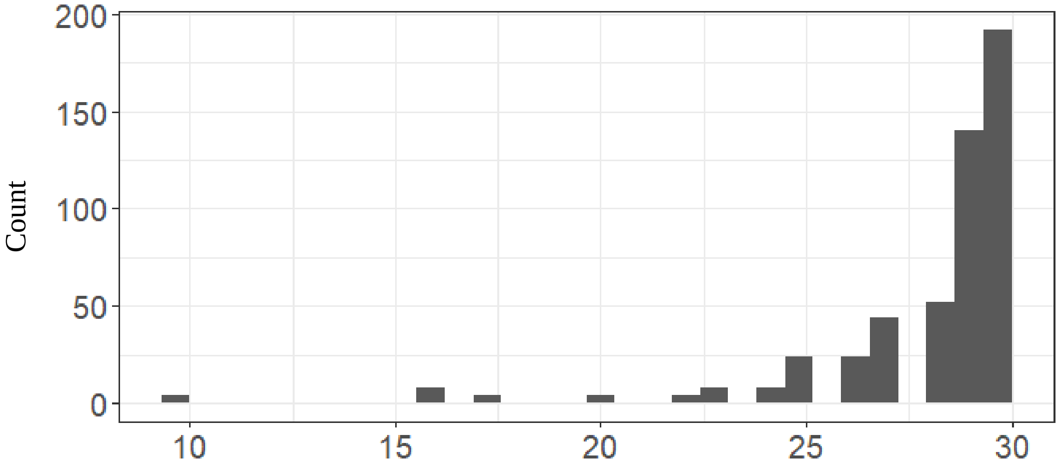


Figure 6.4. Distribution of health literacy (UK-S-TOFHLA) scores.



6.3.1.i. Probability Distribution of Outcome Variables

As mentioned in Chapter 5 (section 5.3.1.i), it is important to discuss the distribution of the outcome variables, as the function and the shape of the distribution determines the model class that should be used to fit the data. In this study, the categories of both outcome measures, perceived understanding and perceived effort, have a natural ordering of the levels, from 1 to 9. However, the distances between the different points on the scales, in terms of the perceptions of the participants, cannot be easily determined. This is because it is not clear how far apart category levels such as “extremely well understood” to “fairly well understood” are (implicitly) judged by participants to be.

In addition, judgement scales tend to be characterized by ceiling or floor effects, whereby participants tend to choose responses at, or close to, one of the two limits of the scale (Agresti, 2010). Thus, as demonstrated in Figure 6.2 (section 6.3.1), the responses are typically skewed. Nonetheless, it is relatively common to treat judgement scales ratings as normally distributed interval data, but this is problematic as doing so inflates Type I error rates (Bürkner & Vuorre, 2018). This is because the potential skew is ignored in the analyses, and it is assumed that the data contain more information than they do. Consequently, for both outcome variables, I chose to conduct analyses assuming that the ratings produced by participants corresponded to an ordinal probability distribution rather than to a normal distribution. This approach can be understood as a generalization of the linear model (akin to the use of binomial logistic regression to analyze accuracy) tailored to be appropriate to the analysis of ordinal data.

Critically, using ordinal models to model ordinal data enables more accurate estimation of the effects than any model which assumes metric or categorical responses (Bürkner & Vuorre, 2018). However, there are several distinct ordinal model classes to choose from (for an overview see Bürkner & Vuorre, 2018). One of these classes, the cumulative model, assumes that the observed ordinal outcome variable represents the categorization of a latent continuous variable. It models this categorization by assuming that there are several thresholds at which the outcome variable is partitioned. This categorization is commonly used to model Likert-scale data, when ordered labels are used to collect judgements about a potentially continuous latent variable (Bürkner & Vuorre, 2018). In this study, it is reasonable to assume that the recorded effort and understanding judgement scores result from the (internal, implicit) categorization by participants of a latent continuous variable corresponding to their opinions about the effort required to understand, or their understanding of, the texts they read. Consequently, I chose to model the data using ordinal cumulative models. However, it is important to mention that for some secondary analyses, specifically examining of correlations between the predictor and outcome variables, I treated the ordinal judgements as continuous. I briefly discuss these correlations next.

6.3.1.ii. Correlations

Table 6.4 shows the correlations between the different predictors and the three judgement scales. First, the correlation between perceived effort and perceived understanding is only moderately high. This indicates that the responses to these two judgement scales may be partially overlapping in variance, but that perceived effort is likely to measure something distinguishable from perceived understanding. Second, the self-reported English language proficiency levels are moderately correlated to health literacy, suggesting that those with high levels of English language proficiency, and native English readers, may have higher health literacy levels than their lower proficiency counterparts.

Critically, the correlation between the FRE and RDL2 has increased dramatically from r = .18 in Study 1 to r = .72 in Study 2. Thereby, the correlation between the FRE and RDL2 exceeded the frequently accepted threshold for collinearity of .7 (Dormann et al., 2013) (see also Chapter 5, section 5.3.1.ii). However, the standard errors of FRE and RDL2 estimates were not meaningfully distorted by this correlation, as judged using the variance inflation factor (Kabacoff, 2011) (see also Chapter 5, section 5.3.1.ii). Thus, readability scores of both readability formulae were retained in all models. Nonetheless, the change in correlation between FRE and RDL2 from Study 1 to Study 2 suggests that the lower number of health-related texts does not represent the population of health-related texts from which these texts were sampled very well. Consequently, the correlation coefficient between readability estimates that were derived from different readability formulae may be spuriously high compared to the true correlation coefficient of all health-related texts. This is because as the number of observations decreases, the correlation coefficient becomes increasingly unstable and likely to yield inaccurate estimates (Schönbrodt & Perugini, 2013). In other words, the smaller the sample of texts, the more the variability between texts is reduced, and the greater the probability grows of obtaining spuriously large correlation coefficients.

Table 6.4. Correlations between reader characteristics, readability measures, and metacomprehension judgements.

	Perceived understanding	Perceived effort	Perceived ease	Age	English proficiency	Health literacy	Education level	FRE
Perceived understanding								
Perceived effort	-.68***							
Perceived ease	.87***	-.76***						
Age	.24***	-.13**	.21***					
English proficiency	.20***	-.09*	.10*	.19***				
Health literacy	.39***	-.25***	.25***	.14**	.49***			
Education level	.20***	-.11*	.12**	-.23***	.15***	.28***		
FRE	.08	-.18***	.10*	.01	.01	.01	-.01	
RDL2	.06	-.10*	.08	.01	.01	.03	-.02	.72***

Note. Significance values are based on Pearson's *r* . * = *p* < .05; ** = *p* < .01; *** = *p* < .001.

Correlations are indicative of potential trends, but to further understand the plausible relations between the predictors and outcome variables shown in Table 6.4, it is important to construct models that can make predictions and treat the judgement scales as ordinal data. I discuss these models and interpret the estimates of these models next.

6.3.2. Bayesian Models

Using Bayesian cumulative mixed-effects models, I examined the factors that influenced variation in the ratings of the two judgement scales: perceived understanding of health-related texts, and perceived effort exerted to understand health-related texts. The predictors included person-level variables, such as age, English language proficiency, health literacy, and education level, and estimates of text readability, specifically the FRE (Flesch, 1948) and the RDL2 (Crossley et al., 2008) scores. I analysed 516 observations — four sets of text judgements per person — using the brm function of the brms package (Bürkner, 2017; 2018) in R (R Core Team, 2019). As I hypothesized that the effects of participant attributes could modulate the impact of the effects of text readability, I included interaction terms corresponding to the interactions between the effects of readability formulae and the effects of individual differences.

6.3.2.i. Prior Distributions

As mentioned in Chapter 5 (section 5.3.2.i), throughout the studies in this thesis I decided to use weakly-informative regularising priors. Including such priors in models improved computational stability by giving the models enough information to avoid inappropriate inferences, while still allowing reasonable variation in the potential estimated effects of predictor variables (Depaoli & van de Shoot, 2017; Gelman & Henning, 2017). However, what priors would be considered as weakly-informative in ordinal regression models depends on the scales used in these models (Bürkner & Vuorre, 2018). Critically, there are currently no specific guidelines for assigning prior distributions for perceived understanding and effort nine-point judgement scales for Bayesian cumulative models. Thus, I opted for what I reasoned would be weakly-informative priors for coefficients of predictor variables. Specifically, I chose priors with normal distributions as the normal distribution is relatively generalisable and frequently used as a prior distribution for the effects of predictors in cumulative models (Bürkner & Vuorre, 2018).

Importantly, unlike Study 1 (Chapter 5) the data collected for Study 2 had a multilevel structure. Specifically, every participant provided four observations, as each participant had to judge their effort exerted on, and understanding of, four health-related texts. Critically, the participants were nested within texts as they were allocated to one of two blocks of four texts at the beginning of the study. Consequently, it was assumed that participants could vary at random in their judgements on the judgement scales, and that participants' judgements of the two sets of texts could also vary at random. To account for this random variation in rating judgements, all the models were fitted with maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013). This means, that the by-individual and by-individual-nested-within-texts random intercepts were fitted with terms corresponding to random variation in the slopes of all individual differences' effects like language proficiency, and all textual measures effects like readability.

Maximal random effects structure allowed models to accurately estimate the effects of predictors while accounting for random variation in the rating judgements associated with unexplained differences between sampled participants and sets of texts. However, in addition to assigning prior distributions to predictors of the model, it was also necessary to assign prior distributions corresponding to reasonable expectations about the potential variances encompassing the random effects. The first prior specification step involved setting a Lewandowski-Kurowicka-Joe (LKJ; Lewandowski, Kurowicka, & Joe, 2009) prior for the plausible correlations between the random effects. The shape parameter of the LKJ prior was set to 2 to prevent the models from allowing extreme correlations, such as ±1, while still permitting relatively high correlations, if warranted by the data, such as the observed pair-wise between RDL2 and the FRE (*r* = .72).

In setting the prior for the variances of random effects, I started with a model with a weakly-informative normal prior with a mean of zero and standard deviation of 10. Next, following the guidelines of Chung, Rabe-Hesketh, Dorie, Gelman, and Liu (2013) for random effects priors, I built a set of candidate models. Chung et al. favour using weakly-informative gamma priors on random effect parameters because, in using the gamma priors, it is possible to assume that random effects with zero variance are impossible. The Gamma distribution has a shape parameter α and a rate parameter, the reciprocal of the scale, λ. Chung et al. recommend gamma priors with $\alpha = 2$ and $\lambda \rightarrow 0$ (where → corresponds to values of λ approaching zero), or alternatively $\alpha = 3$ and $\lambda \rightarrow 0$, where $\lambda \rightarrow 0$ could be .1, .01, .001, 10^{-4}, 10^{-5} and so on. Thus, I fitted the perceived understanding and effort models with different sets of random effects priors. I discuss the optimal models that I arrived at using this approach next.

6.3.2.ii. Perceived Understanding Models

To answer RQ6.2 and RQ6.3, I fitted a series of Bayesian cumulative models. The model building process and sensitivity analyses are briefly described in section 6.3.2.iv, following. In this section, I present a summary of the final model, showing the plausible effects of the predictors of judgements of perceived understanding of health-related texts (Table 6.5). An effect can be said to be plausible or credible because the model, given the data, and the assumptions outlined, yields a posterior distribution that assigns the maximum probability to the coefficient for the effect (the reported Estimate in the table), allocating the bulk of the probability mass to alternate candidate effect coefficient estimates that are either negative only or positive only but not both. In other words, effects can be said to be plausible because, if the 95% lower and upper credible interval bounds include 0 then, essentially, the model cannot tell us if the effect of a variable increases or decreases outcome values (a state equivalent to not knowing if there is an effect at all).

One way of thinking about the estimated effects, reported in Table 6.5, is in terms of estimated changes in judgements. Consequently, to aid the interpretation of the summary table, an additional column, the Expected Judgement Change (EJC), was added to demonstrate the likely impact of the reported effects on self-rated understanding. The EJC scores were calculated by taking into account both the effects' estimates and the eight intercept parameters corresponding to the log-odds thresholds representing different judgement scores.

Table 6.5 shows three plausible influences on the perceived understanding of health-related texts. (Table 6.6 in Appendix C shows the random effects structure of the final model). With each unit increase in age, the log-odds of participants reporting a higher category of perceived understanding (for a text) increased by an average of 9.35 (95% CIs [2.65, 16.31]). In real terms, as shown in the EJC column, this means that older individuals were predicted to rate their understanding of health-related texts higher (rated understanding = 9) than younger individuals (understanding = 8). Similarly, with each unit increase in health literacy, the perceived understanding log-odds were predicted to increase by an average of 11.50 (95% CIs [1.26, 22.27]). Therefore, those with higher health literacy were predicted to rate their understanding of health-related texts as higher (9) than those with lower health literacy (8). In addition, the more educated participants were also predicted to rate their perceived understanding in a higher category (9), on average, than the less educated participants (8) (95% log-odds CIs [2.49, 16.07]). Critically, none of the interactions appeared to indicate credible modulation of text effects by individual differences effects, given the sample data.

Table 6.5. Summary of the final model (Perceived Understanding 3.1).

Coefficients	Estimate	EJC	Est. Error	L-95%	U-95%	Probable (sign)
Intercept [1]	-44.16	-	4.49	-53.43	-35.89	
Intercept [2]	-36.48	-	3.49	-43.62	-29.90	
Intercept [3]	-30.66	-	2.96	-36.59	-25.14	
Intercept [4]	-28.19	-	2.78	-33.80	-23.01	
Intercept [5]	-20.99	-	2.28	-25.63	-16.70	
Intercept [6]	-16.20	-	1.98	-20.20	-12.49	
Intercept [7]	-9.36	-	1.67	-12.66	-6.09	
Intercept [8]	4.37	-	1.94	1.00	8.56	
Age	9.35	8 → 9	3.50	2.65	16.31	(+)
English proficiency	2.49	8 → 8	3.77	-4.45	10.43	
UK-S-TOFHLA	11.50	8 → 9	5.30	1.26	22.27	(+)
Education level	9.01	8 → 9	3.44	2.49	16.07	(+)
FRE	2.10	8 → 8	1.49	-0.74	5.16	
RDL2	2.76	8 → 8	1.63	-0.33	6.09	
Age:FRE	2.17	8 → 8	2.94	-3.54	7.97	
Age:RDL2	0.94	8 → 8	3.18	-5.18	7.30	
English proficiency:FRE	-2.30	8 → 8	2.82	-7.84	3.19	
English proficiency:RDL2	1.96	8 → 8	2.97	-3.83	7.79	
UK-S-TOFHLA:FRE	0.11	8 → 8	3.47	-6.81	6.86	
UK-S-TOFHLA:RDL2	-1.53	8 → 8	3.70	-8.78	5.82	
Education level:FRE	-2.80	8 → 8	2.71	-8.25	2.39	
Education level:RDL2	3.89	8 → 8	2.84	-1.61	9.61	

Note. EJC refers to Expected Judgement Change.

6.3.2.iii. Perceived Effort Models

The analysis of perceived effort followed the same approach as the analysis of perceived understanding. Table 6.7 shows estimates of the predictors that influenced the perceived effort required to understand health-related texts (Table 6.8 in Appendix C shows the random effects parameters of the final model). Critically, the FRE was the sole credible predictor of perceived effort. With each unit increase in the FRE, the perceived effort (required to understand a text) log-odds were predicted to change by -8.56 (95% CIs [-12.61, -4.94]). However, the effects of FRE on perceived effort were relatively weak, as in practice, an increase in FRE score was not predicted to lower the expected judgement. Similar to the perceived understanding model, none of the interaction effects were plausible as judged using 95% credibility intervals.

Table 6.7. Summary of the final model (Perceived Effort 3.1).

Coefficients	Estimate	EJC	Est. Error	L-95%	U-95%	Probable (sign)
Intercept [1]	-13.91	-	3.39	-21.29	-8.05	
Intercept [2]	2.60	-	2.20	-2.13	6.62	
Intercept [3]	10.90	-	2.08	6.67	14.95	
Intercept [4]	15.98	-	2.20	11.68	20.35	
Intercept [5]	24.00	-	2.61	19.00	29.30	
Intercept [6]	28.53	-	2.93	22.93	34.48	
Intercept [7]	37.77	-	3.65	30.78	45.09	
Intercept [8]	47.82	-	4.80	38.89	57.74	
Age	-4.41	2 → 2	4.19	-12.56	3.93	
English proficiency	-1.13	2 → 2	5.18	-12.47	8.15	
UK-S-TOFHLA	-11.58	2 → 2	6.27	-24.09	.61	
Education level	-5.37	2 → 2	4.40	-14.29	2.99	
FRE	-8.56	2 → 2	1.96	-12.61	-4.94	**(-)**
RDL2	-2.26	2 → 2	1.96	-6.20	1.51	
Age:FRE	-4.51	2 → 2	3.40	-11.26	2.08	
Age:RDL2	-1.13	2 → 2	3.76	-8.48	6.25	
English proficiency:FRE	-4.26	2 → 2	3.48	-11.24	2.41	
English proficiency:RDL2	0.18	2 → 2	3.89	-7.49	7.87	
UK-S-TOFHLA:FRE	0.11	2 → 2	4.43	-8.65	8.74	
UK-S-TOFHLA:RDL2	2.95	2 → 2	4.85	-6.52	12.63	
Education level:FRE	0.85	2 → 2	3.21	-5.44	7.18	
Education level:RDL2	-6.42	2 → 2	3.55	-13.43	.47	

Note. EJC refers to Expected Judgement Change.

Overall, I found evidence to suggest that age, health literacy, and education may predict perceived comprehension judgements of health-related texts. In addition, the effect of FRE was likely to predict the effort participants felt they exerted to understand health-related texts. Critically, I did not find substantial evidence for interactions between the effects of individual differences and readability levels on these measures of metacomprehension. I discuss these findings later, but first I describe the model selection process and sensitivity checks of the analyses.

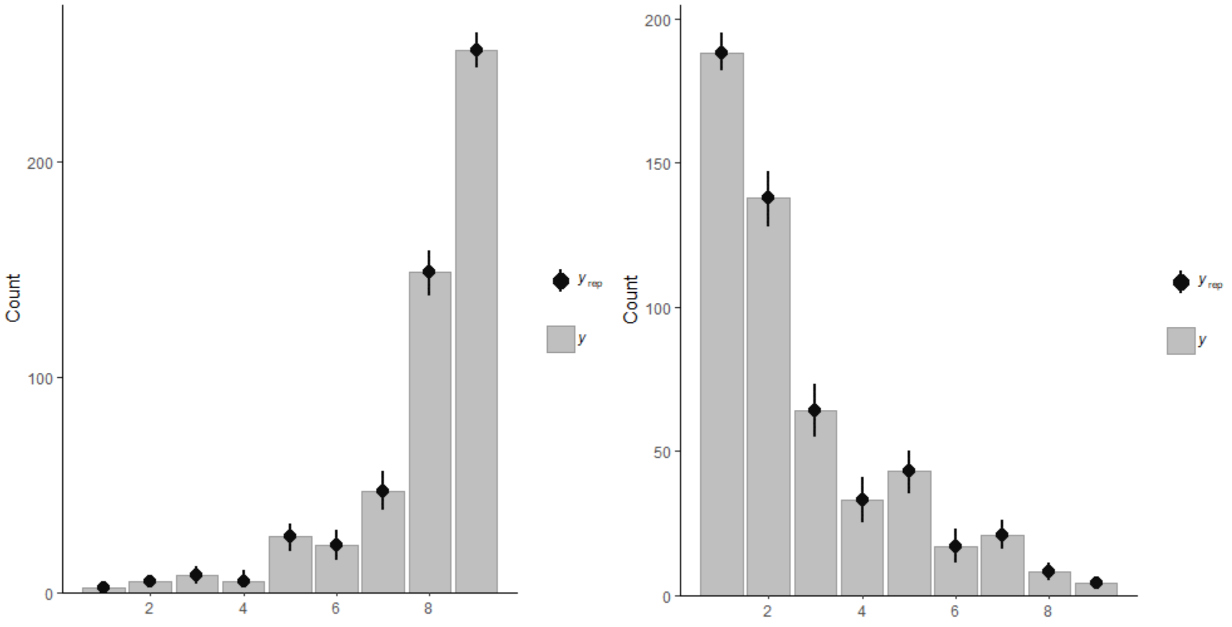
6.3.2.iv. Sensitivity Analyses

In this section, I briefly describe the model selection process. First, I fit a set of models differing in their prior distributions of random effects parameters to check for the sensitivity of the estimates to the choice of the prior (see Table 6.9 and Table 6.10 in

Appendix C). These checks demonstrated that the models' estimates were relatively robust, as in 21 out of 22 of the considered models, the credible estimates were not sensitive to the choice of the prior distribution. Second, I checked $\hat{R}$ chain convergence criterion associated with the fitted models (Gelman et al., 2013; for more details see Chapter 5, section 5.3.2.iv). Only the models reported here satisfied the convergence criteria of $\hat{R} = 1.00$ (Gelman et al., 2013). Thus, in both sets of candidate models, the reported model was determined to be the optimal model, most likely to provide reliable coefficient estimates.

Next, I checked for the presence of local convergence in the converged models (for more details on MCMC and local convergence refer to Chapter 5, section 5.3.2.iv). After doubling the number of iterations used to identify the posterior distribution for the reported models, I found that they still converged and that the effect estimates did not change. This suggests that the estimates of the reported models were relatively insensitive to changes in the number of iterations. Last, I checked the reported models' predictive performance. The posterior predictive check (PPC; Figure 6.5) plots show that the reported models had excellent predictive performance, as the model-implied replicate datasets closely resembled the observed data (Martin & Williams, 2017).

Figure 6.5. PPCs of the perceived understanding and effort models respectively.



Understanding

Effort

Note. Replicate model-implied datasets are plotted as black CIs labelled y_{rep}, the observed data is plotted as grey bars labelled y.

6.4. Discussion

In this study, I aimed to examine whether health-related texts produced by the NHS are perceived to be understandable by individuals from different backgrounds. I also investigated the effects of variation in reader characteristics and text readability levels on self-rated judgements of perceived comprehension and effort. My analyses showed that perceived understanding and perceived effort judgements were influenced by different sets of plausible predictors. In addition, there was no evidence to support the hypothesis that the effects of participant attributes could modulate the impact of the effects of variation in text readability levels. I discuss my findings in the following.

I asked, “What is the self-rated understanding of health-related texts used in this study?” (RQ6.1). On average, the subjective judgements of perceived comprehension revealed that the sample of individuals participating in this study thought that their understanding of health-related texts was relatively high, across the texts they read (Figure 6.2 and Table 6.3). These descriptive findings support $H_{6.1}$ because they suggest that health-related texts, regardless of readability levels, may be perceived to be relatively easy to understand. However, this is problematic as both readability formulae indicate that the sample of texts used in this study should vary in comprehensibility levels (Table 6.2), but the perceived understanding judgements were not associated with differences in the estimated readability of the texts (section 6.3.2.ii).

One potential explanation for the lack of evidence for the effects of variation in readability on self-rated judgements of understanding may be that readers make their metacomprehension judgements based on the apparent ease of text processing rather than on perceived understanding (Begg et al., 1989; Dunlosky et al., 2006). Indeed, variation in FRE (Flesch, 1948) was found to predict the self-reported judgements of perceived effort required to understand health-related texts (section 6.3.2.iii). Overall, texts with high FRE scores were judged as requiring marginally less effort to understand than those with lower FRE scores (Tables 6.2 and 6.3). This suggests that a high proportion of short words and sentences in a text may signal to readers that the text they are reading should be easy to understand, and that they do not have to engage in active processing to understand that text (Crossley et al., 2017; O’Reilly & McNamara, 2007; van den Broek & Helder, 2017). In turn, if perceived effort predicts the actual comprehension of health-related texts, the guidelines recommending the use of short sentences and words (e.g., Plain English Campaign, 2018) may be contributing to differences in tested comprehension (investigated in Chapter 7).

In my study, as well as examining the effects of textual readability measures, I also investigated whether some reader characteristics predict self-rated comprehension of health-related texts (RQ6.2). I found that high relevant background knowledge, as measured using the UK-S-TOFHLA (von Wagner et al., 2007), was predictive of high perceived comprehension judgements. There is a potential two-fold explanation for this finding. First, high-background-knowledge readers may be more likely to be skilled readers who in turn are more likely to engage in active processing (e.g., O'Reilly & McNamara, 2007) to self-regulate their comprehension (Thiede & Anderson, 2003; Thiede et al., 2010). Consequently, metacomprehension judgements of individuals with higher levels of relevant background knowledge may be more accurate at predicting their comprehension compared to judgements of those with lower levels of relevant background knowledge (Griffin et al., 2009).

Second, research evidence indicates that those with high relevant-background knowledge and reading skill tend to understand more of the text read than those with lower relevant-background knowledge and reading skill (e.g., Ozuru et al., 2009). Therefore, the theorised increase in metacomprehension accuracy of high-background-knowledge readers (e.g., Griffin et al., 2009; Ozuru et al., 2009) may be reflected in the higher perceived comprehension judgements of those with higher health literacy levels. This is because adults with higher health literacy levels tend on average to outperform adults with lower health literacy levels on comprehension measures (e.g., Chin et al., 2018). Thus, those with high health literacy levels may be more accurate at judging their comprehension, and their comprehension of health-related texts is likely to be higher, compared to those with lower health literacy levels (investigated in Chapter 7).

In addition to health literacy, high perceived comprehension was also found to be predicted by high education level. One explanation for this could be that, as mentioned in Chapter 2 (section 2.2), the more educated adults may be more likely to regularly evaluate their understanding and self-regulate comprehension breaks by engaging in active processes, such as rereading, than the less educated adults (Thiede et al., 2010; Zabrucky et al., 2012). In turn, the engagement in active processes when reading is likely to be beneficial, as it is thought to improve the accuracy of metacomprehension judgements as well as actual comprehension (e.g., O'Reilly & McNamara, 2007; Zabrucky et al., 2012). Consequently, adults with higher levels of education may judge their perceived comprehension higher, and they may understand more than their less educated counterparts (the effects of educational attainment on tested comprehension are examined in Chapter 7).

Importantly, the perceived comprehension of health-related information was also found to be predicted by participants' age. Specifically, older individuals were more likely to judge their comprehension to be higher compared to younger participants. One explanation for this could be that older adults may be better comprehenders than younger adults, as they may be more efficient at self-regulating their understanding, potentially due to more reading experience (Miles & Stine-Morrow, 2004) or more health knowledge (Griffin et al., 2009; O'Reilly & McNamara, 2007). An alternative explanation could be that older adults may provide more inaccurate metacomprehension judgements than younger adults. Specifically, older adults may be more likely to over-estimate their actual comprehension, due to potential age-related differences in comprehension monitoring (e.g., Dunlosky et al., 2006; Miles & Stine-Morrow, 2004). To examine which one of these explanations is more plausible, the next study considers the effects of age on tested comprehension (Chapter 7).

Overall, in answering RQ6.2, $H_{6.2}$ is only partially supported. This is because the reported evidence suggests that the RDL2 was not associated with perceived effort and understanding of health-related texts, whereas the FRE was a plausible predictor of perceived effort only (plausible, or credible, in terms of the data and the analysis assumptions). Individual differences, except for English language proficiency, were found to be plausible predictors of perceived understanding of health-related texts, but they were not plausible predictors of perceived effort. The potential reason for the lack of evidence for the effect of English language proficiency on perceived understanding judgements is explored in the limitations section later. Critically, based on these findings, it is difficult to determine whether metacomprehension judgements are related to the ease of text processing as this study did not investigate the effects of perceived effort and perceived comprehension on actual comprehension. However, the correlation between the perceived effort and understanding judgements indicates that there is some shared variance between the two constructs these judgement scales measured. Thus, it may be the case that both judgement scales are partially indicative of metacomprehension and may predict tested comprehension (investigated in Chapter 7).

My research also sought to answer the question whether the effects of variation in readability interact with the effects of individual differences to predict self-rated judgements of perceived comprehension (RQ6.3). Critically, I found lack of evidence for the probable effects of these interactions on ratings of both judgement scales ($H_{6.3}$). Thus, it is difficult to argue that $H_{6.3}$ is even partially supported. Considering the hypothesised interaction effects of textual measures of readability and individual differences (e.g., Francis et al., 2018; Kulesz et al., 2016), one would expect the interaction effects to be more predictive of metacomprehension. However, this study did not measure tested comprehension, instead it focused on a proxy of reading comprehension, namely perceived comprehension or metacomprehension. This may be the reason for the lack of credible evidence for the theorised effects of readability estimates by individual differences interactions, as the studies that found these effects focused on tested reading comprehension rather than metacomprehension (e.g., Liu et al., 2009).

Critically, the perceived comprehension and effort judgements may be predicted by different factors to tested comprehension (Kauchak & Leroy, 2016). Furthermore, even if variation in readability scores of health-related texts predicts perceived effort judgements, it is not known whether this variation predicts tested comprehension of health-related texts

(Leroy & Kauchak, 2014). One reason for this is that readers' metacomprehension judgements of texts may not always reflect their actual comprehension levels of these texts (e.g., Maki, 1998; Dunlosky et al., 2005). Therefore, it may be the case that the probable effects of variation in FRE on perceived effort do not reflect changes in comprehension performance (Kauchak & Leroy, 2016). Conversely, the potential effects of variation in text features and texts readability levels on tested comprehension may not be reflected in readers' metacomprehension judgements (e.g., Maki, 1998). Consequently, further research with a larger number of text features by individual differences interactions is needed to: investigate the effects of these interactions on tested comprehension of health-related information; to determine whether the predictors of perceived comprehension also predict tested comprehension; and to establish whether perceived comprehension predicts tested comprehension. This motivated the next study of this thesis (discussed in Chapter 7).

6.4.1. Limitations

This study had two main limitations that future research can avoid. First, a very low sample (N = 2) of beginner level ESL speakers prevented me from making reliable claims about the effects of English language proficiency on metacomprehension of health-related information. Previous research found evidence for differences in metacomprehension strategies use between ESL readers of health-related texts of varying English proficiency (Hong-Nam & Page, 2014). Thus, it is highly probable that the small sample of beginner ESL readers resulted in an under-estimation of the strength of the effect of English language proficiency on both perceived comprehension and perceived effort. Second, the measure of health literacy was not as sensitive as I envisaged it to be when I was designing this study, specifically the UK-S-TOFHLA (von Wagner et al., 2007) scores were at ceiling. If the measure of health literacy used in this study had been more sensitive, it is likely that it would have more discriminant validity enabling more differentiation between individuals. To overcome these limitations, future research could recruit individuals from more heterogenous English language proficiency backgrounds and employ a more sensitive health literacy measure.

6.4.2. Implications

The findings presented in this study are likely to be useful for writers of health information in terms of understanding the factors that influence evaluations of health-related texts. This is important, as the Information Standard (NHS England, 2018a) requires health-related information writers to ask for feedback from the end users when writing health-related texts and some NHS Trusts comply with these requirements by asking a minimum of two patient reader panel members for feedback when producing health-related texts (e.g., Burrow & Forrest, 2015). However, it is currently not clear whether reader panel members can effectively determine what would make the texts easy to understand for the end users.

The effects of probable predictors of perceived comprehension suggest that, on average, individuals' perceived comprehension is not predicted by texts' readability, as assessed using the FRE (Flesch, 1948) or the RDL2 (Crossley et al., 2008). Thus, it may be the case that the indices underlying these two measures of readability, specifically, average word frequency and length, average sentence length, sentence syntax similarity, and content word overlap, are limited predictors of perceived understanding. An alternative explanation might be that due to the possible ceiling effects, in the ratings of perceived comprehension, the effects of variation in readability, between the texts, could not have been detected. Consequently, it is currently unclear what text features the perceived comprehension judgements are based on.

In contrast to perceived understanding judgements, there is evidence to indicate that the perceived effort judgements, a proxy for ease of text processing, are predicted by variation in the proportion of short words and sentences in health-related texts. Therefore, it may be the case that texts with a high incidence of short words and sentences are judged more favourably by reader panel members than texts with a lower proportion of short words and sentences. Writing texts high in FRE is recommended by some of the guidelines NHS writers are encouraged to follow (e.g., Plain English Campaign, 2018). However, it is not clear whether these guidelines predict tested comprehension, or whether they merely influence the judgements of perceived effort required to understand health-related texts. This needs to be investigated further, as the relative accuracy of effort judgements could be low, and effort judgements may not be predictive of comprehension (e.g., Dunlosky et al., 2006).

In addition to the probable effects of readability estimates, the probable effects of variation in reader characteristics also led to further questions. Critically, the probable predictors of perceived understanding indicate that older individuals, and those with higher levels of health literacy or education, are more likely to rate their understanding of health-related texts as higher than younger individuals, and those with lower levels of health literacy or education. It is therefore credible that the views of text comprehensibility vary depending on individuals' background. The finding that health literacy was found to predict the perceived comprehension judgements is particularly problematic as a natural recommendation based on this would be that reader panel members should not have high health literacy levels if their evaluations of texts guide the production of health-related texts for the general population. However, this is not realistically achievable as reader panel members are likely to develop high health literacy levels over time through repeated exposure to health-related texts.

Due to the reported effects of health literacy on perceived comprehension, it is not clear whether the evaluations of long-term reader panel members help in the process of development of understandable health-related texts for the general population. One way of minimising the effects of relevant background knowledge on evaluations of health-related texts, could be ensuring that each reader panel member is not asked to evaluate more than one health-related text about each specific condition. This may reduce the probability of reader panel members gaining relevant background knowledge specific to health-related texts that they are evaluating. Nonetheless, this is also likely to be relatively ineffective at ensuring that the reader panel members do not develop high levels of health knowledge. This is because the health literacy assessment used in this study asked general health questions, and correctly answering those was sufficient to predict high levels of perceived understanding of relatively unrelated health-related texts.

In addition to health literacy, the probable effects of age and education on perceived comprehension are also problematic. This is because if one assumes that reader panel members consist of relatively homogenous groups of people, in terms of age and education level, it may be the case that reader panel members judge health-related texts to be on average easier to understand compared to the rest of the population. Critically, it is not known whether health-related texts are actually easier to understand for older individuals with higher levels of health literacy and education, and whether such individuals can make recommendations that improve tested understanding of other people. It may be the case that perceived understanding of health-related texts varies between readers of different backgrounds, but that tested understanding does not follow the same pattern (e.g., Dunlosky et al., 2006; Maki, 1998). This is because, amongst other plausible reasons, older individuals could be over-estimating their understanding of health-related texts (e.g., Miles & Stine-Morrow, 2004).

In conclusion, if the NHS Trusts continue using reader panel members' evaluations of health-related texts, it is paramount that they be aware that these panels should comprise of individuals of varying age and educational backgrounds. This is because these reader characteristics are likely to predict the perceived comprehension judgements which may impact evaluations of health-related texts. Importantly, the perceived understanding levels of older and highly educated individuals, and the subsequent evaluations of health-related texts, may not reflect the perceived understanding, and text evaluations, of all potential NHS patients. In addition to diversifying reader panels in terms of age and educational level, it may be beneficial for NHS Trusts to focus on recruiting reader panel members from non-health-related backgrounds. This is because perceived understanding, and potential text evaluations, of individuals with higher levels of health literacy are also likely to be different to perceived understanding, and potential text evaluations, of less health literate patients. However, it is important to mention that these suggestions are based on perceived understanding judgements alone, and not tested comprehension of health-related texts. Critically, the usefulness of perceived understanding judgements in predicting tested comprehension of health-related texts is still to be determined. This is investigated in the next chapter (Chapter 7).

Chapter 7: How the effects of Individual Differences Interact with the effects of Text Features to Predict Comprehension of Health-Related Information

This chapter concentrates on comprehension of health-related texts and describes and discusses the third study included in this thesis. It is important to mention that Study 3 was developed considering the findings discussed in the previous two chapters. First, based on the findings of Study 1 (discussed in Chapter 5), it is of theoretical and practical interest to examine whether the text feature predictors of textual measures of readability also predict variation in tested comprehension, and whether variation in readability predicts tested comprehension. Critically, if readability formulae were found to predict tested comprehension, then readability scores of health-related texts could be used as a relatively good proxy for tested comprehension of health information. Second, based on the findings of Study 2 (discussed in Chapter 6), it is important to investigate whether tested comprehension has the same predictors as judgements of perceived comprehension, and perceived effort required to understand texts, and whether metacomprehension judgements predict tested comprehension. Importantly, if tested comprehension has the same predictors as metacomprehension judgements, and metacomprehension judgements predict comprehension performance, then reader panel members judgements could potentially be used as effective comprehension performance predictors.

Overall, Study 3 aimed to investigate how variation in reader attributes and text features, predicts comprehension of health-related information to answer the overarching research questions of this thesis (Chapter 4, section 4.3; section 7.1.2 in this chapter). To strengthen the case for such an investigation, this chapter starts with a short literature review that builds on the literature reviewed in Chapters 1 to 4. This is followed by a method and results sections, based on which the research questions are answered in the discussion section. The chapter ends with a brief discussion of the implications of the findings.

7.1. Literature Review

There are reasons to believe that health communication could be more effective if it were designed not only for the average patient (e.g., Liu et al., 2009). Specifically, predictions that are relatively accurate for an average person only, may be inaccurate for everyone who does not fall into the average patient category. Nonetheless, the guidelines used by the National Health Service (NHS) (e.g., NHS England, 2018a; Plain English Campaign, 2018), assume that the same writing recommendations improve understanding of written health communication uniformly across different groups of the population. However, theoretical and empirical research findings, discussed next, indicate that health-related texts should be adapted for different groups of users.

Critically, the effects of variation in reader attributes are thought to predict comprehension (e.g., Ozuru et al., 2009), and the effects of reader attributes are theorised, and were found, to interact with the effects of text features to predict comprehension (e.g., Francis et al., 2018; Liu et al., 2009). Consequently, there appears to be scope for improving the guidelines for writing health-related documents or suggesting new recommendations based on empirical research findings. However, there is lack of research findings that could guide this process, as only a small number of relatively small-scale studies has been conducted to investigate how the effects of reader attributes and text characteristics interact with each other to predict reading comprehension of health-related texts (e.g., Liu et al., 2009). This research gap motivated the study discussed in this chapter.

To investigate the effects of credible predictors of reading comprehension of health-related texts, it is necessary to identify the candidate variables from a list of reader attributes and linguistic features discussed in the preceding chapters. Next, I draw on the literature review chapters (Chapters 1 to 3) to briefly discuss the effects of variation in theoretically promising reader characteristics on comprehension of health-related texts. The reading comprehension theories discussed in Chapter 1 (sections 1.2 to 1.4), such as the Lexical Quality Hypothesis (LQH; Perfetti, 2007), suggest that the effects of variation in reader characteristics, such as phonological awareness, working memory (WM), and vocabulary knowledge (Kintsch & Rawson, 2007; Perfetti, 2010; Perfetti & Stafura, 2014), are likely to predict comprehension. There are several reasons as to why these individual differences are theorised to be important to comprehension. Phonological awareness is thought to be pertinent to comprehension as it is theorised to be crucial in the development of reading fluency (Perfetti, 1992; Chapter 1, section 1.4). In turn, the attainment of reading fluency is central to comprehension as it is thought to allow the reader to devote more mental resources to generating the meaning of a text read (Perfetti, 1998), thereby to the construction of a logical situation model (Kintsch & Rawson, 2007; Chapter 2, section 2.1.1).

WM constitutes the mental resources that are theorised to be necessary for processing of the textual information read, therefore the construction of propositions, the textbase, and the situation model (Kintsch & Rawson, 2007; Perfetti & Stafura, 2014; Zwaan, 2016) (Chapter 2, section 2.1.1; Chapter 3, section 3.1.1). Thus, the effects of WM on comprehension must be considered alongside the potential effects of phonological awareness, as phonological awareness and WM are likely to be interrelated. This is because without the attainment of reading fluency, more WM resources are likely to be spent on recognising words instead of higher-level processes such as inference-making that are thought to be necessary for the formation of a logical situation model (Kintsch & Rawson, 2007; Perfetti, 2007) (Chapter 1, sections 1.2 to 1.4). Consequently, an investigation of predictors of reading comprehension would be incomplete without the consideration of the effects of phonological awareness and WM.

However, in addition to WM and phonological awareness, another theoretically important predictor of comprehension is variation in vocabulary knowledge (e.g., Perfetti, 2010; Tunmer & Chapman, 2012). Knowledge of word meanings is crucial to forming propositions (Chapter 1, section 1.2), and propositions are key to understanding text as they are a prerequisite to constructing a textbase and a logical situation model of the text read (Kintsch & Rawson, 2007). Thus, without knowing the meanings of the words read, the reader cannot build a complete understanding of the text read. Importantly, phonology is thought to be related to vocabulary as it is theorised to affect word-meaning connections by, for example, creating associations between unfamiliar words and familiar contexts (e.g., Perfetti, 2010) (Chapter 1, section 1.4). In turn, these new associations may allow readers to acquire partial understanding of unfamiliar words, thereby improve individual's knowledge of these words (Perfetti, 2010). Thus, the effects of vocabulary knowledge on comprehension must be studied in conjunction with the effects of phonological awareness.

WM is also thought to be related to vocabulary knowledge (Perfetti & Stafura, 2014; Yang et al., 2005). Specifically, the meaning-to-text integration processes, processes required for textbase formation such as inference making (Kintsch, 1998), of readers with relatively high vocabulary knowledge are theorised to be less WM resource demanding than of those with lower levels of vocabulary knowledge (Perfetti & Stafura, 2014; Yang et al., 2005) (Chapter 1, section 1.4; Chapter 2, section 2.1.2). This is because vocabulary knowledge is thought to modulate the efficiency of the meaning-to-text integration processes. Efficiency is important as the more efficient the execution of meaning-to-text integration processes the less WM resources are thought to be required for the construction of the textbase (Yang et al., 2005; Yang et al., 2007). Consequently, vocabulary knowledge must be studied in conjunction with WM and phonological awareness as they are hypothesised to be interrelated (e.g., Perfetti, 2010; Perfetti & Stafura, 2014; Tunmer & Chapman, 2012).

Critically, some theoretically warranted individual differences are more likely to be predictors of comprehension than others (Chapter 2, section 2.1). The effects of vocabulary knowledge on comprehension were found to be relatively robust amongst those reading in their first language (L1) (e.g., Freed et al., 2017; Tunmer & Chapman, 2012; Van Dyke et al., 2014) and those reading in their second language (L2) (e.g., Brysbaert et al., 2016) (Chapter 2, section 2.1; Chapter 3, section 3.1.3). However, the utility of phonological awareness and verbal working memory in predicting comprehension of adult readers has been questioned (e.g., Freed et al., 2017; Van Dyke et al., 2014). This is because the effects of verbal WM and phonological awareness were found to dissipate in the presence of other predictors of comprehension, such as vocabulary knowledge (Freed et al., 2017) (Chapter 2, section 2.1.1 and 2.1.2). Thus, it is of theoretical interest to examine whether the effects of variation in vocabulary knowledge, phonological awareness, and verbal WM jointly predict comprehension of health-related texts.

In addition to theoretically justified effects of variation in reader attributes, there are also empirically warranted individual differences that may be vital to comprehension. Important, in the context of the linguistically diverse UK population (Office for National Statistics, 2016), is the study of the variation in English language proficiency on comprehension of health-related texts. As mentioned in Chapter 3 (section 3.1.3), there is evidence to suggest that self-rated English language proficiency is likely to predict comprehension of health-related information written in English (Thomson & Hoffman-Goetz, 2010; Todd & Hoffman-Goetz, 2011). One plausible explanation for this could be that individuals with lower levels of English language proficiency, might have had lower levels of exposure to English than those with higher levels of proficiency (Brysbaert et al., 2016; see also Chapter 2, section 2.1.3.i).

Exposure to English matters as lower level of exposure to English may predict lower levels of English language vocabulary (Brysbaert et al., 2016). In turn, English language vocabulary knowledge is theorised, and was found, to be an influential predictor of reading comprehension of English texts for L1 and L2 English readers (e.g., Brysbaert et al., 2016; Freed et al., 2017; Perfetti, 2007; 2010; Tunmer & Chapman, 2012). Consequently, self-rated English language proficiency is likely to be a powerful predictor of comprehension, as it may be a relatively good proxy measure for language exposure and vocabulary knowledge. Nonetheless, the effects of proficiency on comprehension are under-researched as to date few studies have examined the effects of English language proficiency on comprehension of health-related texts (e.g., Todd & Hoffman-Goetz, 2011). This research gap motivated the inclusion of self-rated English language proficiency as a variable in this study.

Notably, the effects of English language proficiency on comprehension of health-related texts must be studied in conjunction with education, as English as Second Language (ESL) readers are likely to vary in educational background (e.g., Todd & Hoffman-Goetz, 2011). This is important as there is evidence to suggest that higher levels of education, regardless of whether education was completed in English or another language, predict higher understanding of health-related texts (Todd & Hoffman-Goetz, 2011). As mentioned in Chapter 2 (section 2.1.3.i), one reason for the effects of education on comprehension may be that English language education is beneficial to the acquisition of English language vocabulary knowledge which is theorised to be vital to successful comprehension (e.g., Brysbaert et al., 2016; LARRC, 2015). Another reason is that education, in English language as well as in other languages, may predict the use of metacomprehension strategies, such as selective rereading, which may be beneficial to comprehension (e.g., Hong-Nam & Page, 2014; Kern, 1994; van den Broek & Helder, 2017; Zabrucky et al., 2012).

Although the use of metacomprehension strategies, such as selective rereading, may be beneficial to comprehension (e.g., Zabrucky et al., 2012), many readers may not know when to use these strategies. There is evidence to suggest that individuals' metacomprehension judgements often diverge from their actual comprehension levels (e.g., Maki, 1998) (Chapter 2, section 2.2). If the accuracy of metacomprehension judgements is low, it is uncertain whether metacomprehension judgements predict comprehension of health-related texts. Concomitantly, it is unclear whether the use of end-user evaluations of health-related texts (NHS England, 2018a; see also section 4.1 of Chapter 4) ensures high-levels of understanding of these texts for the end-users. Thus, for theoretical and practical reasons, it is necessary to investigate the utility of metacomprehension judgements in predicting comprehension of health-related texts. Importantly, the effects of metacomprehension on tested comprehension should be examined in the presence of predictors of metacomprehension of health-related texts, such as education, age, and health literacy (Chapter 6). As mentioned at the beginning of this chapter, if tested comprehension has the same predictors as metacomprehension judgements, and metacomprehension judgements predict comprehension performance, then the producers of health-related documents could potentially use metacomprehension judgements as effective comprehension performance predictors.

However, tested comprehension may not have the same predictors as metacomprehension, and the direction of the effects of these predictors does not have to be the same. For example, the effects of age on metacomprehension reported in Chapter 6 (section 6.3.2.ii) indicate that older adults may think that their understanding of health-related texts is higher compared to that of younger adults, but research evidence shows that older individuals may be less likely to understand health-related texts compared to younger individuals (e.g., Alberti & Morris, 2017; Hannon & Daneman, 2009; Kobayashi et al., 2015; 2016; Liu et al., 2009) (Chapter 3, section 3.1.2; section 3.3). In combination, these findings suggest that the accuracy of metacomprehension judgements of older individuals could be relatively low (e.g., Dunlosky et al., 2006; Miles & Stine-Morrow, 2004) (Chapter 6, section 6.4), and that ageing could have a detrimental effect not only on the speed of processing measures (Chapter 3, section 3.1.2), but also on comprehension of health-related texts (Chin et al., 2011; Kobayashi et al., 2015; 2016). These conclusions are speculative as some studies investigating the effects of age on comprehension of health-related texts used tests of health literacy to assess understanding of health information (e.g., Kobayashi et al., 2015; 2016). Consequently, to verify the findings of some previous research (e.g., Liu et al., 2009), the effects of ageing on comprehension should be measured alongside the effects of health literacy.

Critically, research evidence indicates that functional health literacy is likely to predict comprehension of health-related texts (e.g., Chin et al., 2015; 2018; Liu et al., 2009) (Chapter 3, section 3.1.1). One potential explanation for the effects of health literacy on comprehension may be that functional health literacy encapsulates health knowledge (Chin et al., 2011). In turn, health knowledge may be crucial to comprehension of health-related texts as it constitutes background knowledge in relation to health information. From the theoretical perspective, background knowledge is thought to be important to the development of a logical situation model, and thereby successful comprehension (Kintsch & Rawson, 2007; van Dijk & Kintsch, 1983). Consequently, measurement of health literacy is vital in an investigation of predictors of comprehension of health information.

The effects of reader characteristics, such as variation in health literacy, on comprehension may be modulated by the effects of variation in text features. Indeed, there is some evidence to indicate that readers are more reliant on background knowledge in successful comprehension when texts require them to engage in inference-making due to texts' characteristics, such as low levels of coherence and cohesiveness (e.g., Hamilton & Oakhill, 2014; McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996; van den Broek & Helder, 2017) (Chapter 1, section 1.5). If text coherence and cohesion can have a differential effect on readers with different levels of background knowledge, it is reasonable to assume that the effects of other reader characteristics could modulate the impact of the effects of text features on comprehension of health-related texts (e.g., Francis et al., 2018). I discuss the effects of text features on comprehension next.

Researchers working in the text and discourse framework (Chapter 1, section 1.5) argued that some text properties, such as coherence (e.g., van Dijk & Kintsch, 1983), are important to successful comprehension (e.g., Britton & Gülgöz, 1991; Kintsch, 1988; Kintsch, 1998; Lehman & Schraw, 2002; Linderholm et al., 2000; McNamara & Kintsch, 1996; Vidal-Abarca et al., 2000). However, it is now thought that the effects of text features on comprehension cannot be studied without the consideration of the effects of individual differences. This is because relatively recent empirical evidence indicates that the effects of text features interact with the effects of reader attributes, whereby some textual features are likely to predict comprehension to a varying extent for different groups of the population (e.g., Francis et al., 2018; Kulesz et al., 2016; Liu et al., 2009; McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996; O'Reilly & McNamara, 2007; Ozuru et al., 2009) (Chapter 1, sections 1.5 and 1.6). Critical to this thesis is Liu et al's. (2009) study (discussed in Chapter 3, section 3.3), as their mixed-effects models provide evidence to suggest that variation in Flesch Reading Ease (FRE; Flesch, 1948) and text coherence predicts comprehension of health-related texts differently for different people (see Chapter 3, section 3.3).

Comprehension of health-related texts is likely to be predicted by variation in a range of text features differently for different readers (e.g., Liu et al., 2009). This is important from the perspective of both readability formulae, and NHS guidelines (Chapter 4, section 4.1) for writing comprehensible health-related texts. Specifically, the readability formulae and NHS guidelines assume that text comprehension can be uniformly improved for all individuals by modifying texts such that, for example, texts contain a high proportion of short words and sentences (Flesch, 1948; Plain English Campaign, 2018). However, due to contradictory findings relating to the effects of text features such as word length on comprehension (Chin et al., 2018; Friedman & Hoffman-Goetz, 2007; Leroy & Kauchak, 2014; Liu et al., 2009), the utility of readability formulae, and some NHS guidelines (e.g., Plain English Campaign, 2018), in improving comprehension can be questioned (Kauchak & Leroy, 2016) (Chapter 3, section 3.2.1). Thus, there is a clear theoretical and practical need to examine the joint effects of variation in range of reader characteristics, text features, and readability formulae estimates, on comprehension of health-related texts.

There are several candidate text features, relating to the guidelines used by the NHS health information writers (Chapter 4, section 4.1), that could predict comprehension of health-related texts. Critically, most of these candidate text features were discussed in previous chapters (refer to: Chapter 1, sections 1.5 and 1.6; Chapter 4, section 4.4, Table 4.1; Chapter 5, section 5.2.2, Table 5.1). However, some theoretically promising text features that were included in this study, and were not previously mentioned in this thesis, are discussed and justified later (section 7.2.3.iii). Readability formulae, such as the FRE (Flesch, 1948) and the Coh-Metrix L2 Readability Index (RDL2; Crossley et al., 2008), are promising predictors of comprehension from the theoretical perspective, as they were invented to be used as proxies for text comprehensibility (e.g., Flesch, 1948). In addition, these readability formulae partially map onto some of the guidelines used by the NHS, such as preference for short words (Plain English Campaign, 2018), straightforward words (Marsay 2017a; 2017b), plain language (NHS England, 2018a), and simple words (NHS England, 2018b). However, it is unclear if variation in text features, such as word length, and readability scores of health-related texts predicts tested comprehension of health-related texts (e.g., Chin et al., 2018; Friedman & Hoffman-Goetz, 2007).

One of the problems of using readability formulae estimates to assess difficulty of health-related texts is that different readability formulae may be measuring different constructs of readability (Chapter 5, sections 5.1 and 5.4). This is because they are based on different regression formulae that were calculated using different weights and validated on different datasets (Chapter 3, section 3.2.1). Furthermore, it is unclear whether readability estimates relate to perceived or tested comprehension (e.g., Leroy & Kauchak, 2014) (Chapter 6, section 6.4.2). Revisions of health-related texts that improve tested understanding, do not necessarily meaningfully change the readability scores, and linguistic features, of texts (e.g., Chin et al., 2018). There is some evidence to suggest that comprehension of health-related information can be improved by designing health-related texts in such a way as to minimise the processing demands required to understand these texts by building on individuals' general and health knowledge to scaffold their understanding (Chin et al., 2018). These revisions do not necessarily involve rewriting the passage to change its readability estimates, as they can consist of organisational revisions, such as changing the order of presentation of information and the use of signalling devices such as titles. Thus, as mentioned in Chapter 3 (section 3.1.1), readability may be a relatively bad proxy of comprehensibility, and increasing readability of texts alone may not be sufficient to improve understanding of health-related texts (Chin et al., 2018).

Overall, there is a theoretical and practical need to examine whether variation in theoretically and empirically important reader characteristics, text features, and readability formulae estimates, predicts comprehension of health-related texts. This is because the findings of comprehension researchers are often contradictory (e.g., Chin et al., 2018; Francis et al., 2018; Freed et al., 2017; Friedman & Hoffman-Goetz, 2007; Leroy & Kauchak, 2014; Liu et al., 2009). Thus, it is not clear what text features, and individual differences, are predictors of comprehension of health-related texts. Critically, there is some evidence to suggest that variation in some text features predicts perceived comprehension, but may not predict tested comprehension (e.g., Leroy & Kauchak, 2014) (Chapter 6, section 6.4.2).

Concomitantly, it is unclear whether following the guidelines used by the NHS (e.g., Plain English Campaign, 2018), predicts understanding of health-related texts of all adults. Indeed, theoretical research evidence (e.g., McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996; O'Reilly & McNamara, 2007; Ozuru et al., 2009), and empirical research evidence (e.g., Francis et al., 2018; Kulesz et al., 2016; Liu et al., 2009), indicates that the effects of linguistic features on comprehension are likely to vary between different individuals. Therefore, it is probable that the guidelines used by the NHS health information writers do not improve understanding for all individuals. Consequently, there is a methodological need for an examination of the predictors of comprehension of health-related texts using a mixed-effects model approach (Chapter 1, section 1.6).

Mixed-effects models allow to consider the way that the effects of health-related texts features on comprehension may vary in strength, depending on reader characteristics (e.g., Francis et al., 2018). This is important to this investigation, as the mixed-effects model approach could potentially permit the development of new guidelines that are adapted for different groups of users rather than for average NHS patients only. Critically, to examine the utility of revising health-related texts in accordance with any potential guidelines, it is also important to investigate whether metacomprehension judgements are predictors of actual comprehension of health-related texts. Notably, in the potential absence of the effects of variation in text features on comprehension, reader panel members judgements could potentially be used as effective comprehension performance predictors. In addition, from the practical perspective, if metacomprehension judgements and variation in text features and readability estimates do not predict comprehension of health-related texts, it may be the case that focusing on other interventions, targeting individuals rather than texts, could be more effective in improving health outcomes.

7.1.1. Research Aims

In this study I aimed to examine whether variation in reader characteristics and linguistic features of health-related texts predicts comprehension of health information using mixed-effects models of reading (see Chapter 1, section 1.6). I also aimed to investigate how the effects of textual features on comprehension of health-related information may vary for different kinds of readers. Another purpose of this study was to build on the findings of the previous studies (discussed in Chapters 5 and 6), in order to provide an evidence base for the development of easy-to-understand health-related texts tailored to different groups of users.

7.1.2. Research Questions

RQ7.1. How do reader attributes predict comprehension of written health-related information?

RQ7.2. How do textual characteristics predict comprehension of written health-related information?

RQ7.3. How do the effects of reader attributes and textual characteristics interact in predicting the comprehension of health-related information?

7.1.3. Hypotheses

$H_{7.1}$. Based on the literature reviewed in this thesis, several individual differences variables should predict comprehension of health-related information.

$H_{7.2}$. Text features related to coherence, and both readability formulae, should predict comprehension of health-related texts.

$H_{7.3}$. The effects of reader attributes are likely to interact with the effects of text features to predict reading comprehension of health-related texts.

7.2. Method

7.2.1. Ethics

The ethical approval for the study was granted by the Lancaster University's Research Ethics Committee in June 2016. To comply with ethics, prior to taking part in the study, participants read an information sheet describing the procedure of the study and signed an informed consent form in English (see Appendix D). Data collection took place between February 2017 and January 2018, and each testing session lasted between 60 to 90 minutes depending on the speed of participant's progress.

7.2.2. Participants

The sample of participants consisted of 200 (N_{female} = 117, N_{male} = 83) English-speaking adults, living in the UK, aged 20 to 88 years (M_{age} = 42.58, *SD* = 16.40). Half of the participants were native English speakers and the rest were L1 Polish speakers of ESL with varying levels of self-rated English language proficiency; 36 ESL participants self-rated their English proficiency as at a beginner level; 45 as intermediate; 19 as advanced. The participants also varied in their educational background: 34 completed secondary school only; 116 completed further education only; 20 were students; and 30 completed higher education (Table 7.1).

The decision to recruit only L1 Polish speakers as ESL readers was based on two practical reasons. First, L1 Polish speakers constitute the biggest foreign-born language minority in the UK (Office for National Statistics, 2016). Consequently, the findings of this study could potentially be applicable to a significant proportion of the UK's native and foreign-born population. Second, my L1 is Polish. Thus, at the data collection stage I could interact with L1 Polish low-proficiency ESL readers in their L1 to establish a relatively good rapport. Furthermore, I was able to make some of the tests more accessible to low-proficiency ESL readers, as I was able to score their responses in Polish. Overall, in addition to native English speakers, it made practical sense to focus on L1 Polish speakers, as I would not have been able to score answers to comprehension questions for any other language group of readers.

Table 7.1. Participants by English proficiency, education, and age.

Proficiency	Education	Number	Mean age (*SD*)
Native	Higher education	10	45.10 (17.51)
	University student	10	21 (.90)
	Further education	46	49.70 (16.92)
	Secondary school	34	57.85 (12.95)
Advanced	Higher education	5	39 (6.19)
	University student	10	21 (1)
	Further education	4	41.50 (14.94)
	Secondary school	-	-
Intermediate	Higher education	13	36.69 (10.37)
	University student	-	-
	Further education	32	34.31 (8.64)
	Secondary school	-	-
Beginner	Higher education	2	29.50 (4.55)
	University student	-	-
	Further education	34	41.06 (11.14)
	Secondary school	-	-

Note. Proficiency refers to self-reported English proficiency. Number refers to the number of participants within a particular group.

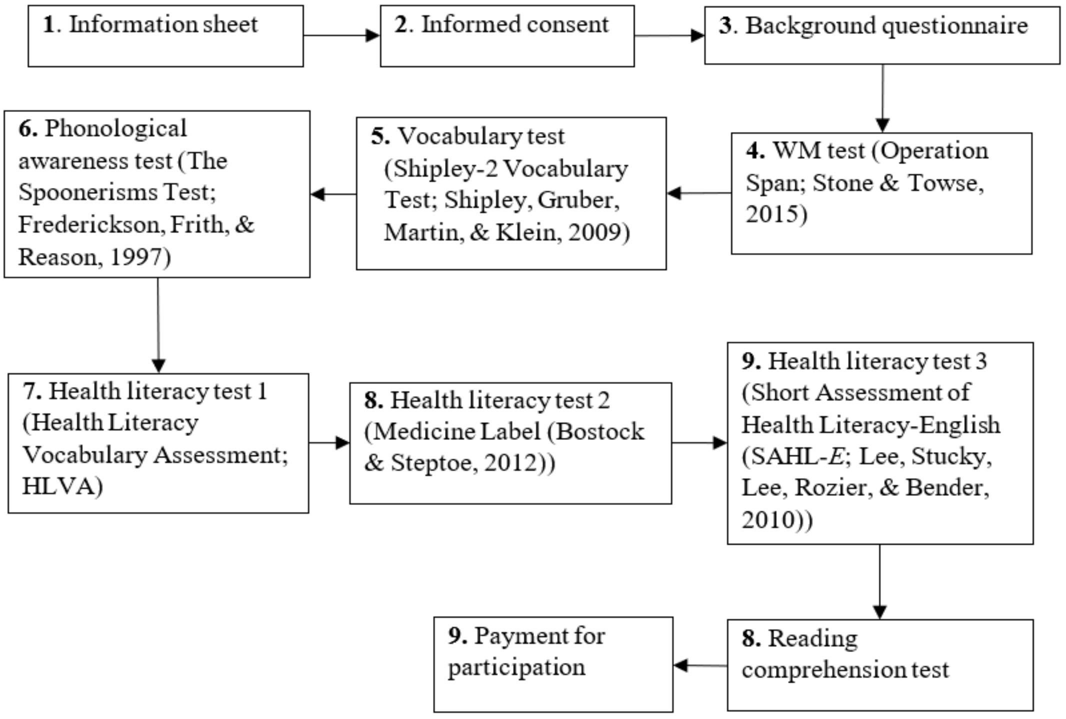
7.2.3. Materials and Procedure

7.2.3.i. Study Overview

All participants completed: a background questionnaire, a WM task, a test of vocabulary knowledge, an assessment of phonological awareness, three assessments of health literacy, and a reading-comprehension-of-health-related-information-test (see Appendix D for the list of materials; see Figure 7.1 for a visual illustration of the research procedure). The participants were tested on three measures of health literacy to mitigate the risk of potential health literacy ceiling effects. Health literacy ceiling effects were thought to be likely as they appeared in Study 2 with a standardised measure of health literacy, namely the UK-S-TOFHLA (von Wagner et al., 2007; see Chapter 6, section 6.2.2.i). Consequently, different measures of health literacy were used in this study. I used two standardised measures of health literacy alongside a new measure involving oral production of definitions of varying in frequency health-related terms that I developed for the purpose of this study (described later in this section).

The order of administration of all the measures was counterbalanced. For half of the participants the order was as shown in Figure 7.1, whereas for the rest the order was reversed. In addition to counterbalancing the administration of all measures, the order of reading comprehension questions on the reading comprehension test was also counterbalanced. Specifically, at the beginning of each testing session, every participant was assigned to one of four counterbalancing conditions. In each condition, the order of the health-related texts and questions differed. Next, I describe the study materials in detail.

Figure 7.1. Research procedure.


1. Information sheet
2. Informed consent
3. Background questionnaire
4. WM test (Operation Span; Stone & Towse, 2015)
5. Vocabulary test (Shipley-2 Vocabulary Test; Shipley, Gruber, Martin, & Klein, 2009)
6. Phonological awareness test (The Spoonerisms Test; Frederickson, Frith, & Reason, 1997)
7. Health literacy test 1 (Health Literacy Vocabulary Assessment; HLVA)
8. Health literacy test 2 (Medicine Label (Bostock & Steptoe, 2012))
9. Health literacy test 3 (Short Assessment of Health Literacy-English (SAHL-*E*; Lee, Stucky, Lee, Rozier, & Bender, 2010))
8. Reading comprehension test
9. Payment for participation

Background Questionnaire. The background questionnaire collected information about participants': age; gender; native language; self-rated English language proficiency (beginner, intermediate, and advanced); and educational background (Appendix D). It also asked three screening questions which were used as a brief measure of health literacy (Chew, Bradley, & Boyko, 2004). These three questions were found to be the best predictors of inadequate health literacy from a sample of 16 screening questions (Chew et al., 2008). In addition, they were validated against two standardised tests of health literacy, including the S-TOFHLA (Baker et al., 1999; Chew et al., 2008). Each of the three screening questions includes a five-item response scale. Depending on the screening question, responses of "Never", "Occasionally", "Extremely", and "Quite a bit", indicate adequate health literacy, whereas the remaining options on the response scale suggest inadequate health literacy. Each response is associated with a score ranging from 1 to 5.

Operation Span. Stone and Towse's (2015) operation span task was built using Tatool, a Java-based framework (von Bastian, Locher, & Ruffin, 2013). This task is based on Daneman and Carpenter's (1980) paradigm to capture simultaneous memory and processing operations of verbal WM. I used this task because there is evidence to suggest that tests which measure WM capacity and processing, such as reading and operation span, are better predictors of comprehension than tests that measure WM capacity alone (e.g., Daneman & Merikle, 1996). In the operation span task participants are presented with numbers made of one or two digits that must be remembered and recalled at the end of the trial in the correct order (scoring is described in section 7.2.3.ii). For every item that must be remembered there is a processing stage succeeding it. The processing stage involves a mathematical operation, such as "1 + 2 = 3", where the participant has to decide whether the solution is correct or incorrect. The operation can be a division, subtraction, multiplication or addition. Each time, both items to be remembered, and mathematical operations, are randomly generated.

Shipley-2 Vocabulary Test. The Shipley-2 Vocabulary Test (Shipley et al., 2009) is a standardised 40-item multiple-choice test. For each item participants must select a synonym of that item from one of four available options. If the participant selects the correct synonym, they receive one point for that item. For example, for an item *Quotidian* the options are *travesty/everyday/calculation/promise*, where choosing *everyday* scores one point as it is the synonym of the item. The test begins with frequent, well known words that are used in daily conversations, to progressively rarer words which are seldom used. The maximum score a participant can achieve is 40.

The Spoonerisms Test. The Spoonerisms test (Frederickson et al., 1997; Walton & Brooks, 1995) was originally developed as a measure of phonological awareness of older children, but it can also be used to assess adults as standardisation norms extend to adulthood. It is comprised of two sections. The first section uses semi-Spoonerisms where the participant must replace the first sound of a word with a different sound, for example *red* with a /b/ gives *bed*. The second section uses full Spoonerisms where the participant is required to swap the first sounds of two words, for example *fed man* gives *med fan*. The words are presented orally, and each section has a time limit of three minutes. Both parts are discontinued if a participant makes three mistakes in a row or takes more than three minutes on a section. The total score for the Spoonerisms test is based on the combined scores a participant achieves in the first and second section. A participant can score a maximum of 10 points in the first section, and 20 points in the second section. Thus, the maximum total score is 30.

HLVA. Research evidence indicates that some of the current health literacy measures, such as the S-TOFHLA, may be prone to exhibiting ceiling effects (Morrison, Schapira, Hoffmann, & Brousseau, 2014). Consequently, the standardised tests of health literacy may not be as sensitive to detecting variation in health literacy as it is claimed that they are (e.g., Baker et al., 1999). To overcome this potential limitation, I developed the Health Literacy

Vocabulary Assessment (HLVA) with the intention of constructing a measure of adults' health literacy that has relatively high discriminative power.

During the HLVA administration, the experimenter reads out a list of 22 varying in frequency medical terms, obtained from Oxford's Concise Medical Dictionary (Martin, 2015), which the participants can also see and read themselves. The 22 chosen words provide a relatively wide range of high and low frequency items (see Table 7.2 in Appendix E), as judged using the frequency values obtained from the BNC (BNC Consortium, 2007) and SUBTLEX-UK (van Heuven et al., 2014) corpora. Consequently, in theory, the 22 words should distinguish between participants of different health literacy levels relatively well. After hearing each item in English, participants are asked to verbally define it. The test has no time limit, but the administration takes approximately 10 minutes. It should be noted that within the field of applied linguistics, permitting ESL speakers to respond to test items in their L1 is encouraged (e.g., Bowles, 2018; Mackey & Gass, 2016; Pavlenko, 2007). This is because asking ESL readers to provide responses in English is likely to result in incomplete answers if these readers cannot express their full range of thoughts due to limited English language proficiency (Bowles, 2018; Mackey & Gass, 2016; Pavlenko, 2007). Thus, to allow L1 Polish readers to express their thoughts fully, regardless of their English language proficiency, a decision was made to accept responses in Polish as well as English, and a scoring key has been devised for answers in both languages (scoring is described in section 7.2.3.ii).

MEDCO Medicine Label. This measure of health literacy consists of a four-item comprehension assessment based on instructions similar to those found on a packet of common painkillers (Bostock & Steptoe, 2012). During the administration, participants are asked to read a made-up medicine label, and are asked four questions relating to the usage of that medicine, such as "What is the maximum number of days you may take this medicine?".

Correct answers score 1, whereas incorrect answers score 0. The maximum score for this measure is 4.

SAHL-*E*. The Short Assessment of Health Literacy-*English* (SAHL-*E*; Lee et al., 2010) is a standardised measure of health literacy which involves showing participants 18 flashcards. Each flashcard has a medical test term, a key word with a related meaning, and a distractor word unrelated in meaning to the test term. The participants are asked to pronounce the test term, and the word which is most closely associated with that word. Consequently, the test measures both comprehension and pronunciation of health-related terms. Specifically, on each flashcard, participants can score half a point for their accuracy of pronunciation and half a point for their understanding of the test term. The maximum score is 18.

Reading Comprehension Test. I developed this test for the purpose of measuring reading comprehension of health-related texts used in this study. It is based on four health-related texts from a sample of eight texts used in Study 2 (described in Chapter 6, section 6.2.2.i). These eight texts were chosen from an opportunity sample of 86 health-related texts which were analysed in Study 1 (described in Chapter 5, section 5.2.1). The four health-related texts were selected based on their contrasting readability scores, as well as their relatively different mean perceived understanding and effort judgements (Chapter 6, section 6.3.1; Table 7.3).

Table 7.3. Characteristics of four chosen health-related texts.

Characteristics	Text Number			
	2	3	5	7
Perceived Understanding*	8.03	7.71	8.09	7.74
Perceived Effort*	2.52	3.45	2.37	2.82
FRE	51.45	45.61	70.27	70.23
RDL2	9.47	10.89	20.82	14.6

Note. * mean judgements.

The test incorporates two judgement scales, used in Study 2, to examine whether self-reported perceptions of understanding, or metacomprehension, predict tested comprehension. The selected judgement scales are: perceived effort exerted on understanding each health-related text, ranging from 1 (no effort at all) to 9 (a lot of effort); and perceived understanding of each health-related text, ranging from 1 (not well understood at all) to 9 (extremely well understood). As mentioned in Chapter 6 (section, 6.2.2.i), nine-point judgement scales were chosen as they were found to outperform their counterparts with fewer categories in terms of criterion validity, test-retest reliability, and discriminative power (Preston & Colman, 2000).

To have as valid measure of comprehension as possible, I tested comprehension of health-related information using both multiple-choice and open-ended questions (Cain & Oakhill, 2006). I decided to use two multiple-choice questions with three response items, and four open-ended questions requiring a verbal response, per text. This decision was influenced by the feedback that I received during the pilot study, which I discuss in the next paragraph. In addition, to replicate the real-life context in which health information is read, thereby ensuring high ecological validity, I permitted individuals to refer to texts when answering comprehension questions. Critically, to allow ESL readers to express their full range of thoughts (Bowles, 2018; Mackey & Gass, 2016; Pavlenko, 2007), L1 Polish participants could answer the open-ended comprehension questions in either Polish or English or using a combination of the two languages (scoring is described in section 7.2.3.ii).

Pilot Study. I consulted with language testing experts prior to designing the reading comprehension test. I was advised to use multiple-choice questions with three response items instead of four as research evidence indicates no meaningful changes in the psychometric properties of three option multiple-choice items when compared to the traditional multiple-choice questions with four or more options (e.g., Royal & Dorman, 2018). I also consulted with a medical expert at Blackpool NHS Trust, to determine whether the open-ended questions were appropriate questions to ask given each health-related text, and to establish whether the answer key was appropriate. As a consequence of these discussions, both the questions and answers were altered in accordance with suggestions of the medical expert.

In addition, I tested all the materials with four participants, prior to the data gathering process, to highlight ambiguities for which adjustments could be made. I found that participants were scoring near ceiling on the reading comprehension of health-related texts measure. Thus, I decided to alter the scoring procedure on that measure to make the test more difficult, requiring the answers to two open-ended questions per text being conditional on a two-component response to be scored as correct. I discuss the scoring procedure of the reading comprehension test in the next section, where I describe the scoring process in the order of the administration of the measures, thereby starting with the Operation Span (Stone & Towse, 2015).

7.2.3.ii. Scoring Choices

Operation Span. The Operation Span (Stone & Towse, 2015) test generates data in the form of an excel file. For interpretation, this data must be further processed using software, such as R (R Core Team, 2019). Further processing involves calculating the desired metric for the measurement of verbal WM. I made the decision to use the proportion of correctly recalled responses across all trials as a measure of WM in this study, as research evidence indicates that the proportion correct scoring method is a more reliable than other scoring methods, such as maximum span (e.g., Friedman & Miyake, 2005).

HLVA. The scoring procedure for the HLVA has been devised using the definitions of the 22 medical terms from the Cambridge English Dictionary (Cambridge University Press, 2018) and Oxford’s Concise Medical Dictionary (Martin, 2015). For the purpose of this study and given that half of the participants were L1 Polish speakers, the definitions used for scoring correct responses have also been translated into Polish, allowing Polish participants to define each term in Polish if they chose to do so. The scoring sheet (see Appendix D) includes acceptable definitions in both English and Polish of the target items. Each item has a maximum score of 2, and to score full marks the participant must define each item using at least two key components that form each definition (bolded and highlighted in yellow in Appendix D). The maximum score a participant can achieve is 44.

Reading Comprehension Test. As mentioned in section 7.2.3.i, each health-related text contained multiple-choice questions and open-ended questions. For each multiple-choice question, there was only one correct response, whereas for each open-ended question the accuracy of an answer was determined based on the precision of the answer, and, in two open-ended questions per text, the level of detail provided (see Appendix D for the answer key). Although the more difficult open-ended comprehension questions were partially scored, if a participant provided only one part of the answer, they received a score of zero for that question rather than a score of .5. I made the decision to score these questions in this way, as allowing for partial scoring of only two questions per text would be likely to lead to models with degenerate estimates in the analysis. In addition, some of the open-ended questions did not ask for enough detail to require a two-component response. Thus, each question was scored as binary accuracy data (correct vs. incorrect). Consequently, the maximum achievable score for the comprehension test was 24, as there were six questions per each one of the four health-related texts.

It is important to mention that the test had relatively high inter-rater reliability. Specifically, after data collection, the 3200 responses to open-ended comprehension questions were audio recorded and marked by the author of this thesis and another L1 Polish PhD student. The inter-rater reliability indicated strong level of agreement ($Cohen's\ K = .88$). The *Cohen's K* statistic can be interpreted as showing that, adjusted for guessing, approximately 77% of the data obtained using the outcome measure could be rated as being reliable (McHugh, 2012). Next, I briefly describe the choice that guided the inclusion of the chosen predictor variables in the analyses.

7.2.3.iii. Variable Selection

The participant-related predictors were chosen based on the literature reviewed in Chapters 1, 2, and 3. Specifically, the chosen participant-related predictors included: age, self-rated English language proficiency, verbal WM, vocabulary knowledge, phonological awareness, health literacy, educational background, and metacomprehension judgements (Table 7.4 lists these predictors alongside a brief justification for their inclusion). In addition to participant-related predictors, this study included text-level predictors. The chosen text-level predictors were derived using the Coh-Metrix application (Graesser et al., 2004) (see Chapter 5, section 5.2.1, for a description of Coh-Metrix). These predictors were based on previous research discussed in the literature review chapters, as well as the current guidelines employed by the NHS England Trusts (e.g., NHS England, 2018a; Plain English Campaign, 2018). It is important to mention that most of the text features included in this study were already discussed, and justified, in the preceding chapters of this thesis (Chapter 4, section 4.4, Table 4.1; Chapter 5, section 5.2.2, Table 5.1). Thus, given the limited word count and to avoid repetition, only the additional text-level predictors, and the associated justifications, that were not mentioned in previous chapters of this thesis are discussed here (Table 7.5).

Table 7.4. Individual differences predictors included in Bayesian models of comprehension

Reader characteristics	Justification for inclusion
Vocabulary knowledge	• Theoretically, vocabulary knowledge is thought to be critical to comprehension (e.g., Kintsch & Rawson, 2007; Perfetti, 2010; Tunmer & Chapman, 2012) (Chapter 1). This is because it is hypothesised that vocabulary knowledge enables the reader to form propositions (Kintsch & Rawson, 2007) (Chapter 1, section 1.2). In turn, propositions are thought to be crucial to understanding text as they are a prerequisite to constructing a textbase and a logical situation model of the text read (Kintsch & Rawson, 2007). • In addition to theoretical accounts, research evidence consistently indicates that vocabulary knowledge is a reliable predictor of comprehension in the presence of other individual differences predictors (e.g., Freed et al., 2017).
Verbal WM	• WM is theorised to be important to comprehension as it is thought to be necessary for processing of the textual information read, therefore the construction of propositions, the textbase, and the situation model (Kintsch & Rawson, 2007; Perfetti & Stafura, 2014; Zwaan, 2016) (Chapter 2, section 2.1.1; Chapter 3, section 3.1.1). • However, research evidence is conflicting, as there is robust evidence to suggest that WM does not predict comprehension in the presence of vocabulary knowledge (e.g., Freed et al., 2017; Van Dyke et al., 2014) (cf. Liu et al., 2009). • Thus, WM is argued to be spuriously related to comprehension through its' association with direct predictors of comprehension, such as vocabulary knowledge (Freed et al., 2017; Van Dyke et al., 2014).
Phonological awareness	• Phonological awareness is thought to be critical, according to the Simple View of Reading account's (SVR; Tunmer & Chapman, 2012), for successful reading comprehension. Specifically, it is thought to be important to attainment of reading fluency (Perfetti, 1992; Chapter 1, section 1.4). In turn, reading fluency is hypothesised to be central to comprehension as it is thought to allow the reader to devote more mental resources to generating the meaning of a text read (Perfetti, 1998), thereby to the construction of a logical situation model (Kintsch & Rawson, 2007; Chapter 2, section 2.1.1). • However, research evidence is conflicting, as there is robust evidence to suggest that phonological awareness does not predict comprehension in the presence of vocabulary knowledge amongst adult readers (Freed et al., 2017). Indeed, it has been argued that phonological awareness predicts comprehension in the absence of vocabulary knowledge as it shares variance with vocabulary knowledge (Freed et al., 2017).
English language proficiency	• Population of the UK is linguistically relatively diverse (Office for National Statistics, 2016), therefore NHS patients are also of different language backgrounds. • Critically, English language proficiency is theorised to predict English vocabulary knowledge (Brysbaert et al., 2016). In turn, the higher the vocabulary knowledge, the higher the lexical quality and the more efficient lexical access are thought to be (Perfetti, 2007; 2010). Thus, the more proficient L2 readers are likely to be better comprehenders of L2 texts than the less proficient L2 readers (Chapter 2, section 2.1.3i). • There is also evidence to suggest that self-rated English language proficiency is likely to predict comprehension of health-related information written in English (e.g., Thomson & Hoffman-Goetz, 2010; Todd & Hoffman-Goetz, 2011).

Health literacy	• Background knowledge is theorised to be critical to comprehension, as the development of a mental model is thought to require that information provided by the text is integrated with reader's relevant background knowledge (Kintsch & Rawson, 2007). Thus, without some relevant background knowledge the reader cannot fully comprehend the text read (Chapter 1, section 1.2). • Inference-making, which is critical to understanding incoherent and incohesive texts (e.g., Hamilton & Oakhill, 2014), is also thought to be reliant on reader's background knowledge (e.g., van Dijk & Kintsch, 1983; van den Broek & Helder, 2017). • Research evidence indicates that high background knowledge is associated with higher comprehension (e.g., Kulesz et al., 2016), especially amongst those who are highly skilled readers (O'Reilly & McNamara, 2007; Ozuru et al., 2009). • Critically, health literacy encompasses health knowledge (Chin et al., 2011) which can be thought of as relevant background knowledge in the context of comprehending health-related texts. • Research evidence indicates that high health literacy and health knowledge predict higher comprehension of health-related texts (e.g., Chin et al., 2015; 2018). • In addition, health literacy levels are likely to vary between different groups of the population. Importantly, reader panel members' background health knowledge is likely to be higher compared to that of an average first-time NHS patient.
Age	• There is evidence from reading comprehension and health literacy literature to suggest that there are some age-related changes in comprehension of health-related texts (e.g., Alberti & Morris, 2017; Hannon & Daneman, 2009; Kobayashi et al., 2015; 2016; Liu et al., 2009) (Chapter 3, section 3.1.2; section 3.3). Specifically, this evidence indicates that older individuals may be less likely to understand health-related texts compared to younger individuals (Liu et al., 2009). • However, there is also contradictory evidence from the field of cognitive psychology suggesting that ageing is associated with changes in the speed of processing, rather than changes in cognitive and linguistic capacities, including reading comprehension (e.g., Chin et al., 2011; 2015; Davies et al., 2017; Li et al., 2004; Ramscar et al., 2017) • Thus, it is theoretically important to examine whether ageing predicts comprehension of health-related texts. • Examining age effects on comprehension is also important from the practical perspective, as NHS patients are of different ages.
Educational background	• There is evidence to suggest that higher education level is associated with higher reading comprehension of monolingual and ESL readers (e.g., Todd & Hoffman-Goetz, 2011). • One explanation for this may be that education is thought to play an influential role in the attainment of vocabulary knowledge (e.g., LARRC, 2015) (Chapter 2, section 2.1.3i), as individuals are often exposed to new words through formal education. Thus, the more educated individuals may have higher comprehension because of their higher levels of vocabulary knowledge (e.g., Brysbaert et al., 2016). • Another explanation for the effects of education on comprehension could be that the more educated adults are more likely to use of self-regulatory active processes, such as re-reading of problematic information, to repair comprehension breaks (e.g., Hong-Nam & Page, 2014; Kern, 1994; Thiede et al., 2010; van den Broek & Helder, 2017; Zabrucky et al., 2012) (section 7.1; Chapter 2, section 2.2). These strategies may be beneficial to comprehension of health-related texts.

Metacomprehension	• Metacomprehension is theorised to be important to comprehension as it is thought that metacomprehension judgements can affect comprehension by contributing to whether individuals engage in specific reading behaviours that regulate comprehension breaks (Thiede et al., 2010) (Chapter 2, section 2.2.). • In turn, strategies that increase metacomprehension accuracy might enable greater self-regulation of reading behaviour, potentially improving comprehension (Thiede & Anderson, 2003). • However, given the relatively low metacomprehension accuracy reported in previous studies (e.g., Maki, 1998), it is questionable whether metacomprehension judgements predict tested comprehension. • In addition, NHS health-information writers rely on reader panel members' comprehensibility judgements when writing health-related texts. Thus, from the practical perspective, it is critical to examine whether metacomprehension judgements have the same predictors as tested comprehension, and whether they can be used as comprehension performance predictors.

Table 7.5. Additional text feature predictors included in Bayesian models of comprehension

Text Features	**Justification for inclusion**	**Possible relation to NHS guidelines**
Average sentence length	• Sentence length is a theoretically important candidate predictor as it constitutes part of FRE regression formula (Flesch, 1948), and its' potential effects on comprehension are contested. • It is thought that using shorter sentences improve comprehensibility (Flesch, 1948), as relatively long sentences may place greater demands on WM-reliant meaning-to-text integration than shorter sentences (Perfetti, 2007; Perfetti & Stafura, 2014; Yang et al., 2005) (Chapter 1, section 1.6). This may be problematic for comprehension of adults with relatively small WM capacities, such as older adults, as empirical research evidence suggests that ageing is negatively related with measures of processing speed, including WM resources (e.g., Hannon & Daneman, 2009; Kobayashi et al., 2015; Li et al., 2004) (Chapter 3, section 3.3). • However, there is also evidence to suggest that decreasing sentence length may reduce text coherence (e.g., Crossley et al., 2008) (see also Chapter 3, section 3.2), as short sentences often omit cohesive markers. Increasing cohesion and coherence of texts, features associated with comprehension, frequently involves increasing average sentence length (e.g., Crossley et al., 2008; Hamilton & Oakhill, 2014; O'Reilly & McNamara, 2007; Ozuru et al., 2009). Therefore, reducing sentence length may not necessarily have a beneficial impact on comprehension, especially if shortening sentences reduces text coherence and cohesion (Ozuru et al., 2009) (Chapter 1, section 1.6).	• NHS's Brand Identity (NHS, 2015) and the Plain English Campaign (2018) guidelines specified the preference for keeping sentences short.

Temporal connectives	• Temporal connectives are connecting words, such as *then* , *after* , *during* , that link propositions and clarify relations in the text (Kintsch, 1998). • By helping readers to link propositions, connectives are theorised to aid the comprehenders in constructing the textbase (Kintsch & Rawson, 2007). Specifically, connectives are thought to increase text cohesion as there is evidence to suggest that they prompt readers to generate inferences passively when reading for understanding (e.g., Hamilton & Oakhill, 2014; van den Broek & Helder, 2017). Thus, texts that contain a relatively high proportion of connectives are unlikely to require as many reader-initiated processes to reach adequate levels of understanding as those that contain fewer connectives (Chapter 1, section 1.5). • It is thought that the incidence of temporal connectives can potentially reduce the WM demands of meaning-to-text integration processes, as the connections between the text and prior knowledge are theorised to be strengthened (Magliano & Schleich, 2000; Zwaan, 2016).	• No obvious relation.
Deep cohesion	• As mentioned in Chapter 1 (section 1.6), deep cohesion refers to a component score of the incidence of causal, temporal, and logical connectives (Crossley et al., 2011). • It is theorised that texts with high deep cohesion scores should be understood better than texts with low deep cohesion scores (e.g., Crossley et al., 2011; Kintsch & Rawson, 2007). This is because highly cohesive texts are thought to require fewer active reader-initiated processes to understand than texts that are less cohesive (e.g., O'Reilly & McNamara, 2007; van den Broek & Helder, 2017).	• No obvious relation.
Stem overlap	• Another measure of co-reference, meaning text coherence (Kintsch & Rawson, 2007), is stem overlap (Dowell et al., 2016). • Stem overlap is similar, and related, to argument overlap (McNamara et al., 2013). Specifically, it measures the overlap between nouns and content words that share a common lemma, such as *price* and *priced* . • Text coherence is thought to be critical to comprehension (e.g., Kintsch, 1988; Kintsch, 1998; McNamara & Kintsch, 1996). This is because texts that are not coherent may require the reader to establish coherence by building inferences to fill the gap between sentences using their background knowledge (Hamilton & Oakhill, 2014; van Dijk & Kintsch, 1983) (Chapter 1, section 1.5).	• No obvious relation.
Temporality	• Coh-Metrix (Graesser et al., 2004) measures temporality using indices of tense and aspect repetition. Temporality is of theoretical interest, as it is thought that texts that have more consistent temporality, specifically with regards to tense and aspect, are easier to process and understand (Crossley et al., 2012; Magliano & Schleich, 2000; McNamara et al., 2013). Specifically, texts with a high proportion of aspect repetition are thought to minimise the WM demands associated with meaning-to-text integration (Magliano & Schleich, 2000). However, temporality has been investigated in the context of narrative texts. Consequently, it is questionable whether informational texts, such as health-related texts with a high incidence of tense and aspect repetition, are easier to understand than those with a lower incidence (Zwaan, 2016).	• No obvious relation.

Sentence syntax similarity	• Sentence syntax similarity is a theoretically interesting candidate predictor of comprehension as it constitutes part of the RDL2 regression formula (Crossley et al., 2008), and its' potential effects on comprehension are contested. • This is because it is thought that texts with greater between-sentence uniformity of syntactic structures impose lower cognitive demands on the reader, permitting more WM resources to be spent on meaning integration processes that maintain coherence (Perfetti, 2007; Perfetti et al., 2007; Perfetti & Stafura, 2014). • The problem with this measure is that it is assumed that syntactically similar words are similar in ease of syntactic structures. However, especially in the case of health-related texts, it can be argued that a measure of similarity does not necessarily indicate simplicity, as sentences can be similar but syntactically complex (Dowell et al., 2016) (Chapter 3, section 3.2).	• Advocacy for straightforward words, simple words, and plain language (Marsay, 2017b; NHS England, 2018a; 2018b; Plain English Campaign, 2018).

7.2.3.iv. Analysis Choices

I discuss the correlations between predictor variables in this section, as some correlations determined variable selection and analysis choices in this study (correlations between predictors and outcome measure are discussed in section 7.3.1.ii). Specifically, many of the text features were highly correlated with each other (Table 7.6). One potential reason for this has been mentioned in Chapter 6 (section 6.3.1.ii). Namely, it is probable that the correlation coefficients between some text-level predictor variables were spuriously large due to a relatively low sample of health-related texts used in this study (Schönbrodt & Perugini, 2013) (section 7.4.1). Although, to provide a more complete picture of the reading comprehension processes in the context of health-related texts it was important to keep as many text features in the models as possible, some text-level variables had to be removed to improve the stability of the models, and of the estimates of the effects (Dormann et al., 2013).

To improve computational stability, and the stability of the estimates of text-feature effects, I removed the variables that were perfectly, or near perfectly, correlated. I removed sentence semantic overlap, as measured using the Latent Semantic Analysis (see Chapter 5, section 5.2.2 for a description of LSA), since it was perfectly correlated with syntax similarity. I also removed word length as it was perfectly negatively correlated with the FRE.

Table 7.6. Correlations between text features and readability formulae scores of health-related texts.

	RDL2	FRE	Word length	Temporal connectives	All connectives	Stem overlap	Hypernymy noun	Hypernymy verb	Temporality	Deep cohesion	Referential cohesion	Causal connectives	CELEX frequencies	Sentence length	Passive voice	Syntax similarity	LSA	Causal cohesion	Hypernymy noun and verb	Logical connectives	Gerunds	BNC frequencies
FRE	.82***																					
Word length	-.82***	-1.00***																				
Temporal connectives	-.26***	-.61***	.65***																			
All connectives	.22***	.05***	-.11***	-.53***																		
Stem overlap	.23***	-.36***	.34***	.47***	.45***																	
Hypernymy noun	-.86***	-.42***	.42***	-.18***	-.23***	-.66***																
Hypernymy verb	.30***	.68***	-.64***	-.22***	-.65***	-.77***	.08***															
Temporality	-.65***	-.64***	.59***	-.19***	.61***	.19***	.52***	-.77***														
Deep cohesion	-.19***	.17***	-.21***	-.88***	.61***	-.41***	.50***	-.10***	.61***													
Referential cohesion	.80***	.36***	-.38***	-.01	.54***	.74***	-.94***	-.31***	-.23***	-.24***												
Causal connectives	.79***	.65***	-.61***	.21***	-.41***	.06***	-.73***	.59***	-.96***	-.65***	.48***											
CELEX frequencies	.56***	.61***	-.56***	.19***	-.68***	-.29***	-.41***	.83***	-.99***	-.59***	.10***	.92***										
Sentence length	-.79***	-.96***	.94***	.41***	.22***	.41***	.42***	-.82***	.82***	.06***	-.26***	-.79***	-.81***									
Passive voice	-.49***	-.65***	.70***	.94***	-.72***	.16***	.14***	-.05***	-.16***	-.77***	-.35***	.08***	.20***	.42***								
Syntax similarity	.41***	-.08***	.04**	.04**	.77***	.90***	-.68***	-.75***	.28***	-.04**	.86***	-.01	-.40***	.23***	-.29***							
LSA	.44***	-.02	-.02	-.05***	.82***	.86***	-.66***	-.72***	.29***	.04*	.86***	-.02	-.41***	.19***	-.37***	1.00***						
Causal cohesion	.25***	.32***	-.38***	-.84***	.91***	.06***	-.04**	-.32***	.50***	.83***	.34***	-.39***	-.55***	-.04**	-.92***	.47***	.55***					
Hypernymy noun and verb	-.28***	-.07***	.13***	.49***	-1.00***	-.52***	.31***	.66***	-.55***	-.54***	-.61***	.34***	.64***	-.19***	.70***	-.82***	-.86***	-.88***				
Logical connectives	-.89***	-.56***	.53***	-.21***	.07***	-.42***	.94***	-.24***	.77***	.61***	-.79***	-.91***	-.69***	.62***	.04**	-.40***	-.38***	.17***	.01			
Gerunds	-.77***	-.48***	.52***	.41***	-.78***	-.54***	.75***	.30***	.01	-.19***	-.91***	-.25***	.10***	.30***	.70***	-.82***	-.86***	-.69***	.82***	.56***		
BNC frequencies	-.49***	-.14***	.19***	.30***	-.91***	-.70***	.60***	.63***	-.32***	-.25***	-.83***	.06***	.43***	-.07***	.59***	-.94***	-.97***	-.73***	.94***	.32***	.93***	
Comprehension test score	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note. Significance values are based on Pearson's *r* . * = *p* < .05; ** = *p* < .01; *** = *p* < .001.

Verb and noun hypernymy was perfectly negatively correlated with the incidence of all connectives. Thus, I decided to remove it as there were other measures of hypernymy in the model. Last, I removed temporality from the model, as it was extremely highly correlated with CELEX word frequencies, and there was a related measure of the incidence of temporal connectives. Although, some problematic correlations remained, I took additional steps to mitigate potential collinearity issues between text-level predictor variables (Chapter 5, section 5.3.1.ii), which are discussed later in the results section (section 7.3.2.iv).

High correlations between text features were not the only culprits of potential collinearity problems, as there were also collinearity issues due to high correlations between individual differences variables which affected the analysis choices (Table 7.7[1]). The standard errors associated with self-reported English proficiency, Shipley-2 vocabulary, and SAHL-*E* measure of health literacy were distorted by multicollinearity as measured using the Variance Inflation Factor (VIF; Chapter 5, section 5.3.1.ii). Since these three variables were relatively highly correlated with each other (Table 7.7), to some extent they measured the same underlying construct. Consequently, I decided to standardise them, to ensure that they were on the same scale, and merge them. After merging, I mean averaged the new variable, which I named English language proficiency, as the new variable could be perceived to be a proxy for it. This is because it was reliant on self-reported English language proficiency, and measures of general and medical English vocabulary knowledge (Brysbaert et al., 2016).

Similarly, to measures of proficiency and vocabulary, perceived understanding and perceived effort judgements were highly correlated and shared a relatively large proportion of variance, indicating that to a certain extent they measured the same construct. Furthermore, the VIF indicated that they suffered from collinearity as the standard errors associated with

[1] It is important to clarify that education level may not relate to education in the UK. This is probably why the correlation between education level and English language proficiency is negative (see Table 7.7).

their estimates were severely distorted, suggesting that their estimates in the models were likely to be unstable and unreliable (Dormann et al., 2013). Therefore, in order to prevent biasing the estimates and improve computational stability, I merged the effort and perceived understanding ratings to create a new variable, which I named metacomprehension.

Table 7.7. Correlations between individual differences.

	Education level	English proficiency	Age	Health literacy (HLVA)	Health literacy (SAHL-*E*)	WM	Shipley vocabulary	Phonological awareness	Health literacy (MEDCO)	Perceived understanding	Perceived effort	Screening question (THLQ1)
English proficiency	-.05***											
Age	-.38***	.19***										
Health literacy (HLVA)	.22***	.65***	.06***									
Health literacy (SAHL-*E*)	-.01	.88***	.14***	.70***								
WM	.24***	.13***	-.54***	.24***	.20***							
Shipley vocabulary	-.06***	.83***	.29***	.67***	.85***	.11***						
Phonological awareness	.15***	.60***	-.07***	.64***	.68***	.44***	.66***					
Health literacy (MEDCO)	.20***	.45***	-.15***	.55***	.55***	.27***	.50***	.50***				
Perceived understanding	.06***	.75***	.13***	.60***	.75***	.15***	.71***	.58***	.44***			
Perceived effort	.03*	-.71***	-.18***	-.50***	-.68***	-.11***	-.69***	-.52***	-.38***	-.85***		
Screening question (THLQ1)	.00	.63***	.08***	.45***	.59***	.11***	.49***	.34***	.32***	.52***	-.46***	
Comprehension test score	.25***	.70***	-.15***	.69***	.76***	.34***	.65***	.63***	.62***	.65***	-.55***	.54***

Note. Significance values are based on Pearson's *r* . * = *p* < .05; ** = *p* < .01; *** = *p* < .001.

In addition to changes due to potential collinearity issues, some variables were removed for the reason of parsimony. Specifically, to limit the number of variables included in the model, thereby reducing the model complexity, only one of the three health literacy screening questions (Chew et al., 2004), from the background questionnaire, was included in the analyses. Namely, question "How often do you have someone help you read hospital materials?" (THLQ1) was retained, whereas the others were discarded. This was justified on the grounds of research evidence suggesting that of the three questions considered, THLQ1 was the most sensitive measure of health literacy (Chew et al., 2004). In addition, evidence indicated that combining the three screening questions did not improve the measure's sensitivity to detect inadequate health literacy (Chew et al., 2004). Thus, a logical choice was to retain the question with the highest discriminant validity.

As in the previous two studies included in this thesis (Chapters 5 and 6), all the predictor variables were scaled by two standard deviations and centred to have a mean of zero (Gelman, 2008). As mentioned in Chapter 5 (section 5.3.2.i), this allowed simple comparisons among predictors. In the case of categorical variables, specifically self-reported English proficiency and education level, prior to scaling, the responses were first converted to numeric variables. As explained in Chapter 6 (section 6.2.2.ii), standardising categorical predictors meant that the models assumed that there is a linear relationship between the effects of variation in education and English language proficiency levels on comprehension. Like in Study 2 (discussed in Chapter 6), native English and advanced English language speakers were assumed to have the highest level of English language proficiency, followed by intermediate and beginner English language speakers respectively.

In summary, the study had a repeated measures design, with each participant answering six comprehension questions, two multiple-choice and four open-ended, about each of the four health-related texts. All participants were exposed to all assessments of individual differences, and they read the same texts. As I hypothesized that the effects of participant attributes could modulate the impact of the effects of text features, in the primary analysis I included interaction terms corresponding to the interactions between the effects of text features and the effects of individual differences to answer RQ7.3. Overall, in the primary analysis, I created a set of 21 different Bayesian mixed-effects logistic models (discussed in detail in the results section). In these models the predictors were kept constant, and the outcome measure constituted the probability of correctly answering reading comprehension questions. Next, I discuss my analyses.

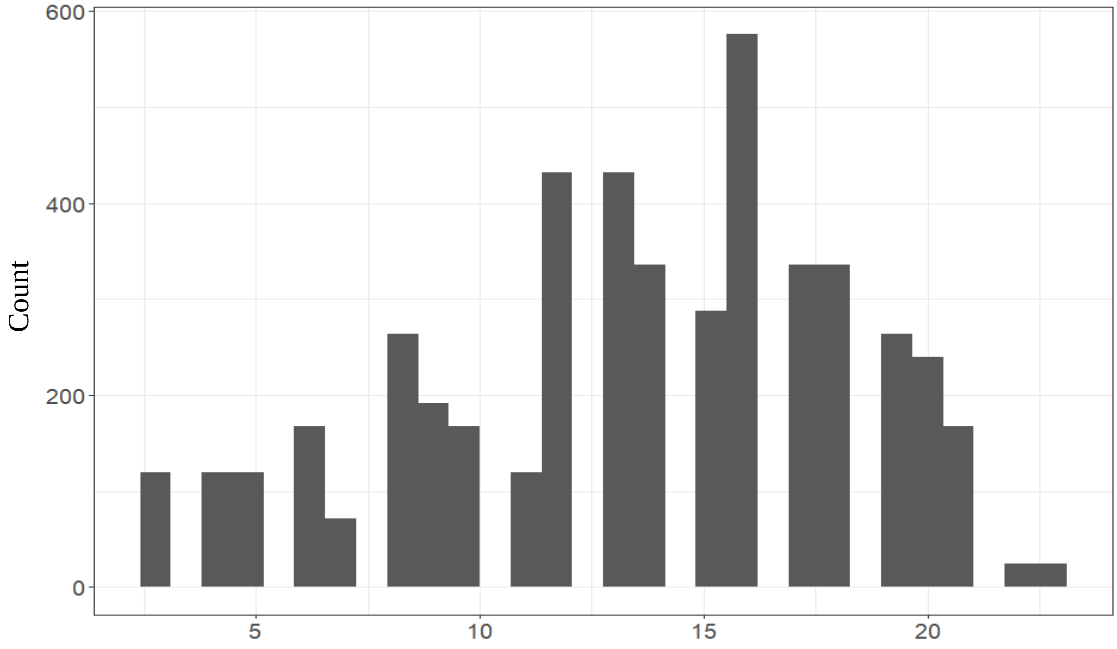
7.3. Results

In this section, first, I discuss the distribution of scores as distributions determine the model classes used in the analyses. Second, I briefly describe the correlations between the text- and person-level predictors of comprehension and the scores on the reading comprehension test. Third, I interpret the estimates calculated by the Bayesian models to answer the research questions posed in this thesis (Chapter 4, section 4.3) and in this chapter (section 7.1.2). Last, I discuss the Bayesian models building process, sensitivity and exploratory analyses.

7.3.1. Descriptive Statistics

In the following description, I briefly discuss participants' performance by self-reported English language proficiency, and education level, on the reading comprehension of health-related information test (Table 7.8; Figures 7.2, 7.3, and 7.4). It is important to mention, that scores on the comprehension test do not constitute a continuous, normally distributed, variable (see section 7.3.2). Nonetheless, in this section they have been treated as interval data. Figure 7.2 shows that the reading comprehension of health-related information test, appeared to be relatively good at discriminating between participants of different ability. This is because the distribution of scores looks approximately normal, without ceiling or floor effects (Chapter 6, section 6.3.1.i), and there is a wide range of scores (see also Table 7.8). Specifically, in the sample of all participants, the scores varied from 3 to 23, where higher scores corresponded to better understanding of health-related texts.

Figure 7.2. Distribution of reading comprehension scores.



One practical implication of the distribution shown in Figure 7.2 is that not all participants understood health-related texts well, even if they might have thought that their understanding of such texts was relatively high (Chapter 6). Especially problematic is the finding that, on average, beginner level ESL participants who completed further education only, understood only around 25% of what they read (Table 7.8). In general, understanding of low-proficiency ESL readers was lower compared to advanced-proficiency and native English readers, regardless of the self-reported education level (Figure 7.3). Amongst the L1 English speakers, on average, those with the lowest educational background had the lowest performance (Figure 7.4; Table 7.8). L1 English readers with secondary school education answered only around 58% of the questions correctly, meaning that they probably did not understand a significant proportion of the key information in the text. Overall, the better educated and the more proficient participants were likely to understand more.

Table 7.8. Comprehension test scores per English language proficiency and educational background.

		Number	Mean	*SD*	Minimum	Maximum
English proficiency	Educational background					
Native English	Higher education	10	17.1	3.18	11	21
	University student	10	18	2.37	13	21
	Further education	46	15.83	2.88	9	21
	Secondary school	34	13.74	3.29	8	20
Advanced	Higher education	5	18.20	2.14	16	22
	University student	10	18.40	2.58	13	23
	Further education	4	15.25	1.79	13	18
	Secondary school	-	-	-	-	-
Intermediate	Higher education	13	14.85	4.39	4	20
	University student	-	-	-	-	-
	Further education	32	12.31	2.9	5	16
	Secondary school	-	-	-	-	-
Beginner	Higher education	2	8.50	.51	8	9
	University student	-	-	-	-	-
	Further education	34	6.56	2.75	3	12
	Secondary school	-	-	-	-	-

Note. The maximum score achievable is 24.

Figure 7.3. Distributions of reading comprehension scores per self-reported English proficiency.

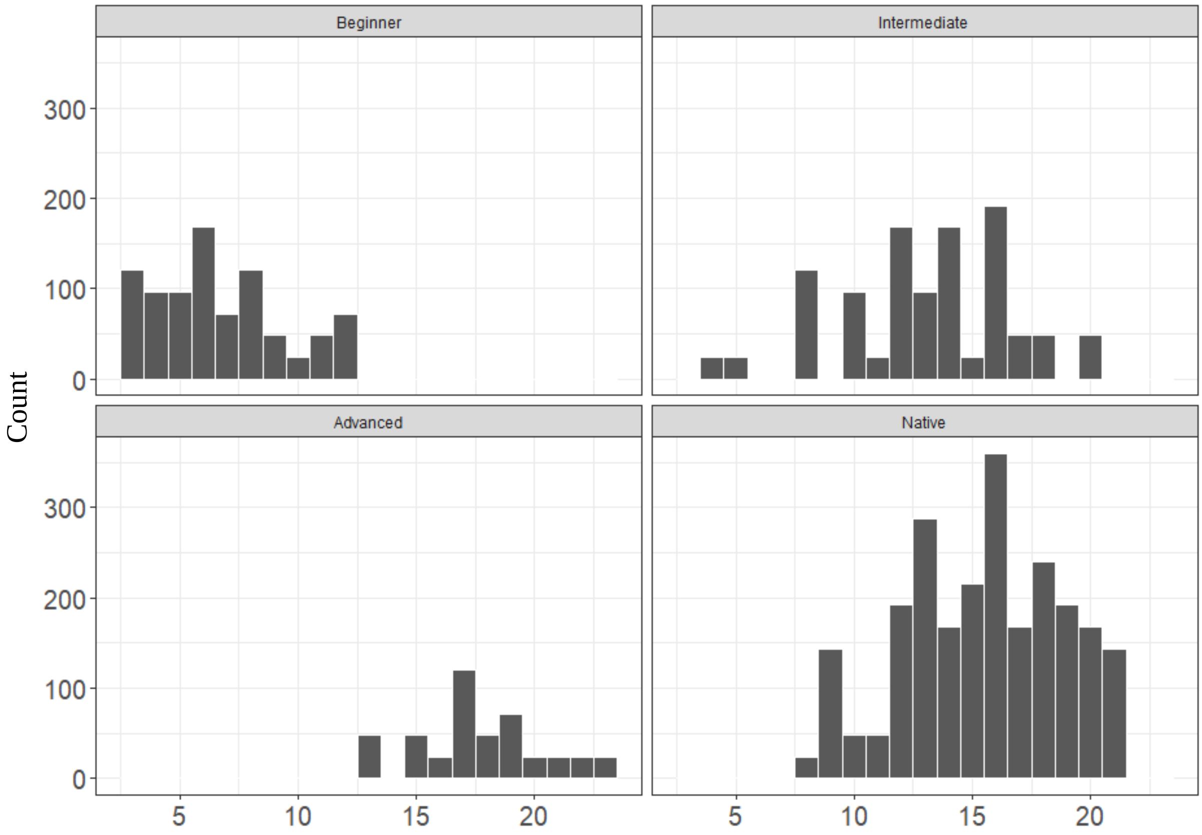
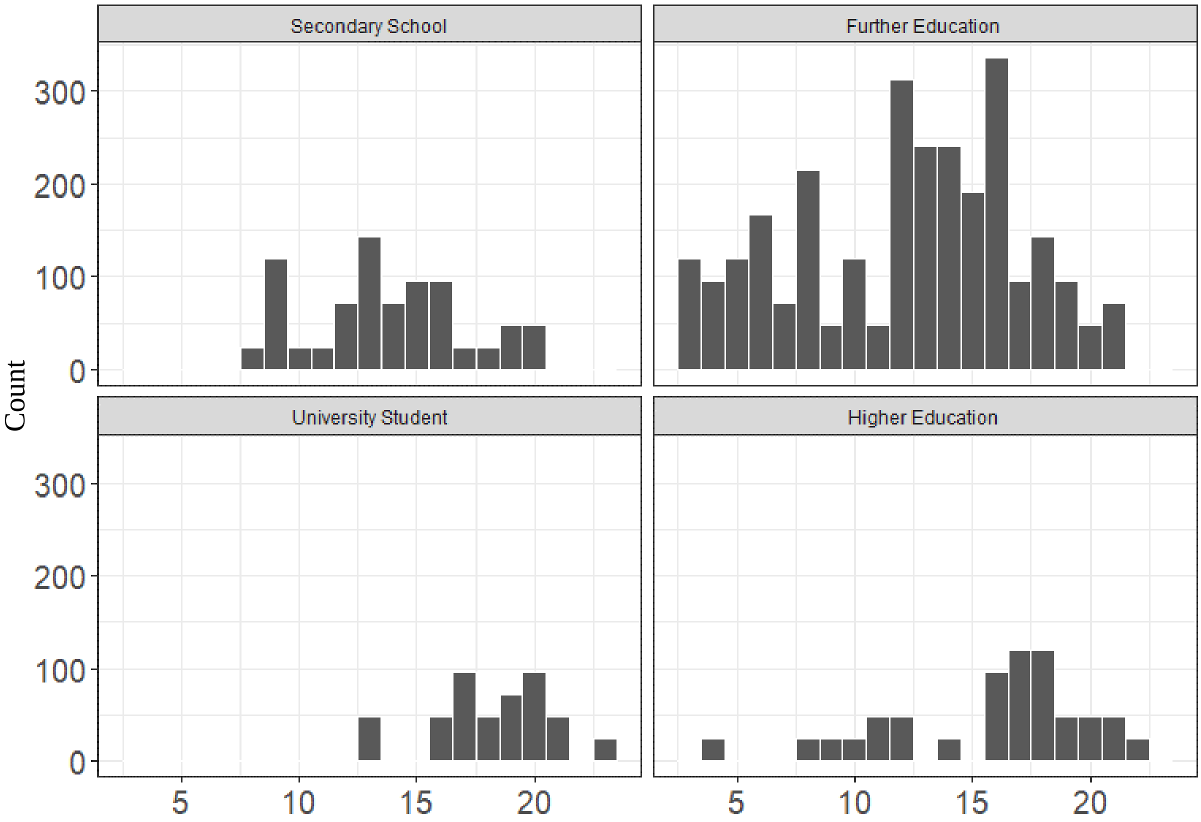


Figure 7.4. Distributions of reading comprehension scores per self-reported education level.



7.3.1.i. Probability Distribution of Outcome Variable

As mentioned in the previous chapters (Chapter 5, section 5.3.1.i; Chapter 6, section 6.3.1.i), it is important to discuss the distribution of the outcome variable, as the function and the shape of the distribution determines the model class that should be used to fit the data. Although, Table 7.8 (section 7.3.1) shows the summed scores for all questions for the four health-related texts that were presented to participants, aggregating the responses to comprehension questions and treating them as interval data is not appropriate. This is because, in this study, each response was scored at the comprehension question level, whereby an answer to a question could have been scored only as either correct or incorrect. This constitutes binary data that follows a Bernoulli distribution, rather than interval level data which may follow a normal distribution (Bolker et al., 2009). Therefore, the modeling approach adopted in this study should reflect the probability of getting a comprehension question right (Bolker et al., 2009). Logistic models with the Bernoulli distribution for the outcome variable allow to model the probability of getting a question right (Bolker et al., 2009). Thus, logistic models were used to model question-level accuracy data in this study. However, to examine correlations between the predictor and outcome variables, I aggregated the binary data from comprehension test for each participant. I briefly discuss some of these correlations next.

7.3.1.ii. Correlations

Table 7.7 (section 7.2.3.iv) shows that all the individual difference variables used in this study were correlated with reading comprehension performance. However, some of the most promising predictors of health comprehension performance may be English proficiency, vocabulary, health literacy, phonological awareness, and perceived comprehension. High scores on each one of these measures were relatively strongly correlated with high reading comprehension performance, suggesting that these individual differences may be important to comprehension. Critically, age was correlated with lower understanding of health-related texts, and lower WM, but with higher vocabulary knowledge, and higher HLVA and SAHL-*E* (Lee et al., 2010) scores. Since, the HLVA and SAHL-*E* health literacy tests seem to target health knowledge, the correlations associated with age suggest that vocabulary and health knowledge may accumulate with ageing, but that the processing of information, including comprehension, may deteriorate with ageing.

In addition to the correlations between individual difference variables, it is important to discuss the correlations between text feature variables (Table 7.6, section 7.2.3.iv). The correlations between the text features were particularly problematic in terms of multicollinearity (section 7.2.3.iv; Chapter 5, section 5.3.1.ii). For example, FRE and the RDL2 correlation was larger than the one reported in Study 2 (Chapter 6, section 6.3.1.ii), and much larger than the one reported in Study 1 (Chapter 5, section 5.3.1.ii). Similarly, the correlations between other text features were relatively large (Table 7.6, section 7.2.3.iv), and many exceeded the frequently accepted threshold for diagnosing collinearity (Dormann et al., 2013) (see also Chapter 5, section 5.3.1.ii).

As mentioned in Chapter 6 (section 6.3.1.ii), it is likely that the text feature correlations were spuriously high due to the relatively low number of health-related texts used in this study which might have led to unstable and inaccurate correlation estimates (Schönbrodt & Perugini, 2013). However, given that the effects of text features were paramount to this investigation I decided to keep most text feature predictors, including those that were highly correlated, in the model (see section 7.2.3.iv for those that were excluded). Although retaining highly correlated text-feature predictors might have led to unstable estimates of text-feature effects, I carried out additional analyses to mitigate for this possibility. I describe these additional analyses later (section 7.3.2.iv).

Critically, none of the text features were correlated with the reading comprehension scores (Table 7.6, section 7.2.3.iv). It may be the case that variation in the text features had a marginal effect on comprehension of health-related texts, given the sample of texts used in this study. However, it is important to note that correlations are indicative of potential trends, but do not generate predictions. Thus, any trends discussed in this section are speculative. To further understand the relations between the effects of variation in individual differences and texts features on comprehension, it is important to construct models that can make predictions and treat comprehension questions as accuracy data. I discuss these models and interpret the estimates of these models next.

7.3.2. Bayesian Models

Using Bayesian mixed-effects logistic models, I examined the text-level and reader-level factors that predicted the changes in comprehension of health-related texts. I analysed 4800 observations — 24 comprehension questions per person — using the brm function of the brms package (Bürkner, 2017; 2018) in R (R Core Team, 2019). As I hypothesized that the effects of participant attributes could modulate the impact of the effects of variation in text features, I included interaction terms corresponding to the interactions between the effects of variability in text features and the effects of individual differences. Next, I discuss and justify my prior distribution choices for the parameters of my models.

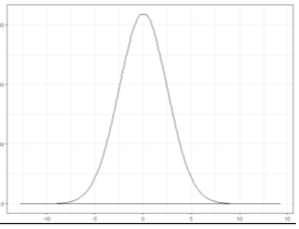
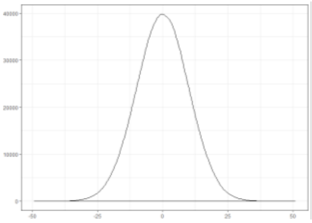
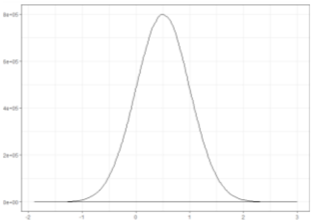
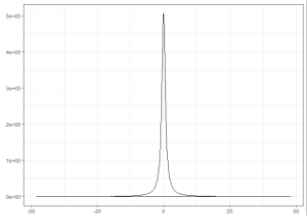
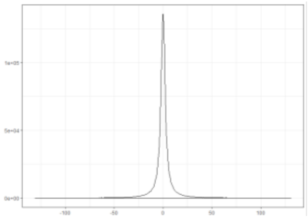
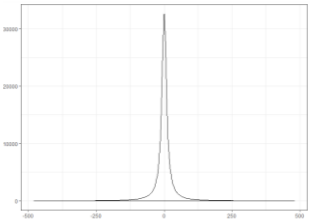
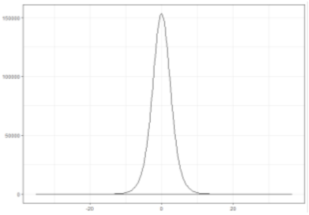
7.3.2.i. Prior Distributions

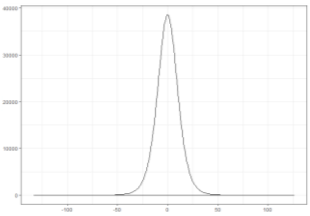
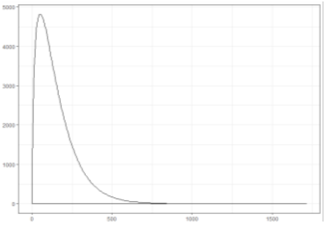
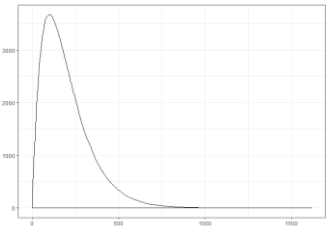
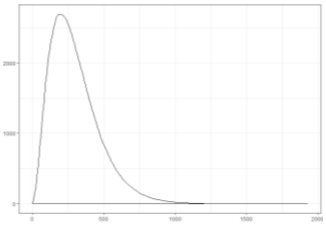
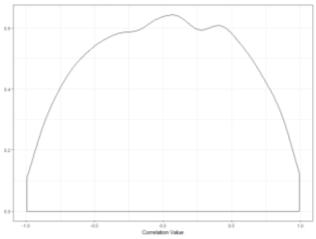
As in the previous studies (Chapter 5, section 5.3.2.i; Chapter 6, section 6.3.2.i), in this study I decided to use weakly-informative regularising priors. Importantly, like Study 2 (Chapter 6), this study also had a multilevel structure. Specifically, every participant provided 24 observations, as each participant had to read four health-related texts and had to answer six comprehension questions per text. Consequently, it was assumed that participants could vary at random in the accuracy of their answers on the comprehension questions. In addition, it was also assumed that comprehension questions could vary in difficulty within each health-related text, but also between health-related texts.

To account for this random variation in the accuracy of answering comprehension questions, all the models were fitted with a maximal random effects structure justified by the data (Bates et al., 2018; Barr et al., 2013; Matuschek et al., 2017). This means that the by-individual and by-comprehension-question random intercepts were fitted with terms corresponding to random variation in the slopes of all individual differences' effects like age. This allowed me to accurately estimate the effects of predictors while accounting for random variation in the probability of getting a comprehension question right associated with the differences between participants and comprehension question difficulty posed within and between the four health-related texts.

Due to the multilevel structure of the data, prior distributions had to be assigned not only for the effects of the predictor variables, but also for the random effects (Chapter 6, section 6.3.2.i). Critically, the prior distributions considered for parameters of the models used in this study were different to those used in Chapter 6 (Section 6.3.2.i), as Bayesian mixed-effects logistic models have some general regularizing priors' guidelines (e.g., Gelman, Jakulin, Pittau, & Su, 2008; Ghosh, Li, & Mitra, 2018). Consequently, I adopted some of these guidelines to build a set of candidate models (see Table 7.9 for justification and visualisation of the considered prior distributions in different models). Next, I discuss what I consider to be the optimal model that I arrived at using the prior distributions described in Table 7.9.

Table 7.9. Prior distributions of predictors and random effects by their variant, purpose, justification, and visual representation.

Prior	Parameters	Prior Parameter Value	Purpose	Justification	Visualisation*
Normal	Predictors	(0, 2.5)	To regularise the estimates of logistic regression predictors	This prior is relatively robust and is known to produce relatively plausible posterior distributions (Ghosh et al., 2018)	
Normal	Intercept	(0, 10)	To regularise the intercept	Same as above, with standard deviation of 10 being commonly used for the intercept as it is less informative (Ghosh et al., 2018)	
Normal	Intercept	(.5, .5)	To regularise the intercept	This prior is more informative but is thought to be more appropriate than Normal(0, 10). Specifically, if the probability of an event, such as answering comprehension question correctly, is anywhere between 0 and 1, a tighter prior with a mean of .5 is thought to produce less biased estimates than a wider prior with a mean of 0 (Gelman, 2018)	
Cauchy	Predictors and random effects	(0, .75)	To regularise the estimates of logistic regression parameters	There is evidence to suggest that this prior regularises better than Normal (Gelman et al., 2008), but it is relatively highly informative, meaning that it is likely to affect the estimates more than weakly-informative normal prior	
Cauchy	Predictors and random effects	(0, 2.5)	To regularise the estimates of logistic regression parameters	A less informative version of the above. As this prior is slightly less informative than the above, it is less likely to have unreasonable influence on the estimates (Gelman et al., 2008)	
Cauchy	Intercept	(0, 10)	To regularise the intercept	A weakly-informative prior that has been shown to regularise the intercept relatively well (Gelman et al., 2008)	
Student-*t*	Predictors	(7, 0, 2.5)	To regularise the estimates of logistic regression predictors	There is evidence to indicate that this prior is more robust than a Normal prior, and is known to produce relatively plausible posterior distributions (Ghosh et al., 2018)	

Student-*t*	Intercept	(7, 0, 10)	To regularise the intercept	Same as above, but slightly less informative (Ghosh et al., 2018)	
Gamma	Random effects	(1.5, .01)	To regularise the estimates of random effects	Gamma priors assume that random effects with variances of zero are implausible (Chapter 6, section 6.3.2.i; Chung et al., 2013; Chung, Gelman, Rabe-Hesketh, Liu, & Dorie, 2015). All the three variants are relatively robust. The (1.5, 01) specification is thought to make more precise, meaning more informative, predictions, whereas the (3, .01) is thought to be less precise but allow for a greater plausible range in the variance of the random effects (Chung et al., 2013; 2015). The (2, .01) variant offers a middle ground between precision and flexibility	
Gamma	Random effects	(2, .01)			
Gamma	Random effects	(3, .01)			
Lewandowski-Kurowicka-Joe (LKJ; Lewandowski et al., 2009)	Correlations	1	To regularise the expected amount of correlation among the parameters	LKJ prior with a shape parameter of 1, permits relatively high correlations between parameters, since some of the correlations between the text features were very high, but makes extreme correlations relatively implausible	

Note. *These distributions are based on visualisations of simulated data. The Cauchy prior visualisations were widened to make the plots easier to look at.

7.3.2.ii. Optimal Model of Reading Comprehension of Health-Related Information

To answer the research questions posed at the beginning of this chapter (section 7.1.2), I fitted a series of Bayesian logistic mixed-effects models. The model building process and sensitivity analyses are briefly described later (section 7.3.2.iii). In this section, I present a summary of the optimal model, showing the plausible effects of the predictors of tested comprehension of health-related texts (Table 7.10; Figure 7.5). It is important to mention that Table 7.10 in this section shows only the effects of variation in reader characteristics and text features and not their interactions with each other. This is because there was no evidence for modulation of the effects of text features by individual differences, and the number of interactions estimated was vast. The estimates of the interaction effects are in the appendix (Table 7.11 in Appendix E).

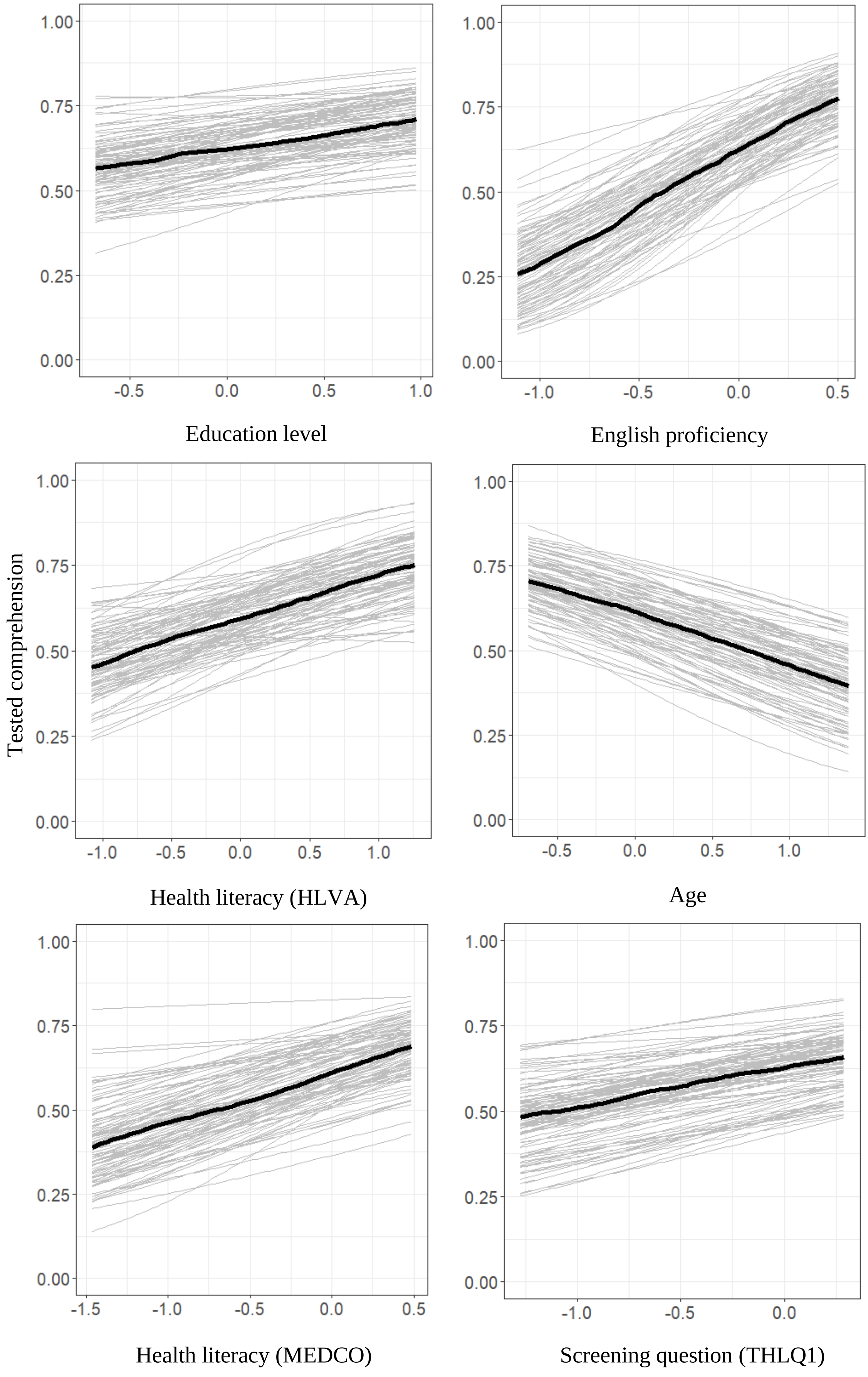
Table 7.10. Summary of the optimal model (Model 19.1).

Coefficients	Estimate	Est.Error	L-95%	U-95%	Estimate OR	L-95% OR	U-95% OR	Probable (sign)
Intercept	.45	.34	-.23	1.13	1.57	.79	3.10	
Education level	.39	.19	.02	.75	1.48	1.02	2.12	*(+)
English proficiency	1.42	.39	.65	2.17	4.14	1.92	8.76	*(+)
HLVA	.55	.24	.10	1.03	1.73	1.11	2.80	*(+)
Age	-.64	.20	-1.02	-.25	.53	.36	.78	*(-)
WM	.08	.19	-.29	.45	1.08	.75	1.57	
Phonology	.00	.23	-.46	.46	1.00	.63	1.58	
MEDCO	.62	.22	.21	1.06	1.86	1.23	2.89	*(+)
Metacomprehension	.29	.21	-.12	.69	1.34	.89	1.99	
THLQ1	.46	.19	.10	.83	1.58	1.11	2.29	*(+)
RDL2	-.06	2.32	-4.58	4.48	.94	.01	88.23	
FRE	-.15	2.26	-4.55	4.28	.86	.01	72.24	
Temporal connectives	.15	2.23	-4.21	4.46	1.16	.01	86.49	
All connectives	-.04	2.34	-4.63	4.49	.96	.01	89.12	
Stem overlap	.13	2.23	-4.22	4.50	1.14	.01	90.02	
Hypernymy noun	-.02	2.30	-4.53	4.45	.98	.01	85.63	
Hypernymy verb	-.12	2.24	-4.48	4.29	.89	.01	72.97	
Deep cohesion	-.10	2.25	-4.55	4.29	.90	.01	72.97	
Referential cohesion	.01	2.33	-4.56	4.55	1.01	.01	94.63	
Causal connectives	-.04	2.30	-4.53	4.50	.96	.01	90.02	
CELEX frequency	-.07	2.26	-4.49	4.31	.93	.01	74.44	
Sentence length	.14	2.27	-4.36	4.60	1.15	.01	99.48	
Passive voice	.16	2.29	-4.31	4.63	1.17	.01	102.51	
Syntax similarity	.06	2.30	-4.44	4.57	1.06	.01	96.54	
Causal cohesion	-.08	2.31	-4.63	4.51	.92	.01	90.92	
Logical connectives	.02	2.29	-4.50	4.50	1.02	.01	90.02	
Gerunds	.04	2.34	-4.53	4.60	1.04	.01	99.48	
BNC frequency	.00	2.31	-4.55	4.58	1.00	.01	97.51	

Note 1. OR refers to Odds Ratio. *Note 2.* English proficiency variable constitutes self-assessed English language proficiency, English language vocabulary, and a vocabulary-based assessment of health literacy (see section 7.2.3.iv). *Note 3.* Metacomprehension variable constitutes of self-rated perceived understanding of, and perceived effort required to understand, health-related texts (see section 7.2.3.iv). *Note 4.* HLVA is health literacy vocabulary assessment; WM is working memory; MEDCO is a medicine-label-based health literacy assessment; THLQ1 is a screening question used to rapidly assess health literacy; RDL2 is Coh-Metrix L2 Readability Index (Crossley et al., 2008); and FRE is Flesch Reading Ease (Flesch, 1948).

Table 7.10 shows the coefficients of the optimal model, including six plausible predictors of tested comprehension of health-related texts, whereas Table 7.12 (Appendix E) shows the random effects structure of the optimal model. To aid the interpretation of the coefficients in Table 7.10, the log-odds estimates are supplemented with OR (Odds Ratio) estimates which were calculated by exponentiating the log-odds coefficients.

Figure 7.5. Probable predictors of tested comprehension of health-related texts.



The optimal model explained 45% of the variance associated with reading comprehension of health-related texts, for new data (LOO-R^2 = .45). The random effects accounted for most of the variance (34%), indicating that a lot of variation in individuals' comprehension accuracy was down to random differences between participants, and differences in difficulty between comprehension questions within each text and between the different texts. The rest of the variance in reading comprehension accuracy (11%) was accounted for by the predictor variables. However, nearly all the 11% was accounted for by reader characteristics variables, as the effects of text features and the effects of interactions between reader characteristics and text features, explained less than 1% of the variance in new data. I discuss the plausible effects of predictors of comprehension next.

First, with each increase in education level individuals were on average 1.48 (95% OR CIs [1.02, 2.12]) times more likely to provide a correct answer to comprehension questions (Figure 7.5). Thus, the more educated individuals were more likely to understand health-related texts than the less educated individuals. Second, English proficiency level, as measured using a combination of self-reported English language proficiency, health vocabulary (SAHL-*E*; Lee et al., 2010), and general vocabulary (Shipley-2; Shipley et al., 2009) was found to be the most plausible predictor of reading comprehension of health-related texts (Figure 7.5). Specifically, with each unit increase in English proficiency, readers were on average 4.14 (95% OR CIs [1.92, 8.76]) times more likely to correctly answer comprehension questions. Consequently, the more proficient English language readers were more likely to understand health-related texts than the less proficient English language readers.

In addition, all measures of health literacy, specifically the HLVA (section 7.2.3.i), MEDCO medicine label (Bostock & Steptoe, 2012), and the screening question (THLQ1; Chew et al., 2004), were found to be plausible predictors of comprehension of health-related texts. On average, individuals were 1.73 (95% OR CIs [1.11, 2.80]), 1.86 (95% OR CIs [1.23, 2.89]), and 1.58 (95% OR CIs [1.11, 2.29]), times more likely to answer comprehension questions correctly with each unit increase in health literacy as assessed using the HLVA, MEDCO, and THLQ1, respectively (Figure 7.5). Overall, readers with higher health literacy levels were more likely to understand health-related texts than those with lower health literacy levels.

Last, age was the only plausible predictor which was found to negatively predict comprehension of health-related texts (Figure 7.5). Specifically, with each unit increase in age, readers were on average 1.89 (1/.53) (Adjusted 95% OR CIs [2.78, 1.28]) times less likely to correctly answer comprehension questions. Consequently, the older individuals were less likely to understand health-related texts than the younger readers. Critically, no other reader characteristics, and no text features, were found to be plausible predictors of comprehension of health-related texts in the presence, and accounting for the influence of, the other predictors. Likewise, as mentioned at the beginning of this section, there was no evidence for the plausible modulation effects of participant attributes on the effects of variation in text features (refer to Table 7.11 in Appendix E). I explore this further later (section 7.3.2.iv), but first I discuss the model building process and the sensitivity analyses.

7.3.2.iii. Sensitivity Analyses

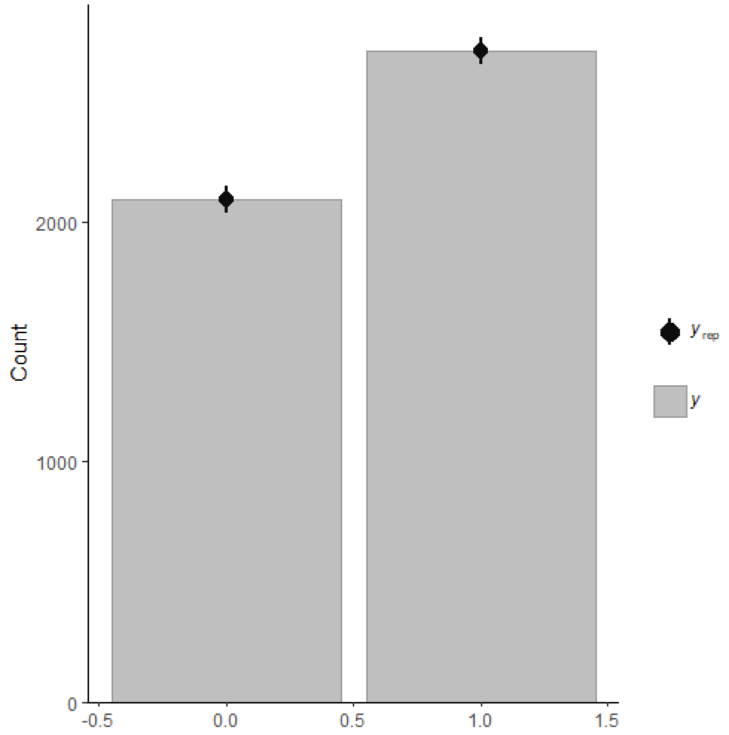
In this section, I briefly describe the model selection process. First, I fitted a set of models differing in their prior distributions of parameters to check for the sensitivity of the estimates to the choice of the prior (Table 7.13 in Appendix E). These checks demonstrated that the models' estimates were relatively robust, as in most of the considered models, the credible estimates were not sensitive to the choice of the prior distribution. The effect of education was sensitive to some sets of priors but given that it remained plausible in the majority of model variants, it is more probable that the effect is there than it is not.

Second, I checked $\hat{R}$ chain convergence criterion associated with the fitted models (Gelman et al., 2013), LOO Information-Criterion (LOOIC), and the number of effective samples generated by the MCMC algorithm for each model parameter (for an explanation see Chapter 5, section 5.3.2.iv). I have used a combination of these criteria to find the model that best fitted the data but did not overfit it (refer to Chapter 4, section 4.5.1 for over-fitting). The reported optimal model performed best using the combination of these criteria. Critically, although some models had lower LOOIC values than the reported model (Table 7.13, Appendix E), they had a relatively small number of samples (see Chapter 5, section 5.3.2.iv for discussion about samples), indicating that their predictions did not generalise very well to the real-world. In turn, the reported optimal model satisfied the convergence criterion, had a relatively low LOOIC, and had a relatively large number of samples, indicating that its predictions did generalise to the real-world relatively well (also evident in Figure 7.6, discussed below).

Next, I checked for the presence of local convergence in the reported model (for more details on MCMC and local convergence see Chapter 5, section 5.3.2.iv). After doubling the number of iterations used to identify the posterior distribution for the reported model, I found that the model converged and that the effect estimates did not change. This suggests that the estimates of the reported model were relatively insensitive to changes in the number of iterations. Last, I checked the predictive performance of the optimal model. The posterior predictive check (PPC; Figure 7.6) plot demonstrates that the reported model had excellent predictive performance, as the model-implied replicate datasets closely resembled the observed data (Martin & Williams, 2017).

However, theoretically (e.g., Francis et al., 2018; Liu et al., 2009), and practically given the writing guidelines (e.g., NHS England, 2018a; Plain English Campaign, 2018), it is unexpected that the effects of text features were not detected in the primary analyses and in the sensitivity checks. One of the potential reasons for the lack of plausible text-feature effects may be collinearity issues between text-level predictor variables (see sections 7.2.3.iv and 7.3.1.ii). Specifically, the inflation of standard error associated with collinearity (Chapter 5, section 5.3.1.ii), and the potential for relatively unreliable and unstable estimates of text-feature effects due to collinearity (e.g., Dormann et al., 2013). Thus, to investigate whether collinearity might have influenced the estimates of text-feature effects in the optimal model, I ran additional exploratory analyses. I discuss these next.

Figure 7.6. PPCs of the reported optimal model (Model 19.1).



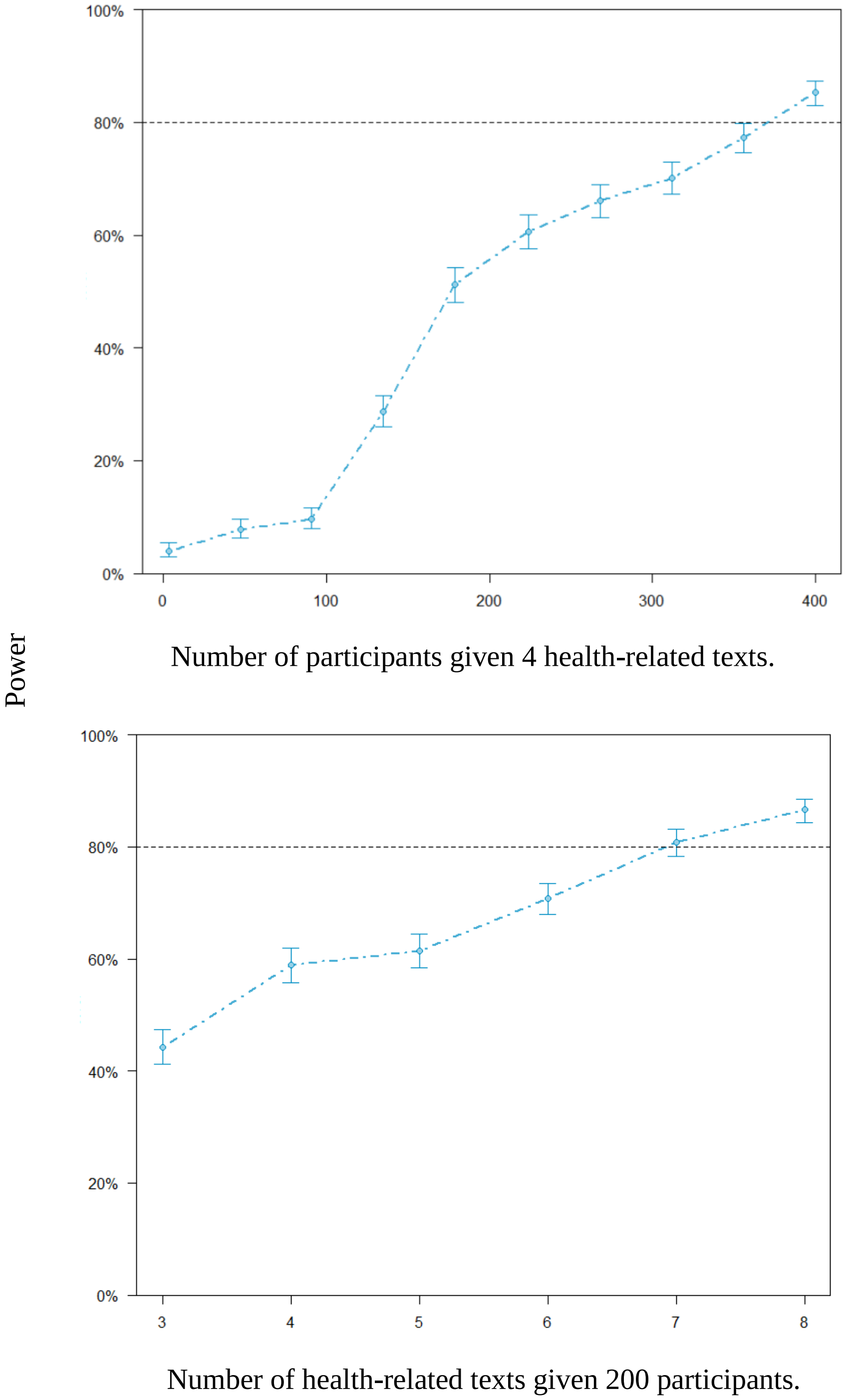
Note. Replicate model-implied datasets are plotted as black CIs labelled y_{rep}, the observed data is plotted as grey bars labelled y.

7.3.2.iv. Exploratory Analyses: Text Features

I ran 36 additional variants of the optimal model. Of those variants, 18 models contained individual text-feature effects only, whereas the rest consisted of individual text-feature effects interacted with the effects of all individual differences' predictors. I found that variation in text features did not predict comprehension of health-related texts in any of the model variants. Furthermore, I found no evidence that the effects of participant attributes could modulate the impact of the effects of variation in text features on comprehension. One observation that could be made is that without sharing variance with other text features that were absent from the model, and without variance inflation associated with collinearity, the estimates of the text feature effects were larger in the absolute sense. Nevertheless, the effects remained improbable, and it was theoretically important to investigate further.

In the Bayesian framework, retrospective power analyses can be performed to assess the probability of sign and magnitude errors, as estimates from underpowered studies can be spurious in terms of the direction and size of the effect (Gelman & Carlin, 2014). Thus, to investigate why the effects of text features remained implausible, I performed a post-study exploratory power analysis for text features by individual differences interactions given the data available using the package simr (Green & MacLeod, 2016). I found that the effects of text features by individual differences interactions achieved 60% power given the sample of texts and participants used in this study. Consequently, to be detected, in this study, the effects of text features on reading comprehension of health-related texts would have to be larger than anticipated. Based on simulations, given the same effect size that is estimated in the optimal model, approximately 360 participants would be required, given 4 texts, to achieve 80% power to detect individual differences by text feature interactions if they are truly there. Alternatively, individuals would have to read 7 health-related texts each, keeping the number of participants constant at 200, to achieve the desired power for the detection of interaction effects (Figure 7.7). Overall, one of the reasons for the lack of plausible text feature effects on comprehension may be lack of power to detect these effects.

Figure 7.7. Simulations-based power calculations for detecting the effects of individual differences by text features interactions.

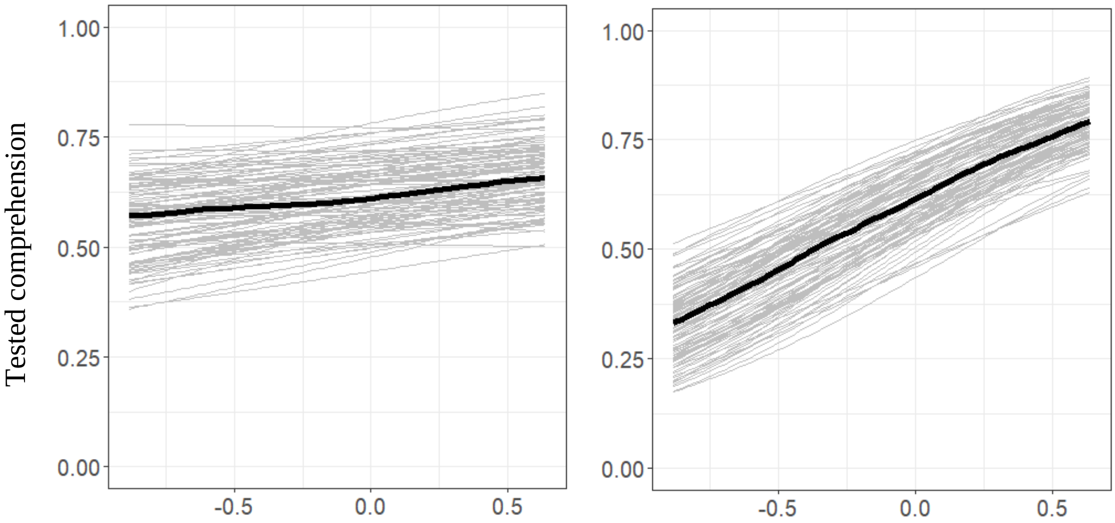


7.3.2.v. Exploratory Analyses: Metacomprehension

In addition to the plausible effects of text features on comprehension, it is important to investigate whether metacomprehension is a plausible predictor of comprehension of health-related texts in the absence of other predictor variables. This is because NHS involves end-users, through for example reader panels, in evaluating comprehensibility of health-related texts (NHS England, 2018a; see also section 4.1 of Chapter 4). Thus, in practice, NHS health information writers rely on evaluations of end users without having information about the reader characteristics of their target group of readers. Consequently, although metacomprehension may of little predictive utility in the presence of individual differences variables such as vocabulary, it may be predictive of comprehension in the absence of other individual differences variables.

Indeed, a variant of the optimal model with random effects, and the predictor of metacomprehension alone, predicted a plausible effect of variation in metacomprehension judgements on tested comprehension. Specifically, with each unit increase in metacomprehension, individuals were predicted to be 3.94 (95% OR CIs [2.32, 6.75]) times more likely to answer comprehension questions correctly. Therefore, those who rated their metacomprehension as higher were more likely to have scored higher on the comprehension of health-related texts test. Figure 7.8 below illustrates the contrast between the predictive utility of metacomprehension depending on the model used for generating the prediction. The left-side plot shows the effects of metacomprehension on tested comprehension using the reported optimal model with all individual differences and text features predictors (section 7.3.2.ii). In turn, the right-side plot demonstrates the same prediction based on the model described in this section, thereby a model without any text-feature and other individual differences predictors. The potential reasons for this discrepancy in predictions are discussed next.

Figure 7.8. Probable effects of metacomprehension on tested comprehension.



Optimal model

Metacomprehension only

7.4. Discussion

In this study, I aimed to investigate the effects of variation in reader characteristics and text features on tested comprehension of health-related texts. My analyses showed that comprehension performance was predicted by education level, English language proficiency, health literacy, and age. In contrast to variation in reader characteristics, there was no evidence for the potential effects of variation in text features on comprehension of health-related texts, in the presence and absence of other covariates. In addition, there was no evidence to support the hypothesis that the effects of participant attributes could modulate the effects of variation in text features. I discuss these findings briefly in the following, and in more detail in the next chapter (Chapter 8). To avoid repetition, theoretical and practical implications of the findings of this study are considered in Chapter 8 (section 8.2).

I asked, “How do reader attributes predict comprehension of written health-related information?” (RQ7.1). I found that several reader attributes were plausible predictors of comprehension of written health-related information, partially supporting $H_{7.1}$. Critically, high English language proficiency, as measured using a combined score of Shipley-2 vocabulary test (Shipley et al., 2009), self-reported English language proficiency, and one of the measures of health literacy (SAHL-*E*; Lee et al., 2010) was predicted to be associated with high reading comprehension of health-related texts.

The high correlations (section 7.2.3.iv) between the vocabulary scores with SAHL-*E* and self-rated proficiency suggest that these variables to some extent measured the same underlying construct. Specifically, here I argue that English language proficiency variable was a proxy for general and health English language vocabulary knowledge. General knowledge, due to involvement of Shipley-2 vocabulary test (Shipley et al., 2009), whereas health knowledge due to the inclusion of SAHL-*E* (Lee et al., 2010). Although the inclusion of the self-rated proficiency variable may be questioned, the correlations suggest that it could be a proxy for health vocabulary knowledge in the specific context of this study (section 7.2.3.iv). Complementarily, self-rated proficiency could also be a proxy of language exposure, which is in turn a proxy of general vocabulary knowledge (Brysbaert et al., 2016). Thus, one of the reasons for the relatively strong effects of English language proficiency variable on comprehension, could be that it encompassed predictors that are thought to, and were found to, be central to comprehension of texts written in English (e.g., Brysbaert et al., 2016; Chin et al., 2018; Freed et al., 2017; Liu et al., 2009; Perfetti, 2010; Todd & Hoffman-Goetz, 2011). I discuss the effects of these theoretically important predictors next.

Health vocabulary knowledge is theorised to be important to comprehension as it is a part of functional health literacy (Chin et al., 2011; Chapter 3, section 3.1.1). In turn, functional health literacy is thought to play a part in successful comprehension of health-related texts (e.g., Chin et al., 2015; 2018; Liu et al., 2009). Indeed, high scores on all measures of health literacy included in this study were found to predict high understanding of health-related texts. One of the reasons for the effects of health literacy on comprehension, may be that health literacy constitutes relevant background knowledge in the context of health. As mentioned in Chapter 3 (section 3.1.1.), relevant background knowledge is thought to be critical to comprehension. This is because relevant background knowledge is theorised to moderate the activation of reader-initiated processes, such as inferences, that are thought to be necessary for meaning integration and construction of a coherent situation model (McNamara & Kintsch, 1996; van den Broek & Helder, 2017).

However, it is important to mention that functional health literacy is also thought to encompass processing of health information (Chin et al., 2011). As mentioned in Chapter 3 (section 3.1.1), this means that not all health literacy measures are the same, as they do not measure the same aspect of health literacy (Kobayashi et al., 2015; 2016). In this study, this is visible upon the examination of the correlations between the different health literacy measures and age (section 7.2.3.iv). Specifically, high scores on the more vocabulary-focused measures of health literacy that required the production of definitions or selection of similar medical words, such as the SAHL-*E* (Lee et al., 2010) and the HLVA (section 7.2.3.i), were positively correlated with high age. In contrast, high scores on the health literacy measure which required more processing, specifically the MEDCO medicine label (Bostock & Steptoe, 2012), were negatively correlated with high age. One explanation for this may be that health knowledge may remain unchanged or accumulate with age (Chin et al., 2009; Gazmararian, Williams, Peel, & Baker, 2003), whereas processing may decline (Kobayashi et al., 2015; 2016) (Chapter 3, section 3.1.1 and 3.1.2). Thus, some aspects of health literacy, such as health knowledge, may be positively affected by ageing, whereas others, such as processing, may be negatively affected by ageing.

Critically, in this study older individuals were predicted to understand less than younger individuals (implications of this are discussed in Chapter 8). One possible explanation for the effects of age may be that the accumulation of health knowledge associated with getting older (e.g., Chin et al., 2009; Gazmararian et al., 2003) is insufficient to offset the negative effects of ageing on processing required to comprehend health-related texts (Kobayashi et al., 2015; 2016). Indeed, as mentioned in Chapter 3 (section 3.3), some research evidence indicates that older individuals are less likely to understand health-related texts compared to younger individuals (Liu et al., 2009). Thus, it may be the case that ageing has a detrimental effect not only on the speed of processing measures (Chapter 3, section 3.1.2), but also on the comprehension of health-related texts (Chin et al., 2011; Kobayashi et al., 2015; 2016).

In addition to the effects of health literacy, and age, a reoccurring theme throughout this thesis is the importance of vocabulary knowledge in successful comprehension (e.g., section 7.1). Vocabulary knowledge is thought to be important to comprehension, as it is theorised to be critical to forming propositions, textbase, and a logical situation model of the text read (Kintsch & Rawson, 2007). In other words, without knowing the meanings of the words read, the reader is unlikely to build a complete understanding of the text read (Kintsch & Rawson, 2007). Consequently, improving individuals' English vocabulary knowledge is likely to have a positive effect on comprehension of health-related texts. One of the ways in which vocabulary knowledge can be developed, is through exposure to language through, for example, education (e.g., Brysbaert et al., 2016; LARRC, 2015).

Indeed, educational background was a predictor of comprehension of health-related texts in this study. Specifically, higher education was associated with higher probability of understanding health-related texts. Critically, the effects of education remained plausible even in the presence of the effects of vocabulary. One possible explanation for this may be that in addition to boosting individuals vocabulary knowledge, higher education may predict the use of metacomprehension strategies, such as selective rereading, which may be beneficial to comprehension (e.g., Hong-Nam & Page, 2014; Kern, 1994; van den Broek & Helder, 2017; Zabrucky et al., 2012).

Critically, the evidence presented in this study indicates that metacomprehension judgements could be used as potential performance predictors of comprehension of health-related texts, but only if there are no better available alternatives to measuring comprehension. One of the reasons for this may be that individuals' metacomprehension accuracy is likely to be relatively low (e.g., Maki, 1998). Consequently, accounting for the effects of all other individual differences variables included in the models built in this study, asking someone how well they understood health-related texts and how much effort they were required to exert to understand these texts does not provide any additional information to what is already given by the other measures. However, in the absence of performance-related predictor variables, asking metacomprehension questions has some predictive utility (see Chapter 8, section 8.2.2, for practical implications).

In addition to the effects of variation in reader characteristics on comprehension, I also investigated the effects of variation in text features, and readability formulae, on comprehension of health-related texts (RQ7.2). Furthermore, I examined how reader attributes could modulate the impact of the effects of text features, and readability formulae, on comprehension (RQ7.3). Overall, I found no evidence for the hypothesised effects of variation in linguistic features on comprehension ($H_{7.2}$), and the hypothesised reader characteristics by linguistic features modulation effects ($H_{7.3}$). This is unexpected given the research evidence for the presence of these interaction effects (e.g., Francis et al., 2018; Kulesz et al., 2016; Liu et al., 2009; McNamara, 2001; McNamara & Kintsch, 1996; McNamara et al., 1996; O'Reilly & McNamara, 2007; Ozuru et al., 2009) (Chapter 1, sections 1.5 and 1.6). One potential explanation for the discrepancy in findings between this and the previous studies is the stringiness of the statistical methods used in this study (see

Chapter 4, section 4.5), compared to other studies (e.g., Liu et al., 2009). Another one is related to the study design, specifically power (discussed in sections 7.4.1 and 7.4.2).

In summary, it is questionable whether variation in text features can predict comprehension of health-related texts (section 7.1), and whether following the guidelines for health-related information writers improves comprehension of health-related texts (e.g., NHS England, 2018a; Plain English Campaign, 2018). Furthermore, the evidence presented in this study suggests that readability-formulae-based scores of health-related texts are unlikely to be good proxies for tested comprehension of health information. Consequently, increasing readability of texts alone may be insufficient to improving understanding of health-related texts (Chin et al., 2018). However, in contrast to relying on readability-formulae-based evaluations of health-related texts, the reliance on evaluations of reader panel members may be justified in the absence of detailed information about the end users (see Chapter 8). Nonetheless, adapting texts to different groups of the population may be more difficult than envisaged by advocates of mixed-effects models of reading (e.g., Francis et al., 2018). There are several possible reasons for this, such as collinearity (Dormann et al., 2013), which also affected the inferences made in this study. I briefly discuss these limitations next.

7.4.1. Limitations

In this investigation, the reported optimal model contained text-feature variables that suffered from multicollinearity. Multicollinearity may be one of the reasons why previous studies considered only a small number of text features in their models (e.g., Kulesz et al., 2016; Francis et al., 2018; Liu et al., 2009). Nevertheless, in this study, efforts were made to account for these issues (section 7.3.2.iv). Unfortunately, due to the low number of texts used in this study, these efforts might have been hindered as retrospective power analyses suggest that this study was underpowered to detect the effects of text features (section 7.3.2.iv). This limited the inference with regards to the effects of text features by individual differences interactions. Critically, the implications of this study being underpowered are serious, as it is likely that a significant proportion of other studies with smaller samples of participants and similar samples of texts are also underpowered.

7.4.2. Implications

To counter the limitations imposed by study design, researchers should use a larger sample of texts to lower the correlations between the effects of the text features (Schönbrodt & Perugini, 2013). Concomitantly, this would also have the effect of increasing the power for detecting the hypothesised effects of text features by individual differences interactions. Retrospective power analyses revealed that with a sample of four health-related texts, approximately 360 participants would be necessary to have enough power to detect the effects of text features by individual differences interactions, 80% of the time, if they were truly there (Figure 7.7, section 7.3.2.iv). Alternatively, keeping the number of participants constant at 200, a sample of 7 health-related texts would be enough to achieve 80% power. Clearly, more research with more power is needed, but another question that must be posed is about the utility of focusing on the effects of text features in improving comprehension. If variation in text features accounts for less than 1% of variance in comprehension (section 7.3.2.ii), it may be the case that resources for improving comprehension could be better spent elsewhere. This possibility, alongside the practical and theoretical implications of the findings, is further discussed in Chapter 8.

Chapter 8: Overall Discussion

The literature reviewed in Chapters 1, 2, and 3 indicates that the effects of individual differences on comprehension of health-related texts are likely to be modulated by the effects of variation in linguistic features (e.g., Francis et al., 2018; Kulesz et al., 2016; Liu et al., 2009). Furthermore, some of the discussed literature suggests that health-related texts could be adapted for different groups of users to optimise understanding (e.g., Francis et al., 2018; Liu et al., 2009). However, in this thesis, no evidence was found for the presence of the hypothesised interaction effects on comprehension of health-related texts. In addition, no evidence was found for the effectiveness of any of the health-related texts writing guidelines adhered to by the NHS in improving comprehensibility of health-related texts (e.g., NHS England, 2015, NHS England, 2018a; 2018b; Marsay, 2017a; 2017b; Plain English Campaign, 2018). Overall, these findings have major implications for reading comprehension theory development as well as for optimising the understanding of health-related texts.

In this chapter, first, I answer the research questions posed in this thesis (Chapter 4, section 4.3). Next, I discuss the theoretical and practical implications of the evidence reported. In summary, my findings suggest that the health-related writing guidelines could be of limited practical significance. I end this chapter, as well as this thesis, with a conclusion that can be reached considering the findings and implications of the studies conducted, and I suggest potential directions for future research.

8.1. Primary Research Questions

8.1.1. RQ1. How do reader attributes predict comprehension of written health-related information?

The evidence reported in Chapter 7 (section 7.3.2.ii) shows that tested comprehension was only predicted by the variation in four individual differences variables: education level, English language proficiency, age, and health literacy. In addition, exploratory analyses (Chapter 7, section 7.3.2.v) indicate that metacomprehension ratings have some utility in predicting comprehension performance in the absence of other individual-level information. As I discussed these findings in Chapter 7 (section 7.4), to avoid repetition I do not consider them again here. Instead, next, I elaborate on the lack of phonological awareness and verbal working memory (WM) effects on comprehension of health-related texts. I study the implications of all the findings later (in section 8.2).

From the perspective of a number of theoretical accounts, verbal WM and phonological awareness are thought to be important to comprehension (e.g., Kintsch & Rawson, 2007; Perfetti, 1992; 1998; 2007; 2010; Perfetti & Stafura, 2014; Tunmer & Chapman, 2012; Zwaan, 2016) (Chapter 1, sections 1.2 to 1.4; Chapter 2, section 2.1.1; Chapter 3, section 3.1.1; Chapter 7, section 7.1). However, the evidence reported in this thesis, as well as in other robust large-scale empirical research (e.g., Freed et al., 2017), indicates that the effects of verbal WM and phonological awareness diminish in the presence of direct predictors of reading comprehension, such as vocabulary knowledge. This suggests that variation in verbal WM and phonological awareness among adult readers may not be a good predictor of comprehension of health-related texts.

There are two main plausible explanations for the lack of phonological awareness and verbal WM effects. The first explanation considers the degree of sensitivity to detect the effects of variation in these variables, given the demands the text imposed on the reader.

Specifically, in the context of phonological awareness, as word reading becomes more fluent and efficient, approaching maximum, the relative proportion of variance in comprehension performance explained by variation in word reading is thought to decrease, whereas linguistic comprehension processes are hypothesised to play a more influential role (Gough & Tunmer, 1986) (Chapter 1, section 1.3). This is because over time, the increasingly diverse and advanced texts, to which developing readers are exposed, make greater demands on higher-level language skills, such as vocabulary knowledge, rather than decoding skills (Garcia & Cain, 2014; Vellutino et al., 2007). Consequently, variation in phonological awareness is likely to play a less important role in predicting text comprehension among adult readers. In addition, variation between individuals is likely to diminish after years of exposure to written text. This is because most adults are fluent readers, meaning that most adults would be likely to have relatively high phonological awareness scores. Thus, due to lack of meaningful variation in phonological awareness among adult readers, detecting the influence of phonological awareness on comprehension is likely to require a much larger sample of participants than the one used in this thesis, and the effect of this influence is likely to be small.

It is also important to mention that, in future research, phonological awareness could be measured using a different task to the Spoonerisms test used in this thesis (Frederickson et al., 1997; Walton & Brooks, 1995) (see Chapter 7, section 7.2.3.i, for a description). In particular, the ability to manipulate sounds within English words may be problematic for English as Second Language (ESL) readers of moderate, and low, English language proficiency (Nenopoulou, 2005). Crucially, the ESL participants involved in the third study included in this thesis (Chapter 7), were first language (L1) Polish speakers. L1 background is another factor which could have made the Spoonerisms task more difficult for Polish participants as, in contrast to English, Polish has a transparent orthography. Transparent orthographies are characterised by a direct relation between graphemes and phonemes. In other words, in transparent orthographies there is a direct relation between the letters and the sounds in the spoken language that the letters represent. This is different to deep orthographies, such as English, that do not have direct letter-to-sound correspondences. The Spoonerisms test requires higher-order phonological awareness skills as well as good knowledge of grapheme-phoneme correspondence in English that may not be as developed in participants of L1 transparent orthographies, such as Polish, compared to L1 English readers (Nenopoulou, 2005). Thus, relatively low proficiency ESL readers coming from a transparent orthographic background in their L1 might not have as developed phonological awareness skills as L1 English readers, resulting in the likely lower performance of these ESL readers compared to L1 English readers on the Spoonerisms task (e.g., Nenopoulou, 2005). Thus, in future investigations, involving the measurement of phonological awareness of ESL readers, administering a phonological awareness task in readers L1 could be considered.

It can be argued that the Spoonerisms test is a measure of reading fluency rather than phonological awareness, but correlations between reading fluency measures and the Spoonerisms test used in this thesis (Frederickson et al., 1997; Walton & Brooks, 1995), amongst dyslexic, non-dyslexic, and ESL adult populations, indicate that these tasks do not measure the same construct (e.g., Nenopoulou, 2005). In addition, amongst L1 English adults, the Spoonerisms test has been found to effectively discriminate between dyslexic and non-dyslexic adults (e.g., Gabay & Holt, 2015; Law, Vandermosten, Ghesquiere, & Wouters, 2014). Consequently, the Spoonerisms test is likely to be an acceptable proxy of phonological awareness amongst L1 English adult readers. As previously mentioned, phoneme manipulation and deletion require relatively high phonological awareness skills.

Similarly, to the effects of phonological awareness on comprehension of health-related texts, the lack of WM effects on comprehension does not indicate that verbal WM is not involved in comprehension. It may simply be the case that, due to the design of the study, the demands on verbal WM were so far reduced that variation between adult readers was relatively unimportant. To replicate the experience of individuals reading health-related texts in the real-world, the participants used in this thesis could refer to each health-related text when answering comprehension questions (Chapter 7, section 7.2.3.i). Consequently, the participants did not have to remember the information present in each text when answering comprehension questions. Critically, this is different to comprehension assessments used in some of the studies which reported effects of WM on reading comprehension of health-related texts. The procedure employed by some of the previous research required participants to answer comprehension questions by recalling the information they read without consulting the text again (e.g., Chin et al., 2018). Thus, the differences in the effects of verbal WM between the findings reported in this thesis, and the evidence reported in other studies (e.g., Chin et al., 2018), could be attributed to the use of different comprehension measures which impose different demands on WM. However, it is also important to mention that L1 background, and English language proficiency of ESL participants, could have impacted on some participants' performance on the WM measure used in this thesis (e.g., Grundy & Timmer, 2017). Therefore, English language proficiency could have confounded the average effect of WM on reading comprehension performance across all participants, but evidence for this is inconclusive (e.g., Calvo, Ibáñez, García, 2016; Lukasik et al., 2018; Yang, 2017).

As mentioned in Chapter 7 (section 7.1), another plausible explanation for the lack of phonological awareness and verbal WM effects could be that these abilities may be hard to distinguish from the effects of vocabulary knowledge, as they are interdependent with vocabulary knowledge (Freed et al., 2017; Perfetti, 2010; Tunmer & Chapman, 2012; Van Dyke et al., 2014) (Chapter 1, sections 1.3 and 1.4; Chapter 2, sections 2.1.1 and 2.1.2). Due to this interdependence, they share variance with each other (Freed et al., 2017; Van Dyke et al., 2014) but vocabulary knowledge is likely to be the most important predictor as the measures of verbal WM and phonological awareness may be, to a certain extent, reliant on the knowledge of word meanings (Chapter 7, section 7.1). Consequently, the utility of phonological awareness and verbal WM in predicting comprehension in the presence of vocabulary knowledge measures is likely to be relatively low. However, this requires further elaboration as the current reading comprehension theories do not fully account for these findings (implications of this are discussed in section 8.2.1.iii). Next, I address the remaining two research questions.

8.1.2. RQ2 and RQ3. How do textual characteristics predict comprehension of written health-related information, and how do the effects of reader attributes and textual characteristics interact in predicting the comprehension of health-related information?

Research evidence suggests that variation in text features may affect different readers differently (e.g., Francis et al., 2018), including in the context of understanding health-related texts (e.g., Liu et al., 2009). However, I found little evidence for the effects of text features on comprehension of health-related texts, and for the hypothesised modulation of the effects of individual differences by text features (Chapter 7, section 7.3). Critically, the effects of text features, and readability formulae, and the effects of potential modulation of reader characteristics by these linguistic features, accounted for less than 1% of the variance in future reading comprehension performance (Chapter 7, section 7.3.2.ii). This is comparable to the relatively small effects associated with variation in text features, and modulations of the effects of individual differences by text features, reported by previous studies (e.g., Davies et al., 2017; Kulesz et al., 2016). Indeed, evidence suggests that the effects of the modulation of the effects of reader characteristics by text features are overshadowed by the effects of individual differences, mainly vocabulary and background knowledge (e.g., Kulesz et al., 2016). Consequently, the effects of text features on comprehension should be considered against the background of large, overarching, effects of variation in individual differences on comprehension (e.g., Davies et al., 2017) (implications of this are discussed in section 8.2.1.iii).

Nonetheless, it is unlikely to be the case that variation in text features does not matter to comprehension. Instead, it is probable that for the sample of texts used in this thesis it is difficult to detect the effects of variation in text features. This is because health-related texts may vary in their text features, but this variation is likely to be relatively small, as all health-related texts are already written with the intention of being easy-to-understand (e.g., Burrow & Forrest, 2015; NHS England, 2018a). In contrast, the variation in text features between texts used in previous experimental studies is likely to be greater, as many of these studies revised difficult-to-read texts to make them more coherent and cohesive (e.g., O'Reilly & McNamara, 2007; Ozuru et al., 2009). Clearly, studying the effects of variation in text features by manipulating theoretically important text features, such as coherence and cohesion (e.g., Kintsch, 1998; Kintsch & Rawson, 2007; Van Dijk & Kintsch, 1983) (see also Chapter 1, section 1.5), is different to examining a sample of written health-related texts. Text manipulations are likely to change the texts to such an extent that the text features differences between the original and revised text versions are likely to be much greater than any detectable differences between health-related texts that are written with the aid of the same writing guidelines (NHS England, 2018a; Plain English Campaign, 2018), and are adapted in accordance with evaluations of reader panel members (e.g., Burrow & Forrest, 2015). Consequently, it is probable that the effects of text features might have been under-estimated in this thesis or over-estimated in some previous studies (e.g., McNamara & Kintsch, 1996; McNamara et al., 1996; O'Reilly & McNamara, 2007; Ozuru et al., 2009).

Overall, it may be the case that manipulations of health-related texts could elicit clearer effects of text features on comprehension. This is theoretically plausible as highly coherent and cohesive texts are likely to require fewer text-based inferences than relatively incoherent and incohesive texts (e.g., Hamilton & Oakhill, 2014), thereby making it easier for the reader to, amongst other things, create propositions and form the microstructure of the text to build a coherent situation model (e.g., Kintsch & Rawson, 2007) (see also Chapter 1, section 1.2). However, detecting the effects of the manipulation of health-related texts on comprehension is likely to be more difficult compared to examining texts that were not produced with explicit writing guidelines and end-user evaluations (e.g., O'Reilly & McNamara, 2007; Ozuru et al., 2009). This is because health-related texts are likely to be on average more coherent and cohesive than texts written without the involvement of end-users and specific writing guidelines (e.g., NHS England, 2018a; Plain English Campaign, 2018). Thus, the potential for improvements in cohesion and coherence may be relatively small compared to biology texts that were manipulated in some previous studies (e.g., McNamara, 2007; Ozuru et al., 2009). Nevertheless, the effectiveness of manipulations of health-related texts in improving comprehension could be examined by future research. Next, I discuss the theoretical implications of the findings reported in this thesis to the study of reading comprehension of health-related texts.

8.2. Implications

8.2.1. Theoretical Implications

The evidence reported in this thesis has three main theoretical implications. Although, these theoretical implications are inter-related and rely on evidence reported in multiple chapters, I discuss these implications in the order of relevance to the studies included in this thesis, and in the chronological order in which these studies were conducted. Consequently, I begin this section with implications of the reported findings with regards to the construct validity of readability estimates, as these were most informed by the first study of this thesis (Chapter 5).

8.2.1.i. Construct Validity of Readability Estimates

The evidence reported in Chapter 5 (section 5.3.2) indicates that the Flesch Reading Ease (FRE; Flesch, 1948) and the Coh-Metrix L2 Readability Index (RDL2; Crossley et al., 2008) readability estimates have different sets of predictors. This is theoretically problematic as both readability formulae claim to be measuring the same construct (Flesch, 1948; Crossley et al., 2008). In addition, it is reasonable to question the construct validity of the readability estimates as text features excluded from readability formulae were found to predict estimates of these formulae (Chapter 3, section 3.2; Chapter 5, section 5.3.2). Furthermore, the existence of a correlation between readability scores was not substantially supported by the data (Chapter 5, section 5.3.1.ii). Thus, as mentioned in Chapter 5 (section 5.4), the reported evidence is consistent with the view that the FRE and RDL2 readability formulae reflect different aspects of linguistic basis of readability.

In addition to concerns about construct validity of the different readability formulae, I found no evidence for direct effects of variation in readability estimates on comprehension (Chapter 7, section 7.3.2). Consequently, the link between readability and comprehension does not appear to be close (cf. Flesch, 1948; Crossley et al., 2008), and it is questionable whether the FRE and RDL2 readability formulae apply to new data that they were not tested on. However, high FRE scores were associated with ratings of the perception that less effort is needed to be exerted by participants to understand health-related texts (Chapter 6, section 6.3.2.iii). This is important as perceived effort shared enough variance with perceived understanding to reason that, to a large extent, they measured the same underlying construct, assumed to be metacomprehension (Chapter 7, section 7.2.3.iv). In turn, evidence from Chapter 7 (section 7.3.2.v) indicates that high metacomprehension, consisting of low perceived effort but high perceived understanding, was predictive of higher comprehension (in analyses in which no other individual differences predictors were included). In other words, the evidence reported in this thesis suggests that the lower the perceived effort, the higher the metacomprehension, and the higher the metacomprehension the higher the comprehension of health-related texts. Thus, it may be the case that variation in readability estimates has a relatively small, indirect, effect on comprehension through its direct effect on metacomprehension.

Critically, current accounts of reading comprehension do not fully explain the potential indirect effects of variation in readability estimates on comprehension (e.g., Kintsch & Rawson, 2007; Perfetti, 2007; 2010; Tunmer & Chapman, 2012). One plausible explanation for the possible indirect effects of FRE variability on comprehension could be that the texts rated as requiring less effort to understand are written in a way that minimises the processing demands placed on the reader. It may be the case that texts containing a high proportion of short words and sentences signal to readers that they do not have to engage in active processing to meet their standards of coherence for the goal of understanding these texts (Crossley et al., 2017; O'Reilly & McNamara, 2007; van den Broek & Helder, 2017). Consequently, readers' low-effort ratings may be reflective of easy-to-understand texts that do not require the readers to engage in resource-demanding active processes required for coherence-building (van den Broek & Helder, 2017). Indeed, texts high in FRE may be easier to process because longer sentences are thought to place greater demands on meaning-to-text integration processes than shorter sentences (e.g., Perfetti, 2007; Perfetti & Stafura, 2014; Yang et al., 2005) (Chapter 1, section 1.6), and longer words are thought to be on average less frequent and more complex than shorter words (e.g., Flesch, 1948; Crossley et al., 2017; McNamara et al., 2013) (Chapter 5, section 5.2.2).

Nonetheless, it is difficult to explain the effects of variation in FRE scores on comprehension since, as mentioned in Chapter 7 (section 7.2.3.iii), there is also research evidence to suggest that decreasing sentence length may reduce text coherence and cohesion (e.g., Crossley et al., 2008; O'Reilly & McNamara, 2007; Ozuru et al., 2009) (see also Chapter 3, section 3.2). Therefore, the evidence reported in this thesis adds to the debate as to whether the effects of word and sentence length are beneficial or detrimental to understanding. In the context of the sample of health-related texts used in this study, it is plausible that the texts with higher proportion of shorter words and shorter sentences were still relatively coherent and cohesive, as these texts were written with the intention of being easy to understand. Thereby, the demands placed on meaning-to-text integration processes required for comprehension might have been lower for such texts compared to relatively coherent and cohesive texts with a lower proportion of short words and sentences (e.g., Yang et al., 2005).

However, it is likely that there is a limit as to the proportion of short words and sentences that texts can include while remaining relatively coherent and cohesive, and that once this limit is exceeded comprehension suffers. This is because it is reasonable to assume that the process of shortening sentences may reduce the proportion of cohesive devices in texts (e.g., Crossley et al., 2008). This could potentially be problematic as relatively low incidence of connectives could lower the overall text cohesion, and possibly coherence (e.g., Ozuru et al., 2009). In turn, lower coherence and cohesion may require readers to engage reader-initiated active processes, such as inference-making, to comprehend the texts read (e.g., Hamilton & Oakhill, 2014). This may be difficult for some readers without the relevant background knowledge that may be necessary to make such inferences (van Dijk & Kintsch, 1983), as weak connections between the text and prior knowledge could increase the WM demands of meaning-to-text integration processes (Magliano & Schleich, 2000; Zwaan, 2016) (Chapter 7, section 7.2.3.iii). Consequently, comprehension of some readers, such as those with relatively low levels of background knowledge, may be negatively impacted by texts that are incohesive, possibly due to containing a high proportion of very short sentences (cf. Ozuru et al., 2009). However, another investigation measuring comprehension, alongside metacomprehension, with an experimental manipulation of text features would be required to examine this possibility. Next, I discuss the implications related to the effects, and construct, of metacomprehension.

8.2.1.ii. What Does Metacomprehension Tell Us Really?

The metacomprehension variable was based on the combined average of perceived effort and perceived understanding judgement scales, because the ratings of these two judgements scales were found to share 72% of the variance with each other (Chapter 7, section 7.2.3.iv). The relatively high proportion of variance these two judgement scales shared indicates that metacomprehension judgements were in part based on the perceived ease of text processing. This has important theoretical implications, as it provides evidence supporting the ease of processing hypothesis which states that readers are more likely to judge their comprehension higher when the text they read is perceived to be easy to process (Begg et al., 1989; Dunlosky et al., 2006) (Chapter 2, section 2.2). Indeed, variation in text features associated with ease of text processing, such as word and sentence length (e.g., Flesch, 1948; Crossley et al., 2017; McNamara et al., 2013; Perfetti, 2007; Perfetti & Stafura, 2014; Yang et al., 2005), predicted perceived effort required to understand health-related texts (Chapter 6, section 6.3.2.iii).

The evidence reported in this thesis is consistent with the view that metacomprehension judgements can be implemented as comprehension performance predictors, as metacomprehension accuracy is good-enough to predict comprehension in the absence of other information (cf. Maki, 1998; Dunlosky et al., 2005) (Chapter 7, section 7.3.2.v). However, the evidence reported in this thesis is also consistent with the view that metacomprehension accuracy is likely to vary between different readers, thereby the utility of metacomprehension judgements in predicting comprehension may be higher for some readers than others. For example, the more health literate and better educated individuals were more likely to judge their understanding of health-related texts as higher than the lower health literate and less educated individuals (Chapter 6, section 6.3.2.ii). In addition, the higher educated and the more health literate individuals were also more likely to have higher comprehension than the less educated and the less health literate readers (Chapter 7, section 7.3.2.ii). Consequently, it is probable that individuals with higher relevant background knowledge, and education, are more accurate at estimating their comprehension performance than the less educated readers with lower levels of relevant background knowledge (e.g., Griffin et al., 2009; Zabrucky et al., 2012). One explanation for this may be the that the standards of coherence of the more educated and health literate readers are higher than those of the less educated and less health literate readers (van den Broek & Helder, 2017). Therefore, it may be the case that the more educated and health literate readers are more likely to engage in reader-initiated active processing to self-regulate their comprehension (Thiede et al., 2010), and thereby improve their metacomprehension accuracy and comprehension performance.

Some readers may be prone to give biased evaluations of their understanding of health-related texts, however, and these evaluations may not be reflective of their levels of comprehension. Specifically, there is evidence to indicate that older individuals may rate their understanding of health-related texts as higher than younger individuals (Chapter 6, section 6.3.2.ii) but older readers' comprehension may be lower than that of younger individuals (Chapter 7, section 7.3.2.ii). One potential explanation for this age-related decrease in metacomprehension accuracy could be that age-related changes may make tasks such as comprehension monitoring more difficult (e.g., Dunlosky et al., 2006; Miles & Stine-Morrow, 2004) (Chapter 6, section 6.4). Critically, the effects of age on comprehension and metacomprehension, have important theoretical implications as they seem to indicate that the accumulation of health knowledge associated with getting older (e.g., Chin et al., 2009; Gazmararian et al., 2003) is insufficient to offset the negative effects of ageing on processing required to comprehend, and evaluate one's understanding of, health-related texts (Chapter 7, section 7.4). This has important practical implications (discussed in section 8.2.2), and it may be the reason why tests that measure reader characteristics were found to be better predictors of comprehension than metacomprehension judgements (Chapter 7, section 7.3.2.ii).

An alternative explanation for the diminution of the effects of metacomprehension in the presence of other individual differences predictors may be that metacomprehension judgements, that is, perceived understanding and effort judgements, are partially based on perceived vocabulary knowledge (Chapter 7, section 7.2.3.iv, Table 7.7). Consequently, the effects of metacomprehension on comprehension may be indirect, akin to the effects of verbal WM and phonological awareness on comprehension (Freed et al., 2017) (Chapter 7, section 7.3.2.ii). Thus, metacomprehension judgements may be a confound or a proxy measure for other individual differences which have direct effects on comprehension, such as vocabulary knowledge.

In the light of the findings reported in this thesis, it can be argued that comprehension can be understood to consist of processes, such as metacomprehension, that help to regulate comprehension processing by interacting with the developing mental representation of the text and thereby influence comprehension (cf. van den Broek & Helder, 2017). Through metacomprehension, it is probable that comprehension is influenced by textual features, such as word and sentence length. These features may signal the need for active processing, or increase in effort, to readers when passive processes alone are not sufficient to understand the text read due to the demands imposed by the text. Thus, metacomprehension judgements may play an important role in comprehension, as they may mediate the deployment of metacomprehension strategies, such as selective rereading, that are thought to regulate comprehension breakdowns (Thiede et al., 2010) (Chapter 2, section 2.2).

Critically, if metacomprehension judgements are partially based on vocabulary knowledge, it may be the case that the highly educated and older readers rate their understanding of health-related texts as higher due to the accumulation of health, and general, vocabulary knowledge. This may be problematic in the context of older adults as their health vocabulary knowledge may be higher than that of younger adults, but their processing capacity required to monitor their understanding of, and understand, health-related texts may be lower than that of younger adults due to age-related changes (cf. Miles & Stine-Morrow, 2004). Thus, possibly due to, in part, the lack of awareness for the need of deployment of reader-initiated strategies to fix comprehension breaks (cf. van den Broek & Helder, 2017), the comprehension of health-related texts of older adults is likely to be lower compared to that of younger adults (Chapter 7, section 7.3.2.ii). However, an additional investigation would be required to examine this possibility. Next, I discuss the implications of the reported evidence on comprehension in the context of health.

8.2.1.iii. Comprehension in the Health Context

The evidence reported in this thesis considering the individual differences predictors of comprehension (Chapter 7, section 7.3.2.ii) is comparable with the account of Freed et al. (2017). This suggests that the variance in phonological awareness and verbal WM is secondary to that of the relatively robust effects of vocabulary, education, age, and health literacy in predicting reading comprehension of health-related texts. Thereby, the reported evidence has implications for evaluating the role of verbal WM and phonological awareness in comprehension among adult readers. It may be the case that readers perform well on verbal WM measures because they have high levels of vocabulary knowledge (Freed et al., 2017), and that the influence of phonological awareness on comprehension diminishes with reading experience (e.g., Cromley & Azevedo, 2007; Cromley et al., 2010; Vellutino et al., 2007). Thus, models of reading that emphasise the role of phonological awareness in comprehension, such as the Simple View of Reading (Gough & Tunmer, 1986; Tunmer & Chapman, 2012) (Chapter 1, section 1.3), may be more applicable to the study of children rather than proficient adult readers. This is not to say that adult readers do not vary in their phonological awareness at all, but that the differences in phonological awareness between adults are likely to be relatively small and associated with variation in vocabulary knowledge (Castles & Friedman, 2014; Freed et al., 2017; Perfetti, 2010) (Chapter 1, section 1.4; Chapter 2, section 2.1.2). Consequently, a case could be made for directing research resources away from examining the effects of variation in verbal WM and phonological awareness when studying comprehension of adult readers.

In contrast to the effects of verbal WM and phonological awareness, the evidence reported in this thesis is consistent with the view that knowledge of word meanings is crucial to successful comprehension (e.g., Kintsch & Rawson, 2007; Perfetti, 2007; 2010; Tunmer & Chapman, 2012). Thus, the importance of English language vocabulary knowledge in predicting comprehension of English health-related texts should be highlighted for both English monolingual (Perfetti, 2010), and ESL readers (Brysbaert et al., 2016). In terms of theoretical implications, the finding that general and health vocabulary knowledge shared a substantial proportion of variance with self-rated English language proficiency, suggests that vocabulary knowledge is a relatively good proxy for English language proficiency or language exposure (Brysbaert et al., 2016; Kuperman & Van Dyke, 2013). This can be interpreted as indicating that the differences between monolingual and ESL readers, and the differences between ESL readers of varying proficiency levels, can be captured relatively accurately using standardised vocabulary tests (Brysbaert et al., 2016). Thus, as mentioned in

Chapter 7 (section 7.4), improving vocabulary knowledge of monolingual, and ESL, readers may be particularly effective at improving their comprehension of health-related texts.

Another theoretical implication is afforded by the finding that several health literacy measures predicted comprehension (Chapter 7, section 7.3.2.ii). As mentioned in Chapter 7 (section 7.4), the separate effects of health literacy indicate that the different tests of health literacy measure different aspects of the same underlying construct of functional health literacy (Chin et al., 2011). Some measures, such as the MEDCO medicine label (Bostock & Steptoe, 2012), appear to be more process oriented, whereas others, such as the health literacy vocabulary assessment (HLVA; Chapter 7, section 7.2.3.i), are more health knowledge reliant. This evidence is supportive of the view that functional health literacy encompasses health knowledge and processing, and that both aspects of health literacy are important to comprehension of health-related texts (Chin et al., 2011) (Chapter 3, sections 3.1.1 and 3.1.2; Chapter 7, section 7.4). To accommodate these findings, some theories of comprehension may have to be extended to incorporate a processing component in the comprehension system (Chapter 3, section 3.1.1) (cf. Ramscar et al., 2017). This is because the correlations between age and the different health literacy measures (Chapter 7, section 7.2.3.iv), the effects of age on comprehension reported in Chapter 7 (section 7.3.2.ii), and empirical research findings (e.g., Liu et al., 2009), indicate that it is likely that linguistic and cognitive capacities, inclusive of processing, are related to comprehension.

There is lack of evidence for direct effects of variation in text features on comprehension of health-related texts, and the proportion of variance accounted for by the effects of text features was found to be marginal (section 8.1.2). I do not claim that variation in text features does not matter, but it may be that, in the context of health-related texts, the effects of variation in text features on comprehension are relatively small (e.g., Kulesz et al., 2016), and indirect (section 8.2.1.i). The theoretical implication of this is that variation in individual differences may be greater than variation in texts features (e.g., Davies et al., 2017). Indeed, it is plausible that highly educated and proficient readers are more likely to understand all types of texts, regardless of how they are written, than the less educated and less proficient readers.

Nonetheless, there is some evidence to suggest that effects of variation in text features on comprehension may have an indirect influence on comprehension through metacomprehension (section 8.2.1.i). Consequently, adequate comprehension theory should consider the potential direct and indirect effects of variation in text features on comprehension, in the context of metacomprehension and standards of coherence (Chapter 2, sections 2.1.3.ii and 2.2). Specifically, how different texts may signal to readers the necessity for engagement in reader-initiated active processing required to understand the text read at the level desired by the reader, and how this engagement, or lack of, may impact comprehension (sections 8.2.1.i and 8.2.1.ii). However, this consideration of the effects of text features on comprehension should be secondary to the discussion of the overarching effects of individual differences on comprehension. I discuss the practical implications of this next.

8.2.2. Practical Implications

The evidence reported in this thesis highlights the need for the NHS to consider that comprehension of written-in-English health-related texts may be inadequate among the elderly, and individuals who are less educated, or have relatively low levels of English language proficiency and health literacy. The relatively strong effects of health literacy and English language proficiency on comprehension suggest that comprehension of health-related texts may be improved with interventions aimed at educating the population on health and improving English language proficiency of monolingual and low-proficiency ESL speakers. Meta-analyses indicate that vocabulary knowledge, a major component of both English language proficiency and health literacy (Chapter 7), is amendable to training (e.g., Scammacca, Roberts, Vaugh, Stuebing, 2015). Thus, preventative interventions could focus on improving the vocabulary knowledge and health literacy of at-risk groups of ESL and monolingual readers.

Although teaching a significant proportion of the population new vocabulary would require a big initial investment, and may not be easily scalable in practice, there is evidence to indicate that improving health vocabulary is associated with an overall reduction in costs of treating patients (e.g., Weiss & Palmer, 2004). Thus, in the long-term, interventions focusing on improving the literacy levels of patients are likely to be cost-effective, and benefit these patients' health outcomes (e.g., Paasche-Orlow & Wolf, 2007). In the short-term, cost-effective interventions could include medical professionals tailoring their communication to patients' education, language, and health literacy levels (Schillinger et al., 2003), and the use of teach-back (Slater, Huang, & Dalawari, 2017). Teach-back is a method which involves asking individuals to repeat back what they understand, in their own words, so that comprehension can be confirmed, and misunderstandings corrected. Like tailoring communication to patients, teach-back has been found to increase comprehension of health-related texts (Schillinger et al., 2003; Slater et al., 2017). However, while the short-term interventions may be effective, these interventions are targeting the symptoms of poor comprehension, rather than addressing the underlying causes. Consequently, a more sustainable approach would be to focus on the long-term interventions aimed at improving vocabulary knowledge and health literacy of at-risk groups of the population.

Concerning the guidelines adopted by the NHS for writing easy-to-understand health-related texts (e.g., NHS England, 2018a; Plain English Campaign, 2018), the evidence reported in Chapter 7 (section 7.3.2.ii) suggests that following these guidelines is unlikely to result in direct changes to understanding of health-related texts. As mentioned in Chapter 4 (section 4.1) no evidence base is cited to support the effectiveness of the NHS guidelines (e.g., NHS England, 2018a; Plain English Campaign, 2018), and the utility of the readability indices have seldom been validated in terms of their capacity to predict comprehension (e.g., Flesch, 1948). Nonetheless, the prevalence of the use of the guidelines, and the widespread use of the readability indices in the context of health-related texts (Wang et al., 2013), made it entirely reasonable to expect to see effects of the associated variables on text comprehension. It is unprecedented that there was no sign of those effects. Overall, the evidence reported in this thesis suggests that the guidelines or the readability indices could be practically useless because the effects of the associated variables are comparatively subtle, meaning of limited practical significance. However, this is a strong claim that requires further evidence obtained from an investigation using a larger number of health-related texts.

In contrast to relying on following the guidelines used by the NHS on how to write easy-to-understand health-related texts (e.g., NHS England, 2018a; Plain English Campaign, 2018), end-user evaluations of health-related texts are likely to be relatively effective at predicting comprehension. This is because the evidence reported in Chapter 7 (section 7.3.2.ii) indicates that, in the absence of other information about the target population, metacomprehension judgements are likely to be useful as effective predictors of comprehension of health-related texts. Thus, one practical recommendation that can be made is to either obtain more person-level information to make more precise comprehension predictions or to continue to rely on end-user evaluations of the comprehensibility of health-related texts as a proxy for tested comprehension.

Importantly, when relying on metacomprehension judgements, care must be taken to ensure that selected groups of end-users, or reader panel members, are demographically diverse. This is because high metacomprehension, as well as high comprehension, were associated with high education and health literacy (Chapter 6, section 6.3.2.ii; Chapter 7, section 7.3.2.ii), suggesting that highly educated and literate individuals are likely to understand more, and may be more accurate at judging their comprehension, than their less educated and literate counterparts. However, it is questionable whether reader panel members should consist of elderly individuals, as older readers were found to report higher perceived understanding compared to younger readers (Chapter 6, section 6.3.2.ii), but their tested comprehension of health-related texts was lower than that of the younger participants (Chapter 7, section 7.3.2.ii). Although, on average, metacomprehension judgements predicted comprehension performance in the absence of other information, the discrepancy between metacomprehension and tested comprehension of older individuals may lead to biased evaluations of health-related texts. Thus, to improve the relative accuracy of metacomprehension judgements of the health-related texts considered, reader panels may benefit from including a combination of predominantly younger, and some older, members with varying levels of education, and health literacy levels.

Critically, focusing on improving texts alone may not be enough to affect health outcomes, health status, or adherence to a set of instructions (e.g., Squiers, Peinado, Berkman, Boudewyns, & McCormack, 2012). Given the importance of individual differences, as demonstrated in the findings reported in this thesis, it is important not to dismiss person-level interventions in favour of text-level interventions. It may be the case that redesigning health-related texts to match the characteristics of readers is just one intervention that should be used to increase comprehension and adherence to health instructions. Person-level interventions, such as teach-back (Slater et al., 2017), could be combined with text-level interventions to improve understanding of health-related texts. However, it may be the case that some patients will not adhere to the recommendations written in text that they do understand, due to potential mediating factors, such as motivation, social support, fatalism, access to health care, decision making skills, trust in the information, emotions, self-efficacy, attitudes, perceived relevance of the message, and perceived effectiveness of the required behaviour (Squiers et al., 2012). For example, a chronic smoker may understand that smoking can cause cancer but may not have the motivation to quit. Thus, additional interventions are likely to be required to influence motivation and the multitude of potential mediating factors between comprehension and adherence to instruction. Overall, a multidisciplinary investigation may be required to offer greater insights into the processes involved in understanding health-relating texts, and in acting on that understanding (e.g., Chin et al., 2018).

8.3. Conclusion

The benefit of mixed-effects analysis of reading comprehension is that it had the potential to afford insights into how health-related texts could be adapted for different groups of users to optimise understanding, by examining how the effects of text features may modulate the effects of reader characteristics. However, no evidence was found for these hypothesised modulation effects, suggesting that adaptations of health-related texts for different users may not readily have detectable impacts on comprehension. The evidence reported in this thesis is consistent with the view that the differences between individuals are more important than the differences between texts (e.g., Davies et al., 2017; Kulesz et al., 2016). The effects of English language proficiency, education, age, and health literacy, were found to be robust predictors of comprehension, highlighting the importance of general and health vocabulary knowledge to successful comprehension of health-related texts. In contrast, the effects of variation in text features of health-related texts were found to have no impact on variation in text comprehension (sections 8.1.2; 8.2.1.i; and 8.2.1.ii), and were overshadowed by the effects of individual differences (Chapter 7, section 7.3.2.ii).

The findings reported in this thesis can be explained by theoretical accounts that consider the variance shared between different reader characteristics to account for lack of direct effects of verbal WM and phonological awareness (e.g., Freed et al., 2017), and account for the developmental nature of the effects of some of these variables, such as phonological awareness, on comprehension of individuals of different ages (e.g., Tunmer & Chapman, 2012). In addition, it may be the case that metacomprehension and standards of coherence are important to understanding the potential effects of variation in text features on comprehension through their effects on metacomprehension (see sections 8.2.1.i and 8.2.1.ii). Thus, an adequate comprehension theory could also incorporate standards of coherence and metacomprehension as the basis for successful comprehension among adult readers (e.g., van den Broek & Helder, 2017) (Chapter 2, sections 2.1.3.ii and 2.2).

The lack of supporting evidence for the effects of text features indicates that the most effective way to improve comprehension of health-related texts is likely to be offered by interventions focusing on individuals rather than on texts (section 8.2.2). Although, text-level manipulations offer the promise of a relatively low-cost easy-fix solution to improving comprehension of at-risk populations, the variance accounted for by text features in this thesis (Chapter 7, section 7.3.2.ii), and some published research (e.g., Davies et al., 2017; Kulesz et al., 2016), shows that focusing exclusively on text-level interventions is not the right approach to improving comprehension. Concentrating on individuals may also be more ethical as, in addition to improving comprehension, interventions focusing on readers are likely to address the root causes of poor comprehension, such as limited vocabulary and health literacy levels, rather than simply making text-level accommodations for these causes.

Nevertheless, text-level interventions are likely to be beneficial for some groups of the population, such as the elderly. This is because, at present, no intervention targeting individuals can reverse the process of ageing, and evidence indicates that older individuals are less likely to understand health-related texts compared to younger individuals (e.g., Chapter 7, section 7.3.2.ii; Liu et al., 2009), potentially due to changes in processing (Chin et al., 2011; Kobayashi et al., 2015; 2016; Liu et al., 2009) (Chapter 3, section 3.1.2; Chapter 7, section 7.4). Therefore, efforts should continue to be made to produce comprehensible texts, in particular to improve the comprehension of the elderly for whom interventions aimed at improving vocabulary knowledge may be insufficient to counteract the age-related changes in processing that is required to understand health-related texts (section 8.2.1.ii).

Overall, there is a practical need for research examining both text- and person-level interventions to improve comprehension of health-related texts amongst individuals who are older, less educated, and who might have lower English language proficiency and literacy levels. In addition, there is a theoretical need to continue reading comprehension research to progress reading comprehension theory development, and to understand what person- and text-level interventions are likely to be optimally effective. However, future research should be aware that robust investigations require large samples of participants and texts. I briefly discuss this issue next.

8.3.1. Limitations

This thesis has shown that investigating the effects of text features on the comprehension of health-related texts, and the potential modulations of individual differences by text features, requires large samples of participants and texts. This is because the effects of text features on comprehension are likely to be relatively small and overshadowed by the effects of variation in reader characteristics (e.g., Davies et al., 2017; Kulesz et al., 2016) (Chapter 7, section 7.3.2.ii). Consequently, as demonstrated using the relatively large sample of participants used in this thesis (Chapter 7, section 7.2.2), the effects of variation in text features are likely to be difficult to detect using small samples of participants and texts. To be able to detect the effects of variation in text features on comprehension, if these effects are truly there, future research should recruit large number of participants and obtain a relatively large number of observations per reader by asking each individual to read a relatively large number of texts.

8.3.2. Directions for Future Research

Although, I found limited support for the effectiveness of variation in text features on comprehension, research into the effects of text features, and the potential modulations of the effects of individual differences by text features, should continue. There are two reasons for this, which could form two alternative hypotheses in a future investigation. First, if the effects of variation in text features are of limited significance in the context of health-related texts, then interventions aimed at improving comprehensibility of health-related texts using text-level manipulations, such as through adoption of NHS guidelines (e.g., NHS England, 2018a), could be practically useless (null hypothesis). Second, and in contrast to the first reason, it may be the case that the relatively small potential effects of text features could have large aggregated consequences for the understanding of health-related texts of the UK's population (research hypothesis). Thus, future research should examine what might be accomplished with manipulations of health-related texts, and a larger sample of readers and texts, to investigate which hypothesis is supported by the obtained evidence. Critically, continuing this research is vital for public health, as lack of comprehension is associated with worse health outcomes (e.g., Baker et al., 2002; Bostock & Steptoe, 2012; Schillinger et al., 2002). Importantly, based on the findings of future research it may be possible to determine what interventions resources should be spent on to improve health outcomes and maximise understanding of health-related texts efficiently and cost-effectively.

References

Agresti, A. (2010). Chapter 3: Logistic regression models using cumulative logits. In *Analysis of ordinal categorical data (2nd ed.)* (pp. 44-87). New Jersey: Wiley.

Alberti, T. L., & Morris, N. J. (2017). Health literacy in the urgent care setting: What factors impact consumer comprehension of health information? *Journal of the American Association of Nurse Practitioners, 29*, 242–247.

Albright, J., de Guzman, C., Acebo, P., Paiva, D., Faulkner, M., & Swanson, J. (1996). Readability of patient education materials: implications for clinical practice. *Applied Nursing Research, 9*, 139–143.

Andrews, S. (2015). Chapter: Individual Differences Among Skilled Readers: The Role of Lexical Quality. In Pollatsek, A., & Treiman, R. (Eds.), *The Oxford Handbook of Reading* (pp. 129-148). New York: Oxford University Press.

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General, 139*, 299–318.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database [CD ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Badarudeen, S., & Sabharwal, S. (2010). Assessing readability of patient education materials: current role in orthopaedics. *Clinical Orthopaedics and Related Research, 468*, 2572–2580.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417-423.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In: *Recent advances in learning and motivation*, Vol. 8 (Bower GA, ed), 47–90. New York: Academic Press.

Baker, D. W., Gazmararian, J. A., Williams, M. V., Scott, T., Parker, R. M., Green, D., Ren. J., & Peel, J. (2002). Functional health literacy and the risk of hospital admission among Medicare managed care enrolees. *American Journal of Public Health, 92*, 1278–1283.

Baker, D. W., Williams, M. V., Parker, R. M., Gazmararian, J. A., & Nurss, J. (1999). Development of a brief test to measure functional health literacy. *Patient Education and Counseling, 38*, 33–42.

Balota, D. A., Cortese, M, J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*(2), 283-316.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. arXiv:1506.04967v2.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255-278.

Beck, I., McKeown, M., & Gromoll, E. (1989). Learning from social studies texts. *Cognition and Instruction, 6*, 99–158.

Beck, I. L., McKeown, M. G., Sinatra, G. M., & Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly, 26*, 251–276.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632.

Berkman, N. D., Davis, T. C., & McCormack, L. (2010). Health literacy: What is it? *Journal of Health Communication, 15*, 9–19.

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., Viera, A., Crotty, K., Holland, A., Brasure, M., Lohr, K. N., Harden, E., Tant, E., Wallace, I., & Viswanathan, M. (2011). Health Literacy Interventions and Outcomes: An Updated Systematic Review (Evidence Reports/Technology Assessments, No. 199). Retrieved from Agency for Healthcare Research and Quality, U.S. Department of Health & Human Services website https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/health-literacy_research.pdf

BNC Consortium. (2007). British National Corpus, version 3 (BNC XML ed.). Retrieved from http://www.natcorp.ox.ac.uk

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24,* 127-135.

Bostock, S., & Steptoe, A. (2012). Association between low functional health literacy and mortality in older adults: longitudinal cohort study. *BMJ, 344,* e1602. Doi:10.1136/bmj.e1602

Bowles, M. A., (2018). Introspective Verbal Reports: Think-Alouds and Stimulated Recall. In Phakiti, A., De Costa, P., Plonsky, L., & Starfield, S. (Eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology* (pp. 339-358). London: Palgrave.

Braze, D., Katz, L., Magnuson, J. S., Mencl, W. E., Tabor, W., Van Dyke, J. A., Gong, T., Johns, C. L., & Shankweiler, D. P. (2016). Vocabulary does not complicate the simple view of reading. *Reading and writing, 29*, 435–451.

Braze, D., Tabor, W., Shankweiler, D., & Mencl, W. E. (2007). Speaking up for vocabulary: Reading skill differences in young adults. *Journal of Learning Disabilities, 40,* 226–243.

Brown, J. D. (1998). An EFL readability index. *JALT Journal*, *20*, 7–36.

Brown, J., Bennett, J., & Hanna, G. (1980). *The Nelson-Denny reading test*. Boston: Houghton Mifflin.

Brown, J., Fishco, V., & Hanna, G. (1993). *Nelson-Denny reading test: Manual for scoring and interpretation, Forms G & H.* Chicago: Riverside Press.

Brown, J. D., Janssen, G., Trace, J., & Kozhevnikova, L. (2012). A Preliminary Study of Cloze Procedure as a Tool for Estimating English Readability for Russian Students. *Second Language Studies, 31*(1), 1-22.

Bruck, M. (1990). Word-recognition skills of adults with childhood diagnoses of dyslexia. *Developmental Psychology, 26,* 439 – 454.

Brysbaert, M., Lagrou, E., & Stevens, M. (2016). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition, 20*(3), 530–548

Bürkner, P-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1), 1-28.

Bürkner, P-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal, 10*(1), 395-411.

Bürkner, P-C., & Vuorre, M. (2018). Ordinal Regression Models in Psychology: A Tutorial. PsyArXiv preprint: https://psyarxiv.com/x8swp/

Burns, M, K., Kwoka, H., Lim, B., Crone, M., Haegele, K., Parker, D. C., Petersen, S., & Scholin, S. E. (2011). Minimum reading fluency necessary for comprehension among second-grade students. *Psychology in the Schools, 48*(2), 124–132.

Burrow, M., & Forrest, M. (2015). *Creating Patient Information for Patients/Relatives/Carers.* Internal Blackpool Teaching Hospitals NHS Foundation Trust report: Unpublished.

Cain, K. (2006). Individual differences in children's memory and reading comprehension: An investigation of semantic and inhibitory deficits. *Memory, 14*(5), 553–569.

Cain, K., & Oakhill, J. V. (1999). Inference making and its relation to comprehension failure. *Reading and Writing, 11*, 489-503.

Cain, K., & Oakhill, J. V. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*, 697-708.

Cain K., Oakhill J. V., Barnes M. A., & Bryant P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition, 29*, 850-859.

Cain, K., Oakhill, J. V., & Bryant, P. E. (2000). Phonological skills and comprehension failure: A test of the phonological processing deficit hypothesis. *Reading and Writing, 13,* 31–56.

Cain, K., Oakhill, J. V., & Bryant, P. E. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31–42.

Calvo, N., Ibáñez, A., & García, A. M. (2016). The impact of bilingualism on working memory: a null effect on the whole may not be so on the parts. *Frontiers in Psychology, 25*(7), 265.

Cambridge University Press. (2018). *Cambridge English Dictionary*, Cambridge Dictionary Online. Retrieved May 18, 2018, from: https://dictionary.cambridge.org/dictionary/english/

Carrell, P. L. (1987) Readability in ESL. *Reading in a Foreign Language*, *4*(1), 21– 40.

Castles, A., & Coltheart, M. 1996: Cognitive Correlates of Developmental Surface Dyslexia: A Single Case Study. *Cognitive Neuropsychology, 13*, 25–50.

Castles, A., & Friedmann, N. (2014). Developmental Dyslexia and the Phonological Deficit Hypothesis. *Mind & Language, 29*(3), 270–285.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49*(2), 278-293.

Charbonneau, D. H. (2013). Health Literacy and the Readability of Written Information for Hormone Therapies. *Journal of Midwifery and Women's Health, 58*, 265-270.

Chew, L., Bradley, K. A., & Boyko, E. J. (2004). Brief questions to identify patients with inadequate health literacy. *Family Medicine, 36*(8), 588–594.

Chew, L., Griffin, J., Partin, M., Noorbaloochi, S., Grill, J., Snyder, A., et al. (2008). Validation of screening questions for limited health literacy in a large VA outpatient population. *Journal of General Internal Medicine, 23*(5), 561–566.

Chin, J., D'Andrea, L., Morrow, D. G., Stine-Morrow, E. A., Conner-Gercia, T., Graumlich, J. F., & Murray, M. D. (2009). Cognition and illness experience are associated with illness knowledge among older adults with hypertension. In *Proceedings of the 53rd annual meeting of the Human Factors and Ergonomics Society 2009*. Santa Monica, CA: Human Factors and Ergonomics Society.

Chin, J., Moeller, D. D., Johnson, J., Duwe, E. A. G., Graumlich, J. F., Murray, M. D., & Morrow, D. G. (2018). A Multi-faceted Approach to Promote Comprehension of Online Health Information Among Older Adults. *Gerontologist, 58*(4), 686–695.

Chin, J., Morrow, D. G., Stine-Morrow, E. A. L., Conner-Garcia, T., Graumlich, J. F., & Murray, M. D. (2011). The process-knowledge model of health literacy: Evidence from a componential analysis of two commonly used measures. *Journal of Health Communication, 16*(3), 222–241.

Chin, J., Payne, B., Gao, X., Conner-Garcia, T. Graumlich, J., Murray, M. D., Morrow, D. G., & Stine-Morrow, E. A. L. (2015). Memory and comprehension for health information among older adults: Distinguishing the effects of domain-general and domain-specific knowledge. *Memory, 23*, 577-589.

Christ, T. (2011). Moving past "right" or "wrong" toward a continuum of young children's semantic knowledge. *Journal of Literacy Research, 43*(2), 130-158.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerative penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika, 78*, 685-709.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models. *Journal of Educational and Behavioural Statistics, 40*(2), 136-157.

Clark, S. K., Jones, C. D., & Reutzel, D. R. (2013). Using text structures of information books to teach writing in the primary grades. *Early Childhood Education Journal, 41*(4), 265-271.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI*, pp. 535-555. New York: Academic Press.

Conway, A. R. A., Kane, M. J., & Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user’s guide. *Psychonomic Bulletin & Review, 12*(5), 769-786.

Cortez, S., Milbrandt, M., Kaphingst, K., James, A., & Colditz, G. (2015). The readability of online breast cancer risk assessment tools. *Breast Cancer Research and Treatment, 154*, 191–199.

Cowan, N. (2010). The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, *19*(1), 51–57.

Crawford, J. R., Parker, D. M., & Besson, J. A. O. (1988). Estimation of premorbid intelligence in organic conditions. *British Journal of Psychiatry, 153*, 178–181.

Cromley, J. G., & Azevedo, R. (2007). Testing and Refining the Direct and Inferential Mediation Model of Reading Comprehension. *Journal of Educational Psychology, 99*(2), 311–325.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading Comprehension of Scientific Text: A Domain-Specific Test of the Direct and Inferential Mediation Model of Reading Comprehension. *Journal of Educational Psychology, 102*(3), 687–700.

Crossley, *S. A.,* Allen, *D., &* McNamara, *D. S. (*2011*).* Text readability and intuitive simplification*: A comparison of* readability formulas. *Reading in a Foreign Language, 23*(1), 84-101.

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research, 16*(1), 89-108.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. *Proceedings of the 29th annual conference of the Cognitive Science Society* (pp. 197–202). Nashville, TN: Cognitive Science Society.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475–493.

Crossley, S. A., McCarthy, P. M., Louwerse, M. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal 91*(2), 15–30.

Crossley, S. A. & McNamara, D. S. (2008). Assessing Second Language Reading Texts at the Intermediate Level: An approximate replication of Crossley, Louwerse, McCarthy, and McNamara (2007). *Language Teaching*, *41* (3), 409–229.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243-263.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes, 54*(5-6), 340-359.

Cummins, J. (2000). Chapter 7. The Threshold and Interdependence Hypotheses Revisited. In *Language, Power and Pedagogy: Bilingual Children in the Crossfire* (pp. 173-200). Clevedon: Multilingual Matters

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Eds.), *Schooling and language minority students: A theoretical framework* (pp. 3-49). Los Angeles: Evaluation, Dissemination, and Assessment Center, California State University, Los Angeles.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42*(8), 895-948.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior, 19*(4), 450–466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3,* 422-433.

Davies, R. A. I., Arnell, R., Birchenough, J., Grimmond, D., & Houlson, S. (2017). Reading Through the Life Span: Individual Differences in Psycholinguistic Effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* DOI: 10.1037/xlm0000366

Davis, T. C., Long, S. W., Jackson, R. H., Mayeaux, E. J., George, R. B., Murphy, P. W., Crouch, M. A. (1993). Rapid estimate of adult literacy in medicine: a shortened instrument. *Family Medicine, 25,* 391–395.

Davis, T. C., Wolf, M. S., Bass, P. F., Middlebrooks, M. Kennen, E., Baker, D. W., Bennett, C. L., Durazo-Arvizu, R., Bocchini, A., Savory, S., & Parker, R. M. (2006). Low literacy impairs comprehension of prescription drug warning labels. *Journal of General Internal Medicine, 21,* 847–851.

Dekeyser, R., & Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkel (ed.) *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, 395-406). Routledge: Taylor and Francis.

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods, 22*(2), 240-261.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first and second language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology, 66*, 843–863.

Dijkstra, A. T., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition, 5*(3), 175-197.

Dirlam, D. K. (1972). Most efficient chuck sizes. *Cognitive Psychology, 3,* 355-359.

Doctorow, J., Wittrock, M. C., & Marks, C.B. (1978). Generative processes in reading comprehension. *Journal of Educational Psychology, 70,* 109-118.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J. R., Gruber, B., Lafourcade, B., Leitao, P. J., Munkemuller, T., McClean, C., Osborne, P, E., Reineking, B., Schroder, B., Skidmore, A. K., Zurrel, D., & Lautenbach, S.

(2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46.

Dowell, N. M. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from Educational material to learning environments at scale. *Journal of Learning Analytics, 3*(3), 72–95.

Dumenci, L., Matsuyama, R. K., Kuhn, L., Perera, R. A., & Siminoff, L. A. (2013). On the Validity of the Shortened Rapid Estimate of Adult Literacy in Medicine (REALM) Scale as a Measure of Health Literacy. *Communication Methods and Measures, 7*(2), 134-143.

Dunlosky, J., Baker, J. M. C., Rawson, K. A., & Hertzog, C. (2006). Does Aging Influence People's Metacomprehension? Effects of Processing Ease on Judgments of Text Learning. *Psychology and Aging, 21*(2), 390-400.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A Brief History and How to Improve Its Accuracy. *Current Directions in Psychological Science, 16*(4), 228-232.

Dunlosky, J., Rawson, K.A., & Middleton, E. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52,* 551–565.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83–87.

Dymock, S. (2007). Teaching tips comprehension strategy instruction: Teaching narrative text structure awareness. *The Reading Teacher, 61*(2), 161-167

Eager, C., & Roy, J. (2017). Mixed Effects Models are Sometimes Terrible. ArXiv preprint: https://arxiv.org/abs/1701.04858

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Elwér, A., Keenan, J. M., Olson, R. K., Byrne, B., & Samuelsson, S. (2013). Longitudinal stability and predictors or poor oral comprehenders and poor decoders. *Journal of Experimental Child Psychology, 115,* 497–516.

Engen, L., & Høien, T. (2002). Phonological skills and reading comprehension. *Reading and Writing: An Interdisciplinary Journal, 15*(7–8), 613–631.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–233.

Francis, D. J., Kulesz, P. A., & Benoit, J. S. (2018). Extending the Simple View of Reading to Account for Variation Within Readers and Across Texts: The Complete View of Reading (CVRi). *Remedial and Special Education, 39*(5), 274-288.

Frederickson, N., Frith, U., & Reason, R. (1997). *Phonological Assessment Battery [PhAB]: Manual and test materials.* London: NFER-Nelson Publishing Company Ltd.

Freed, E. M., Hamilton, S. T., & Long, D. L. (2017). Comprehension in proficient readers: The nature of individual variation. *Journal of Memory and Language, 97,* 135-153.

Friedman, D. B., & Hoffman-Goetz, L. (2007). An Exploratory Study of Older Adults' Comprehension of Printed Cancer Information: Is Readability a Key Factor? *Journal of Health Communication, 12*(5), 423-437.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Researching Methods, 37*, 581–590.

Friedmann, N., & Rahamim, E. (2007). Developmental letter position dyslexia. *Journal of Neuropsychology, 1*, 201–36.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5,* 239–256.

Gabay, Y., & Holt, L. L. (2015). Incidental learning of sound categories is impaired in developmental dyslexia. *Cortex, 73*, 131–143.

Garcia, J. R., & Cain, K. (2014). Decoding and reading comprehension: a meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*(1), 74-111.

Gazmararian, J. A., Williams, M. V., Peel, J., & Baker, D. W. (2003). Health literacy and knowledge of chronic disease. *Patient Education and Counseling, 51*, 267–275.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review, 71*(2), 369–382.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27,* 2865-2873.

Gelman, A. (2015). The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management, 41*(2), 632–643.

Gelman, A. (2018). *Prior Choice Recommendations.* Retrieved October 2018, from https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Gelman, A., & Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9,* 641-651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC Press, London, third edition.

Gelman, A., & Geurts, H. M. (2017). The statistical crisis in science: how is it relevant to clinical neuropsychology? *The Clinical neuropsychologist, 31*(6-7), 1000-1014.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2018). R-squared for Bayesian regression models. *The American Statistician,* doi:10.1080/00031305.2018.1549100.

Gelman, A., & Henning, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society, 180*(4), 967-1033.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*(4), 1360-1383.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy, 19*(555). doi:10.3390/e19100555

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing, 25,* 1819-1845.

Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second languages. *Language Learning, 43,* 5–42.

Ghosh, J., Li, Y., & Mitra, R. (2018). On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression. *Bayesian Analysis, 13*(2), 359-383.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In Cornoldi, C., & Oakhill, J. V. (Eds.), *Reading comprehension difficulties: Processes and remediation* (pp. 1–13). Mahwah, NJ: LEA.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6–10.

Grabe, W. (2014). Key Issues in L2 Reading Development. *In Deng, X., & Seow, R. 4th Centre for English Language Communication Symposium Proceedings. Symposium conducted at the National University of Singapore.*

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher, 40*(5), 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193-202.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Green, A., Khalifa, H., & Weir, C. (2013). Examining textual features of reading texts. *Cambridge English Language Assessment Research Notes, 52,* 24-39

Green, P., & MacLeod, C. J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493-498.

Greenfield, G. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Unpublished doctoral dissertation, Temple University, Philadelphia, PA, United States. (University Microfilms No. 99–38670).

Grober, E., & Sliwinski, M. (1991). "Development and Validation of a Model for Estimating Premorbid Verbal Intelligence in the Elderly". *Journal of Clinical and Experimental Neuropsychology, 13*(6), 933–949.

Grundy, J. G., & Timmer. K. (2017). Bilingualism and working memory capacity: A comprehensive meta-analysis. *Second Language Research, 33*(3), 325–40.

Hakim, H., Provencher, T., Chambers C. T., Driedger, S. M., Dube, E., Gavaruzzi, T., Giguere, A. M. C., Ivers, N. M., MacDonald, S., Paquette, J-S., Wilson, K., Reinharz, D., & Witteman, H. O. (2019). Interventions to help people understand community immunity: A systematic review. *Vaccine, 37*(2), 235-247.

Hall, N., Birt, L., Rees, C. J., Walter, F. M., Elliot, S., Ritchie, M., Weller, D., & Rubin, G. (2016). Concerns, perceived need and competing priorities: a qualitative exploration of decision-making and non-participation in a population-based flexible sigmoidoscopy screening programme to prevent colorectal cancer. *BMJ Open 6*(11):e012304.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. Hong Kong, China: Longman

Hannon, B., & Daneman, M. (2009). Age-Related Changes in Reading Comprehension: An Individual-Differences Perspective. *Experimental Aging Research, 35*(4), 432-456.

Hamilton, S. T., & Oakhill, J. V. (2014). Establishing coherence across sentence boundaries: an individual differences approach. *Language, Cognition and Neuroscience, 29*(10), 1240-1248,

Harm, M.W., & Seidenberg, M.S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662-720.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*, 491–528.

Harte, E., MacLure, C., Martin, A., Saunders, C. L., Meads, C., Walter, F. M., Griffin, S. J., Mant, J., & Usher-Smith, J. A. (2018). Reasons why people do not attend NHS Health Checks: a systematic review and qualitative synthesis. *British Journal of General Practice, 68*(666), 28-35.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioural and Brain Sciences, 33*, 61-83.

Heredia, R. R., Altarriba, J., & Cieślicka, A. B. (2015). *Methods in Bilingual Reading Comprehension Research.* New York: Springer-Verlag.

Hong-Nam, K., & Page, L. (2014). Investigating Metacognitive Awareness and Reading Strategy Use of EFL Korean University Students. *Reading Psychology, 35*(3), 195-220.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127–160.

Hulme, C., Nash, H. M., Gooch, D. C., Arne Lervåg, A., & Snowling, M. J. (2015). The foundations of literacy development in children at family risk of dyslexia. *Psychological Science, 26*(12), 1877–1886.

Hulme, C., & Snowling, M. J. (2016). Reading disorders and dyslexia. *Current Opinion in Pediatrics, 28*(6), 731-735.

Jenkins, J. R., Fuchs, L. S., van den Broeck, P., Espin, C., & Deno, S. L., (2003). Sources of individual difference in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719–729.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science, 4*, 72-115.

Kabacoff, R. I. (2011). Chapter 8: Regression. In *R in Action: Data analysis and graphics with R* (pp. 173-218). Manning Publications: Shelter Island, New York.

Kamalski, J., Sanders, T., & Lentz, L. (2008). Coherence Marking, Prior Knowledge, and Comprehension of Informative and Persuasive Texts: Sorting Things Out. *Discourse Processes, 45*, 323–345.

Kauchak, D., & Leroy, G. (2016). Moving beyond readability metrics for health-related text simplification. *IT Professional, 18*(3), 45-51.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading Comprehension Tests Vary in the Skills They Assess: Differential Dependence on Decoding and Oral Comprehension. *Scientific Studies of Reading, 12*(3), 281-300.

Kelemen, W. L., Frost, P. J., & Weaver, C. A., III. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28,* 92–107.

Kendeou, P., van den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research and Practice, 29*, 10-16.

Kern, R. G. (1994). The role of mental translation in second language reading. *Studies in Second Language Acquisition, 16,* 441-461.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. New York, NY: Cambridge University Press.

Kintsch, W. (1988). The Role of Knowledge in Discourse Processing: A Construction-Integration Model. *Psychological Review, 95,* 163-182.

Kintsch, W., Patel, V., Ericsson, K. A. (1999). The Role of Long-Term Working Memory in Text Comprehension. *Psychologia, 42,* 186-198.

Kintsch, W., & Rawson, K.A. (2007). Comprehension. In M. J. Snowling and C. Hulme, (Eds.), *The Science of Reading: A Handbook* (p. 209-226). Oxford: Blackwell Publishing.

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85* (5), 363-394.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L.G. Nilsson (Ed.), *Perspectives on memory research: Essays in honor of Uppsala University's 500th anniversary* (pp. 329-365). Hillsdale, NJ: Erlbaum.

Kobayashi, L., Wardle, J., Wolf, M., & von Wagner, C. (2016). Aging and Functional Health Literacy: A Systematic Review and Meta-Analysis. *Journals of Gerontology: Psychological Sciences, 71*(3), 445-457.

Kobayashi, L., Wardle, J., Wolf, M., & von Wagner, C. (2015). Cognitive function and health literacy decline in a cohort of aging English adults. *Journal of General Internal Medicine, 30*(7), 958–64.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134.

Kruschke, J. K., & Liddell, T. M. (2018a). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178–206.

Kruschke, J. K., & Liddell, T. M. (2018b). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 155-177.

Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology, 108*, 1078–1097.

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance, 39*(3), 802–823.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757–786.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*(3), 1030-1046.

Language and Reading Research Consortium (LARRC; 2015). Learning to Read: Should We Keep Things Simple? *Reading Research Quarterly, 50*(2), 151-169. doi: 10.1002/rrq.99

Lastine-Sobecks, J. L., Jackson, S. T., & Paolo, A. M. (1998). Identifying the pronunciation of irregularly spelled words: Relation to verbal IQ. *The Clinical Neuropsychologist, 12*, 189–192.

Law, J. M., Vandermosten, M., Ghesquiere, P., & Wouters, J. (2014). The relationship of phonological ability, speech perception, and auditory perception in adults with dyslexia. *Frontiers in human neuroscience, 8*, 482.

Lee, S.-Y., Stucky, B. D., Lee, J. Y., Rozier, R. G., & Bender, D. E. (2010). Short assessment of health literacy–Spanish and English: A comparable test of health literacy for Spanish and English speakers. *Health Services Research, 45*(4), 1105–1120.

Lehman, S., & Schraw, G. (2002). Effects of coherence and relevance on shallow and deep text processing. *Journal of Educational Psychology, 94*, 738–750.

Leroy, G., Endicott, J. E., Kauchak, D., Mouradi, O., & Just, M. (2013). User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of Medical Internet Research, 15*(7), e144.

Leroy G, Kauchak D. (2014). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association, 21*, e169–e172.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100* (9), 1989-2001.

Li, S. C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the Couplings Among Intellectual Abilities and Constituent Cognitive Processes Across the Life Span. *Psychological Science, 15*(3), 155-163.

Lin, L., Zabrucky, K. M., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology, 115,* 187–198.

Linderholm, T., Everson, M., van den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of causal text revisions on more- and less-skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction, 18,* 525–556.

Lisman, I. E., & Idiart, M. A. P. (1995). Storage of 7 + 2 short-term memories in oscillatory subcycles. *Science, 267,* 1512-1515.

Liu, C-J., Kemper, S., & Bovaird, J. A. (2009). Comprehension of health-related written materials by older adults. *Educational Gerontology, 35*(7), 653-668.

Liu, C-J., Yates, K. E., & Rawl, S. M. (2013). Readability and text cohesion of online colorectal cancer and screening information. *American Medical Writers Association Journal, 28*(4), 146-151.

Lorch, R. F., Lorch, E. P., & Inman, W. E. (*1993*). Effects of signaling topic structure on text recall. *Journal of Educational Psychology, 85*(2), 281-290.

Lorge, I. (1939). Predicting reading difficulty of selections for children. *The Elementary English Review, 16*(6), 229-233.

Luck, S. J., & Vogel, E. K. (1998). Response from Luck and Vogel. (A response to "Visual and auditory working memory capacity", by Cowan, N., in the same issue). *Trends in Cognitive Sciences, 2*, 78-80.

Lukasik, K. M., Lehtonen, M., Soveri, A., Waris, O., Jylkkä, J., & Laine, M. (2018). Bilingualism and working memory performance: Evidence from a large-scale online study. *PloS one, 13*(11), e0205916. https://doi.org/10.1371/journal.pone.0205916

MacGregor, J. N. (1987). Short-term memory capacity: Limitation or optimization? *Psychological Review, 94*, 107-108.

MacKay, D. J. C. (2003). Chapter 28: Model Comparison and Occam's Razor. In *Information Theory, Inference, and Learning Algorithms* (pp. 343-355). Cambridge University Press.

Mackey, A., & Gass, S. M. (2016). Common Data Collection Measures. In Mackey, A., & Gass, S. M. (Eds.), *Second Language Research: Methodology and Design (2nd ed.)* (pp. 52-111). New York: Taylor and Francis.

Magliano, J.P. & Schleich, M.C. (2000). Verb aspect and situation models. D*iscourse Processes, 29,* 83–112.

Maki, R. H. (1998). Test predictions over text material. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.), *The educational psychology series. Metacognition in educational theory and practice* (pp. 117-144). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*, 723-731.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language, 94*, 305-315.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Chichester, West Sussex: John Wiley & Sons Ltd.

Martin, E. A. (Ed.). (2015). *Concise Colour Medical Dictionary* (Sixth Edition ed.). Oxford: Oxford University Press.

Martin, S. R., & Williams, D. R. (2017). Outgrowing the Procrustean Bed of Normality: The Utility of Bayesian Modeling for Asymmetrical Data Analysis. PsyArXiv preprint: https://psyarxiv.com/26m49/

Marsay, S. (2017a). *Accessible Information Standard Implementation Guidance.* Retrieved from https://www.england.nhs.uk/publication/accessible-information-standard-implementation-guidance/ Accessed 07th January, 2019.

Marsay, S. (2017b). *Accessible Information Standard Specification.* Retrieved from https://www.england.nhs.uk/publication/accessible-information-standard-specification/ Accessed 07th January, 2019.

McCall, W. A., & Crabbs, L. M. (1926). *Standard test lessons in reading.* Books II, III, IV, and V. New York: Bureau of Publications, Teachers College, Columbia University.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276–282.

McKee, R. A., & Miller, C. C. (2015). Institutionalizing Bayesianism within the organizational sciences: A practical guide featuring comments from eminent scholars. *Journal of Management, 41*(2), 471–490.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99,* 440–466.

McLaughlin, G. H. (1969). SMOG grading—A new readability formula. *Journal of Reading, 12,* 639–646.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55,* 51–62.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini, E. Albro, & R. T. O'Reilly (Eds.), *Measuring up* (pp. 89–119). Lanham, MA: R&L Education.

McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22,* 247–288.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). *Coh-Metrix version 3.0 indices.* Accessed 26 June 2019 at http://cohmetrix.com

McNamara, D. S., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In J. D. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 60–81). Mahwah, NJ: Erlbaum.

Meyer, B. J. F. (2003). Text coherence and readability. *Topics in Language Disorders, 23*, 204–224.

Miles, J. R., & Stine-Morrow, E. A. L. (2004). Adult age differences in self-regulated learning from reading sentences. *Psychology and Aging, 19*, 626–636.

Miller-Matero, L. R., Clark, K. B., Brescacin, C., Dubaybo, H., & Willens, D. E. (2016). Depression and literacy are important factors for missed appointments. *Psychology, Health & Medicine, 21*(6), 686-695.

Morrison, A. K., Schapira, M. M., Hoffmann, R. G., & Brousseau, D. C. (2014). Measuring health literacy in caregivers of children: A comparison of the newest vital sign and S-TOFHLA. *Clinical Pediatrics, 53*(13), 1264–1270.

Narvaez, D., van den Broek, P., & Ruiz, A. B. (1999). The influence of reading purpose on inference generation and comprehension in reading. *Journal of Educational Psychology, 91*, 488–496.

Nation, K. (2007). Children’s Reading Comprehension Difficulties. In M. J. Snowling and C. Hulme, (Eds.), *The Science of Reading: A Handbook* (p. 248-265). Oxford: Blackwell Publishing.

Nation, K., Clarke, P., Marshall, C.M., & Durand, M. (2004). Hidden language impairments in children: parallels between poor reading comprehension and specific language impairment. *Journal of Speech, Hearing and Language Research, 47*(1), 199-211.

Nation, K., Cocksey, J., Taylor, J. S. H., & Bishop, D. V. M. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry, 51*(9), 1031-1039.

Nation, K., & Snowling. M. J. (1997). Assessing reading difficulties: the validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67,* 359-370.

Nation, K., & Snowling. M. J. (1998). Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language, 39,* 85-101.

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading, 27,* 342-356.

Nelson, H. E. (1982). *The National Adult Reading Test (NART): test manual*. Windsor: NFER-Nelson.

Nenopoulou, S. (2005). *Dyslexia versus English-as-an-Additional Language: Literacy and Phonological Skillls* (Doctoral thesis, University of Surrey). Surrey Research Insight Open Access. http://epubs.surrey.ac.uk/709/1/fulltext.pdf

NHS Digital (2018a). Childhood Vaccination Coverage Statistics England, 2017-18. https://files.digital.nhs.uk/55/D9C4C2/child-vacc-stat-eng-2017-18-report.pdf

NHS Digital (2018b). Hospital Admitted Patient Care Activity, 2017-18. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2017-18#key-facts

NHS Digital (2017). Hospital Admitted Patient Care Activity, 2016-17. https://webarchive.nationalarchives.gov.uk/20180307201607/https://digital.nhs.uk/catalogue/PUB30098

NHS Digital (2016). Hospital Admitted Patient Care Activity, 2015-16. https://webarchive.nationalarchives.gov.uk/20180328130140/http://digital.nhs.uk/catalogue/PUB22378

NHS England. (2019). Missed GP appointments costing NHS millions. https://www.england.nhs.uk/2019/01/missed-gp-appointments-costing-nhs-millions/

NHS England. (2018a). Information Standard. https://www.england.nhs.uk/tis/about/the-info-standard/

NHS England. (2018b). NHS Identity Guidelines: Tone of voice. https://www.england.nhs.uk/nhsidentity/identity-guidelines/tone-of-voice/

NHS England. (2015). NHS Brand Guidelines: General Practitioner. Department of Health Branding Team.

NHS England. (2014). NHS Five Year Forward View. https://www.england.nhs.uk/publication/nhs-five-year-forward-view/

Oakhill, J. V. (1984). Inferential and memory skills in children’s comprehension of stories. *British Journal of Educational Psychology, 54,* 31-39.

Oakhill, J. V., Cain, K., & Elbro, C. (2014). What it’s all about. In *Understanding and Teaching Reading Comprehension* (pp. 1-10). Abingdon: Routledge.

Office for National Statistics (2016). *Population of the United Kingdom by Country of Birth and Nationality.* London: Office for National Statistics. Retrieved from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/datasets/populationoftheunitedkingdombycountryofbirthandnationality

Olin, J. T., & Zelinski, E. M. (1997). Age differences in calibration of comprehension. *Educational Gerontology, 23*, 67–77.

O’Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43,* 121-152.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9*(6). Retrieved from http://pareonline.net/getvn.asp?v=9&n=6

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2007). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19,* 228-242.

Paasche-Orlow, M. K., & Wolf, M. S. (2007). The causal pathways linking health literacy to health outcomes. *American Journal of Health Behavior, 31*(1), 19-26.

Paris, S. G., & Hamilton, E. E. (2009). The Development of Children's Reading Comprehension. In S. E. Israel & G. G. Duffey (Eds.) *Handbook of Research on Reading Comprehension* (pp. 32-53). Abingdon, Oxon: Routledge.

Patel, C. R., Cherla, D. V., Sanghvi, S., Baredes, S., & Eloy, J. A. (2013). Readability Assessment of Online Thyroid Surgery Patient Education Materials. *Head & Neck, 35*(10), 1421-1425.

Pavlenko, A. (2007). Autobiographical narratives as data in applied linguistics. *Applied Linguistics, 28*(2), 163-188.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A. (1991). Representations and Awareness in the Acquisition of Reading Competence. In Rieben, L., & Perfetti, C. A. (Ed.), *Learning To Read: Basic Research and Its Implications* (pp. 33-46). Hillsdale, NJ: Lawrence Erlbaum Associates.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*, 357–383.

Perfetti, C. A. (2010). Decoding, vocabulary, and comprehension: The golden triangle of reading skill. In M. G. McKeown & L. Kucan (Eds.), *Bringing reading researchers to life: Essays in honor of Isabel Beck* (pp. 291-303). New York: Guilford.

Perfetti, C. A., Landi, N., & Oakhill, J. V. (2007). The Acquisition of Reading Comprehension Skill. In M. J. Snowling and C. Hulme, (Eds.), *The Science of Reading: A Handbook* (p. 227-247). Oxford: Blackwell Publishing.

Perfetti, C., & Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading, 18*, 22-37.

Plain English Campaign. (2018). How to write in plain English. http://www.plainenglish.co.uk/files/howto.pdf

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Preston, C. C., & Colman, A. M. (2000). Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica, 104*, 1–15.

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

Ramscar, M., Sun, C. C., Hendrix, P., Baayen, H. (2017). The Mismeasurement of Mind: Life-Span Changes in Paired-Associate-Learning Scores Reflect the "Cost" of Learning, Not Cognitive Decline. *Psychological Science, 28*(8), 1171-1179.

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: insights from a multiple case study of dyslexic adults. *Brain: A Journal of Neurology, 126*, 841–65.

Ramus, F., & Szenkovits, G. (2008). What phonological deficit? *The Quarterly Journal of Experimental Psychology, 61*, 129–41.

Raney, G. E., Campbell, S. J., & Bovee, J. C. (2014). Using Eye Movements to Evaluate the Cognitive Processes Involved in Text Comprehension. *Journal of Visualized Experiments: JoVE*, (83), e50780. Advance online publication. http://doi.org/10.3791/50780

Rawson, K. A., & Dunlosky, J. (2002). Are Performance Predictions for Text Based on Ease of Processing? *Journal of Experimental Psychology, Learning, Memory, and Cognition, 28*(1), 69-80.

Rawson, K.A., Dunlosky, J., & Thiede, K.W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*, 1004–1010.

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye Movements During Mindless Reading. *Psychological Science, 21*(9), 1300-1310.

Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: explorations in experimental cognitive psychology. *Neuroscience, 139*(1), 5-21.

Riche, J. M., Reid, J. C., Robinson, R. D., & Kardash, C. A. M. (1991). Text and reader characteristics affecting the readability of patient literature. *Reading Improvement, 28*(4), 287-292.

Roberts, L., & Siyanova-Chanturia, A. (2013). Using Eye-tracking to Investigate Topics in L2 Acquisition and L2 Processing. *Studies in Second Language Acquisition, 35*, 213–235.

Rowlands, G., Protheroe, J., Winkley, J., Richardson, M., Seed, P. T., & Rudd, R. (2016). A mismatch between population health literacy and the complexity of health information: an observational study. *British Journal of General Practice, 65*(635), e379-e386. DOI: https://doi.org/10.3399/bjgp15X685285

Royal, K., & Dorman, D. (2018). Comparing Item Performance on Three- Versus Four-Option Multiple Choice Questions in a Veterinary Toxicology Course. *Veterinary Sciences, 5*(2), 55.

Sabatini, J. P., Sawaki, Y., Shore, J. R., & Scarborough, H. S. (2010). Relationships among reading skills of adults with low literacy. *Journal of Learning Disabilities, 43*(2), 122–138.

Scammacca, N., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities, 48*, 369–390.

Schillinger, D., Grumbach, K., Piette, J., Wang, F., Osmond, D., Daher, C., Palacios, J., Sullivan, G. D., & Bindman, A. B. (2002). Association of health literacy with diabetes outcomes. *Journal of American Medical Association, 288*, 475–482.

Schillinger, D., Piette, J., Grumbach, K., Wang, F., Wilson, C., Daher, C., Leong-Grotz, K., Castro, C., & Bindman, A. B. (2003). Closing the loop: physician communication with diabetic patients who have low health literacy. *Archives of internal medicine, 163*(1), 83-90.

Schnick, T., & Knickelbine, M. (2007). *Using the lexile analyzer for educators and media specialists.* Durham, NC: MetaMetrics, Inc.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47,* 609-612.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition Learning, 4,* 33–45.

Segers, E., & Verhoeven, L. (2016). How logical reasoning mediates the relation between lexical quality and reading comprehension. *Reading and Writing, 29,* 577–590.

Sharpe, K., & O'Carroll, R. (1991). Estimating premorbid intellectual level in dementia using the national adult reading test: a Canadian study. *British Journal of Clinical Psychology, 30*(4), 381–384.

Shipley, W. C., Gruber, C. P, Martin, T. A., & Klein, A. M. (2009). *Shipley-2 manual.* Los Angeles, CA: Western Psychological Services.

Silva, M. T., & Cain, K. (2015). The relations between lower- and higher-level oral language skills and their role in prediction of early reading comprehension. *Journal of Educational Psychology, 107*, 321-331.

Slater, B. A., Huang, Y., & Dalawari, P. (2017). The Impact of Teach-Back Method on Retention of Key Domains of Emergency Department Discharge Instructions. *The Journal of Emergency Medicine, 53*(5), 59-65.

Squiers, L., Peinado, S., Berkman, N., Boudewyns, V., & McCormack, L. (2012). The Health Literacy Skills Framework. *Journal of Health Communication, 17*(3), 30-54.

Stan Development Team. (2018). *Stan Modeling Language Users Guide and Reference Manual, Version 2.18.0.* http://mc-stan.org

Stanovich. K. E. (1988). Explaining the differences between the dyslexic and the garden-variety poor reader: The phonological-core variable-difference model. *Journal of Leaming Disabilities, 21*, 590-612.

Stine-Morrow, E. A. L., Miller, L. M. S., Gagne, D. D., & Hertzog, C. (2008). Self-regulated reading in adulthood. *Psychology and Aging, 23*, 131–153.

Stone, J. M., & Towse, J. N. (2015). A Working Memory Test Battery: Java-Based Collection of Seven Working Memory Tasks. *Journal of Open Research Software, 3*(1), 1-9.

Stoye, G. (2017). UK health spending: Briefing note. *Institute for Fiscal Studies,* IFS Briefing Note BN201.

Street, J. A. (2020). More lexically-specific knowledge and individual differences in adult native speakers' processing of the English passive. *Language Sciences, 78*, 101254.

Street, J. A., & Dąbrowska, E. (2014). Lexically specific knowledge and individual differences in adult native speakers' processing of the English passive. *Applied Psycholinguistics, 35*(1), 97-118.

Taylor, T. H. (2010). Chapter: Ceiling Effect. In Salkind, N. J. (Ed.), *Encyclopaedia of Research Design* (pp. 133-134). Thousand Oaks: SAGE Publications, Inc.

Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28,* 129–160.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of text. *Journal of Educational Psychology, 95*, 66–73.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor Metacomprehension Accuracy as a Result of Inappropriate Cue Use. *Discourse Processes, 47*(4), 331-362.

Thomson, M. D., & Hoffman-Goetz, L. (2010). Cancer Information Comprehension by English-as-a-Second-Language Immigrant Women. *Journal of Health Communication, 16*(1), 17-33

Thorndike, E. L., & McCall, W. A. (1921). *Thorndike-McCall reading scales for Grades 2-12*. New York: Bureau of Publications, Teachers College, Columbia University

Todd, L., & Hoffman-Goetz, L. (2011). Predicting health literacy among English-as-a-second-Language older Chinese immigrant women to Canada: comprehension of colon cancer prevention information. *Journal of Cancer Education, 26*(2), 326-332.

Tokowicz, N., Michael, E. B., & Kroll, J. F. (2004). The roles of study-abroad experience and working memory capacity in the types of errors made during translation. *Bilingualism: Language and Cognition, 7*, 255–272.

Tunmer, W. E., & Chapman, J. W. (2012). The Simple View of Reading Redux: Vocabulary Knowledge and the Independent Components Hypothesis. *Journal of Learning Disabilities, 45*(5), 453-466.

Tunmer, W. E., & Chapman, J. W. (2011). Does set for variability mediate the influence of vocabulary knowledge on the development of word recognition skills? *Scientific Studies of Reading, 16*(2), 122-140.

Upton, T., & Lee-Thompson, L. (2001). The role of the first language in second language reading. *Studies in Second Language Acquisition, 23*, 469-495.

U.S. Department of Health and Human Services. (2010). *Healthy people 2010. Understanding and improving health and objectives for improving health* (2nd ed.). Washington, DC: U.S. Government Printing Office.

Usher, M., Haarmann, H., Cohen, J. D., & Horn, D. (2001). Neural mechanisms for the magical number 4: competitive interactions and non-linear oscillations. *Behavioral and Brain Sciences, 24,* 151-152.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217-239.

van den Broek, P., Bohn-Gettler, C., Kendeou, P., Carlson, S., & White, M. J. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension. In M.T. McCrudden, J. Magliano, & G. Schraw (Eds.), *Relevance instructions and goal-focusing in text learning* (pp. 123–140). Greenwich, CT: Information Age Publishing.

van den Broek, P., Helder, A. (2017). Cognitive Processes in Discourse Comprehension: Passive Processes, Reader-Initiated Processes and Evolving Mental Representastions. *Discourse Porcesses, 54*(5-6), 360-372.

van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081–1087.

van den Broek, P., Rapp, D., & Kendeou, P. (2005). Integrating memory based and constructionist processes in accounts of reading comprehension. *Discourse Processes, 39*(2), 299–316.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension.* Academic Press, Inc: New York.

Van Dyke, J.A., & Johns, C.L., Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition, 131*, 373-403.

van Heuven, W. J. B., Dijkstra, A. T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language, 39,* 458-483.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review, 25*(1), 143-154.

Vaughn, S., Wanzek, J., Wexler, J., Barth, A., Cirino, P. T., Fletcher, J.M., Romain, M.A., Denton, C.A., Roberts, G., & Francis, D. J. (2010). The relative effects of group size on Reading progress of older students with reading difficulties. *Reading and Writing: An Interdisciplinary Journal, 23*(8), 931-956.

Vehtari, A., Gelman, A., & Gabry, J. (2017a). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing. 27*(5), 1413--1432. doi:10.1007/s11222-016-9696-4.

Vehtari, A., Gelman, A., & Gabry, J. (2017b). Pareto smoothed importance sampling. ArXiv preprint: http://arxiv.org/abs/1507.02646/

Vellutino, F. R., Tunmer, W. E., Jaccard, J., & Chen, S. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*, 3–32.

Verhaeghen, P. (2003). Aging and vocabulary scores: A meta-analysis. *Psychology and Aging, 18,* 332–339.

Vidal-Abarca, E., Martínez, G., & Gilabert, R. (2000). Two procedures to improve instructional text: Effects on memory and learning. *Journal of Educational Psychology, 92*, 107–116.

von Bastian, C. C., Locher, A., & Ruffin, M. (2013). Tatool: A java-based open-source programming framework for psychological studies. *Behavior research methods, 45*(1), 108–115.

Von Wagner, C., Knight, K., Steptoe, A., & Wardle, J. (2007). Functional health literacy and health-promoting behaviour in a national sample of British adults. *Journal of Epidemiology and Community Health. 61*(12), 1086-1090.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin 101*, 192–212.

Walton, D., & Brooks, P. (1995). The Spoonerism Test. *Educational and Child Psychology, 12*(1), 50–52.

Wang, L.-W., Miller, M. J., Schmitt, M. R., & Wen, F. K. (2013). Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Research in Social & Administrative Pharmacy, 9*(5), 503–516.

Weir, C., & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension. *Cambridge ESOL: Research Notes, 31*, 2–10.

Weiss, B. D., & Palmer, R. (2004). Relationship between health care costs and very low literacy skills in a medically needy and indigent Medicaid population. *Journal of the American Board of Family Practitioners, 17*, 44-47.

White, S., Milne, E., Rosen, S., Hansen, P., Swettenham, J., Frith, U., & Ramus, F. (2006). The role of sensorimotor impairments in dyslexia: a multiple case study of dyslexic children. *Developmental Science, 9*, 237–55.

Wittrock, M. C., & Alesandrini, K. (1990). Generation of summaries and analogies and analytic and holistic abilities. *American Educational Research Journal, 27*, 489-502.

World Health Organization (2018). Europe Observes a 4 Fold Increase in Measles Cases in 2017 Compared to a Previous Year. http://www.euro.who.int/en/media-centre/sections/press-releases/2018/europe-observes-a-4-fold-increase-in-measles-cases-in-2017-compared-to-previous-year

Yang E. (2017). Bilinguals' Working Memory (WM) Advantage and Their Dual Language Practices. *Brain sciences, 7*(7), 86. https://doi.org/10.3390/brainsci7070086

Yang, C-L., Perfetti, C. A., & Schmalhofer, F. (2005). Less skilled comprehenders' ERPs show sluggish word-to-text integration processes. *Written Language & Literacy, 8*, 233-257.

Yang, C-L., Perfetti, C. A., & Schmalhofer, F. (2007). Event-related potential indicators of text integration across sentence boundaries. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 55–89.

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 53–79.

Yeari, M., van den Broek, P., & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: Evidence from eye-movements. *Reading and Writing, 28*, 1071–1097.

Zabrucky, K., Moore, D., & Agler, L-M. L. (2012). Metacomprehension across the Lifespan: influence of education and instructional support. *Procedia - Social and Behavioral Sciences, 46,* 601 – 604.

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review, 23,* 1028–1034.

Zwaan, R.A., & Radvansky, G.A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.

Appendices

Appendix A: Chapter 5 Tables and Figures

Table 5.7. Sensitivity Checks (FRE).

Model number	Likelihood function	Priors	Probable effects (sign)	Highest Rhat	LOOIC	LOO R2	Chains	Notes
1	Skew-normal	Predictors(Normal(0, 10)), Intercept(Normal(50, 50)),	Referential cohesion (+), Sentence length (-), LSA (-), Logical connectives (+), Gerunds (-)	1.00	577.48	.25	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth. One observation (47) is labelled as problematic.
2	Gaussian (Normal)	Predictors(Normal(0, 10)), Intercept(Normal(50, 50)),	Referential cohesion (+), Sentence length (-), LSA (-), Logical connectives (+), Gerunds (-)	1.00	574.09	.24	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
1.1 (without observation 47)	Skew-normal	Predictors(Normal(0, 10)), Intercept(Normal(50, 50)),	Referential cohesion (+), Sentence length (-), Passive voice (-), LSA (-), Logical connectives (+), Gerunds (-)	1.00	551.58	.27	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth. Four additional observations are labelled as problematic.
2.1* (without observation 47)	Gaussian (Normal)	Predictors(Normal(0, 10)), Intercept(Normal(50, 50)),	Referential cohesion (+), Sentence length (-), Passive voice (-), LSA (-), Logical connectives (+), Gerunds (-)	1.00	551.15	.26	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
2.2** (without observation 47)	Gaussian (Normal)	Predictors(Normal(0, 10)), Intercept(Normal(50, 50)),	Referential cohesion (+), Sentence length (-), Passive voice (-), LSA (-), Logical connectives (+), Gerunds (-)	1.00	551.16	.26	6 chains with 8000 iterations per chain	No transitions beyond maximum treedepth.

Notes. *The chosen model. **Local convergence check of the chosen model.

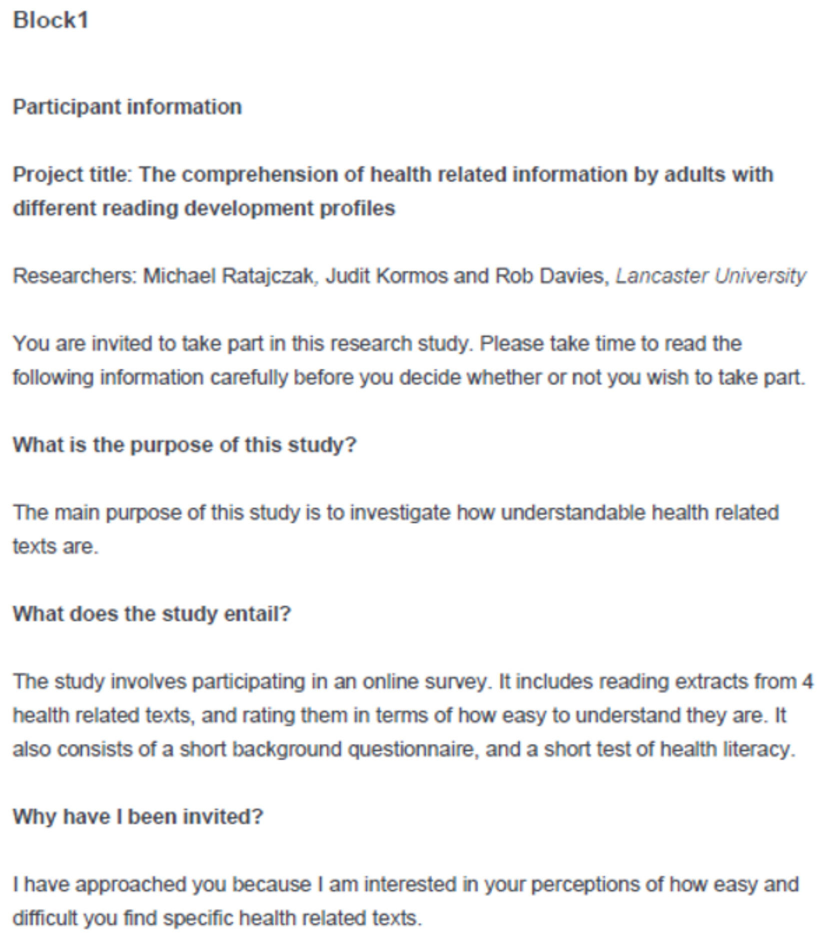
Table 5.8. Sensitivity Checks (RDL2).

Model number	Likelihood function	Priors	Probable effects (sign)	Highest Rhat	LOOIC	LOO R2	Chains	Notes
1	Skew-normal	Predictors(Normal(0, 10)), Intercept(Normal(15, 15)),	Referential cohesion (+), Word frequency (+), Syntax similarity (+), Causal cohesion (+), Hypernymy noun (-)	1.00	387.80	.28	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth. One observation (72) is labelled as problematic.
2	Gaussian (Normal)	Predictors(Normal(0, 10)), Intercept(Normal(15, 15)),	Referential cohesion (+), Word frequency (+), Syntax similarity (+), Causal cohesion (+), Hypernymy noun (-)	1.00	384.66	.30	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth. One observation (72) is labelled as problematic.
1.1 (without observation 72)	Skew-normal	Predictors(Normal(0, 10)), Intercept(Normal(15, 15)),	Referential cohesion (+), Word frequency (+), Syntax similarity (+), Hypernymy noun (-)	1.00	364.67	.45	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth. One additional observation is labelled as problematic.
2.1* (without observation 72)	Gaussian (Normal)	Predictors(Normal(0, 10)), Intercept(Normal(15, 15)),	Referential cohesion (+), Word frequency (+), Passive voice (-), Syntax similarity (+), Hypernymy noun (-)	1.00	362.96	.45	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
2.2** (without observation 72)	Gaussian (Normal)	Predictors(Normal(0, 10)), Intercept(Normal(15, 15)),	Referential cohesion (+), Word frequency (+), Passive voice (-), Syntax similarity (+), Hypernymy noun (-)	1.00	362.74	.45	6 chains with 8000 iterations per chain	No transitions beyond maximum treedepth.

Notes. *The chosen model. **Local convergence check of the chosen model.

Appendix B: Chapter 6 Online Survey

Page 1



Block1

Participant information

Project title: The comprehension of health related information by adults with different reading development profiles

Researchers: Michael Ratajczak, Judit Kormos and Rob Davies, *Lancaster University*

You are invited to take part in this research study. Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the purpose of this study?

The main purpose of this study is to investigate how understandable health related texts are.

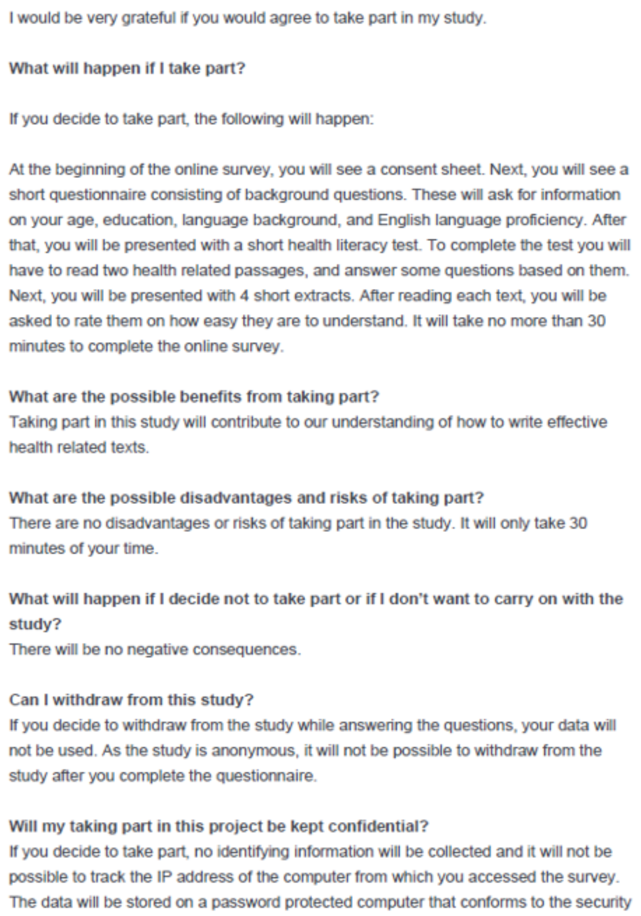
What does the study entail?

The study involves participating in an online survey. It includes reading extracts from 4 health related texts, and rating them in terms of how easy to understand they are. It also consists of a short background questionnaire, and a short test of health literacy.

Why have I been invited?

I have approached you because I am interested in your perceptions of how easy and difficult you find specific health related texts.

Page 2



I would be very grateful if you would agree to take part in my study.

What will happen if I take part?

If you decide to take part, the following will happen:

At the beginning of the online survey, you will see a consent sheet. Next, you will see a short questionnaire consisting of background questions. These will ask for information on your age, education, language background, and English language proficiency. After that, you will be presented with a short health literacy test. To complete the test you will have to read two health related passages, and answer some questions based on them. Next, you will be presented with 4 short extracts. After reading each text, you will be asked to rate them on how easy they are to understand. It will take no more than 30 minutes to complete the online survey.

What are the possible benefits from taking part?
Taking part in this study will contribute to our understanding of how to write effective health related texts.

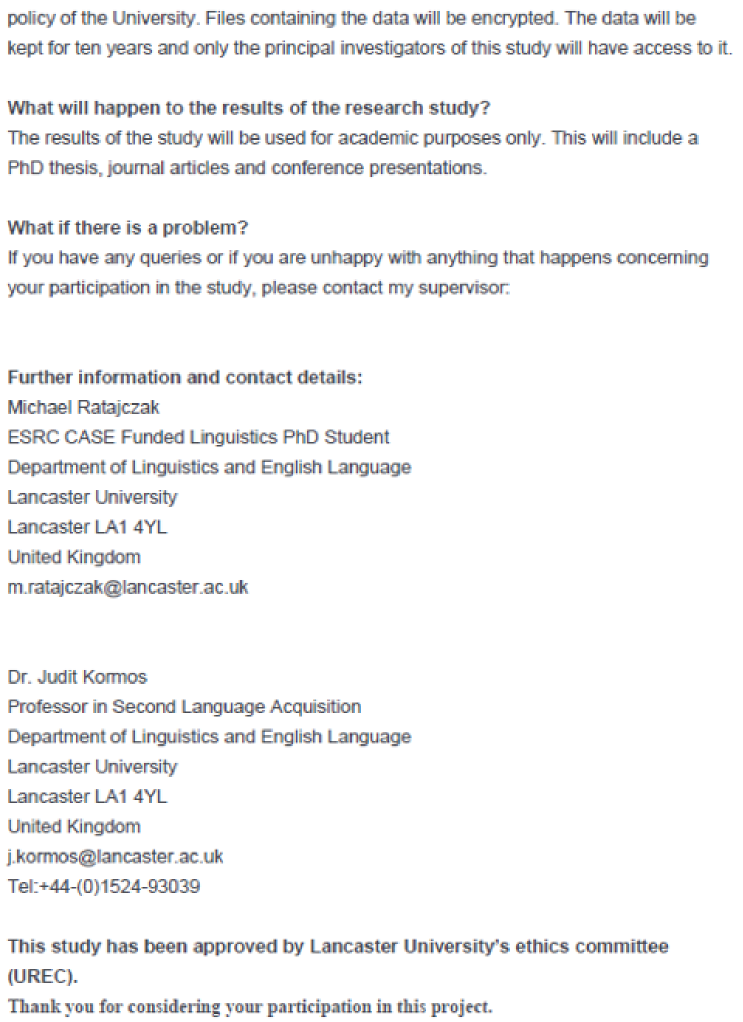
What are the possible disadvantages and risks of taking part?
There are no disadvantages or risks of taking part in the study. It will only take 30 minutes of your time.

What will happen if I decide not to take part or if I don't want to carry on with the study?
There will be no negative consequences.

Can I withdraw from this study?
If you decide to withdraw from the study while answering the questions, your data will not be used. As the study is anonymous, it will not be possible to withdraw from the study after you complete the questionnaire.

Will my taking part in this project be kept confidential?
If you decide to take part, no identifying information will be collected and it will not be possible to track the IP address of the computer from which you accessed the survey. The data will be stored on a password protected computer that conforms to the security

Page 3



policy of the University. Files containing the data will be encrypted. The data will be kept for ten years and only the principal investigators of this study will have access to it.

What will happen to the results of the research study?
The results of the study will be used for academic purposes only. This will include a PhD thesis, journal articles and conference presentations.

What if there is a problem?
If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact my supervisor:

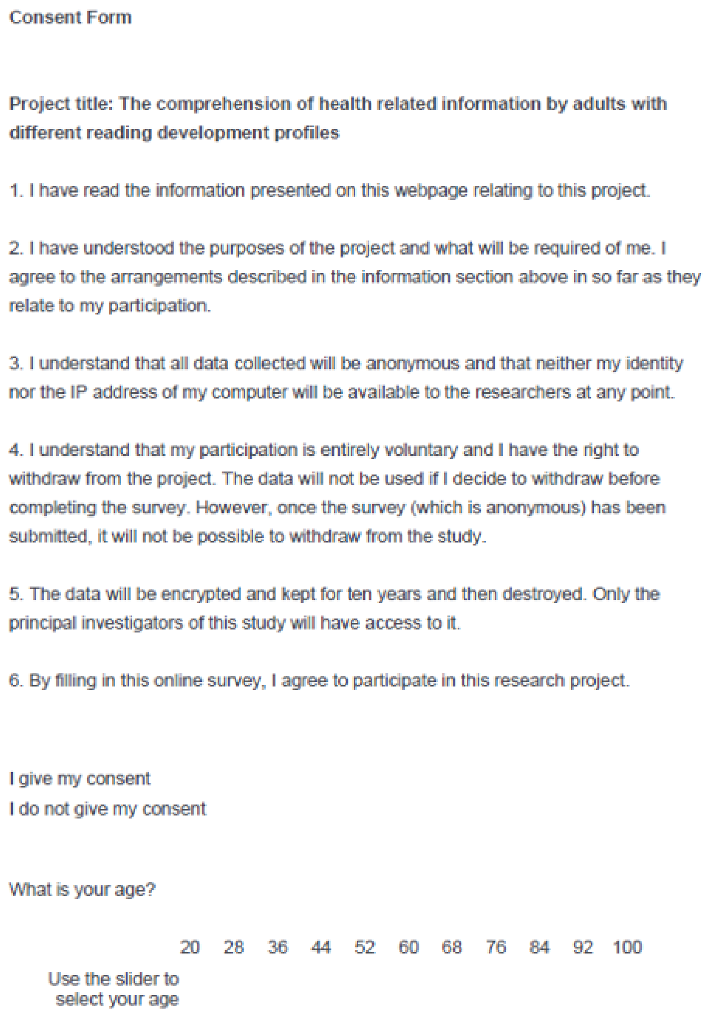
Further information and contact details:
Michael Ratajczak
ESRC CASE Funded Linguistics PhD Student
Department of Linguistics and English Language
Lancaster University
Lancaster LA1 4YL
United Kingdom
m.ratajczak@lancaster.ac.uk

Dr. Judit Kormos
Professor in Second Language Acquisition
Department of Linguistics and English Language
Lancaster University
Lancaster LA1 4YL
United Kingdom
j.kormos@lancaster.ac.uk
Tel:+44-(0)1524-93039

This study has been approved by Lancaster University's ethics committee (UREC).

Thank you for considering your participation in this project.

Page 4



Consent Form

Project title: The comprehension of health related information by adults with different reading development profiles

1. I have read the information presented on this webpage relating to this project.

2. I have understood the purposes of the project and what will be required of me. I agree to the arrangements described in the information section above in so far as they relate to my participation.

3. I understand that all data collected will be anonymous and that neither my identity nor the IP address of my computer will be available to the researchers at any point.

4. I understand that my participation is entirely voluntary and I have the right to withdraw from the project. The data will not be used if I decide to withdraw before completing the survey. However, once the survey (which is anonymous) has been submitted, it will not be possible to withdraw from the study.

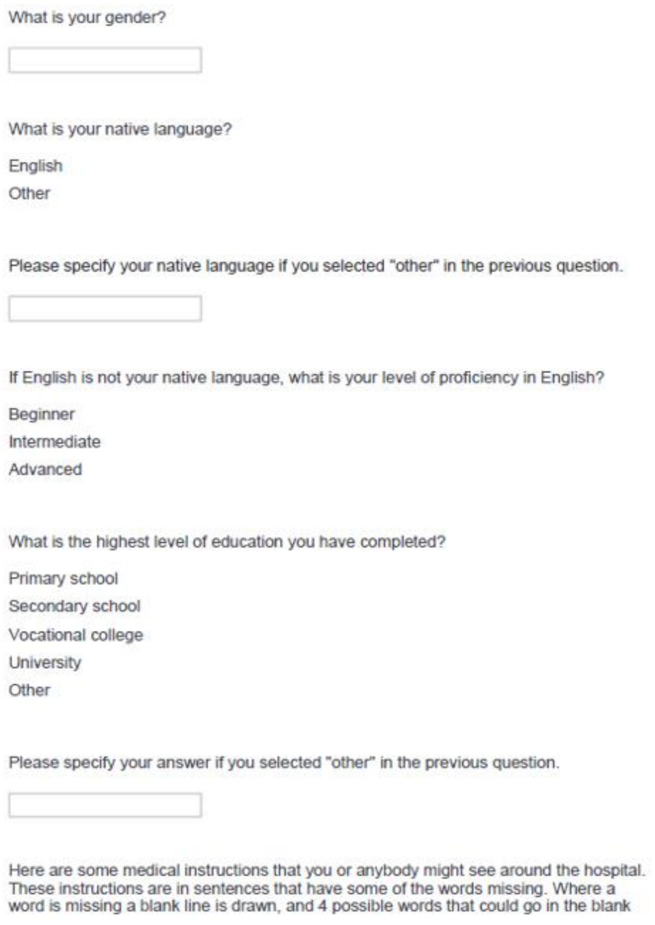
5. The data will be encrypted and kept for ten years and then destroyed. Only the principal investigators of this study will have access to it.

6. By filling in this online survey, I agree to participate in this research project.

I give my consent
I do not give my consent

What is your age?

Page 5



What is your gender?

□

What is your native language?

English

Other

Please specify your native language if you selected "other" in the previous question.

□

If English is not your native language, what is your level of proficiency in English?

Beginner

Intermediate

Advanced

What is the highest level of education you have completed?

Primary school

Secondary school

Vocational college

University

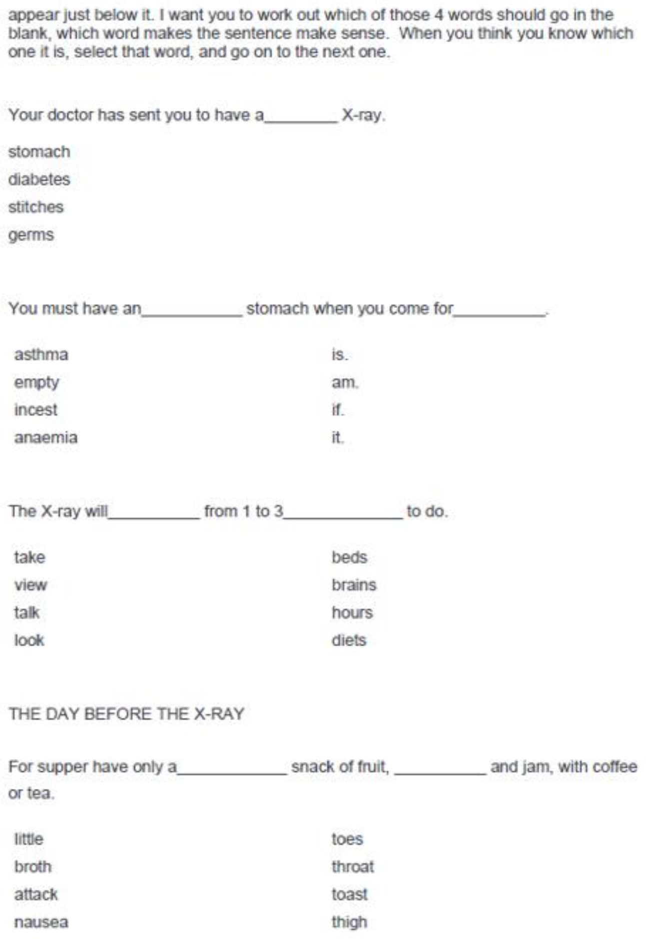
Other

Please specify your answer if you selected "other" in the previous question.

□

Here are some medical instructions that you or anybody might see around the hospital. These instructions are in sentences that have some of the words missing. Where a word is missing a blank line is drawn, and 4 possible words that could go in the blank

Page 6



appear just below it. I want you to work out which of those 4 words should go in the blank, which word makes the sentence make sense. When you think you know which one it is, select that word, and go on to the next one.

Your doctor has sent you to have a________ X-ray.

stomach

diabetes

stitches

germs

You must have an___________ stomach when you come for__________.

asthma	is.
empty	am.
incest	if.
anaemia	it.

The X-ray will__________ from 1 to 3_____________ to do.

take	beds
view	brains
talk	hours
look	diets

THE DAY BEFORE THE X-RAY

For supper have only a____________ snack of fruit, __________ and jam, with coffee or tea.

little	toes
broth	throat
attack	toast
nausea	thigh

Page 7

After____________, you must not_________ or drink

minute
midnight
during
before

easy
ate
drank
eat

anything at_________ until after you have___________ the X-ray.

ill
all
each
any

are
has
had
was

THE DAY OF THE X-RAY

Do not eat______________.

appointment.
walk-in.
breakfast.
clinic.

Do not__________, even________________.

drive
drink
dress
dose

heart.
breath.
water.
cancer.

If you have any____________, call the X-ray____________ on 01616450022.

answers

Department

Page 8

exercises
tracts
questions

Sprain
Pharmacy
Toothache

It has been explained to___________ that during the course of the

my
me
he
she

______________ or procedure, unforeseen conditions may be_______________

syphilis
hepatitis
colitis
operation

revealed
depressed
directed
notified

that necessitate an extension of the_____________ procedure(s) or

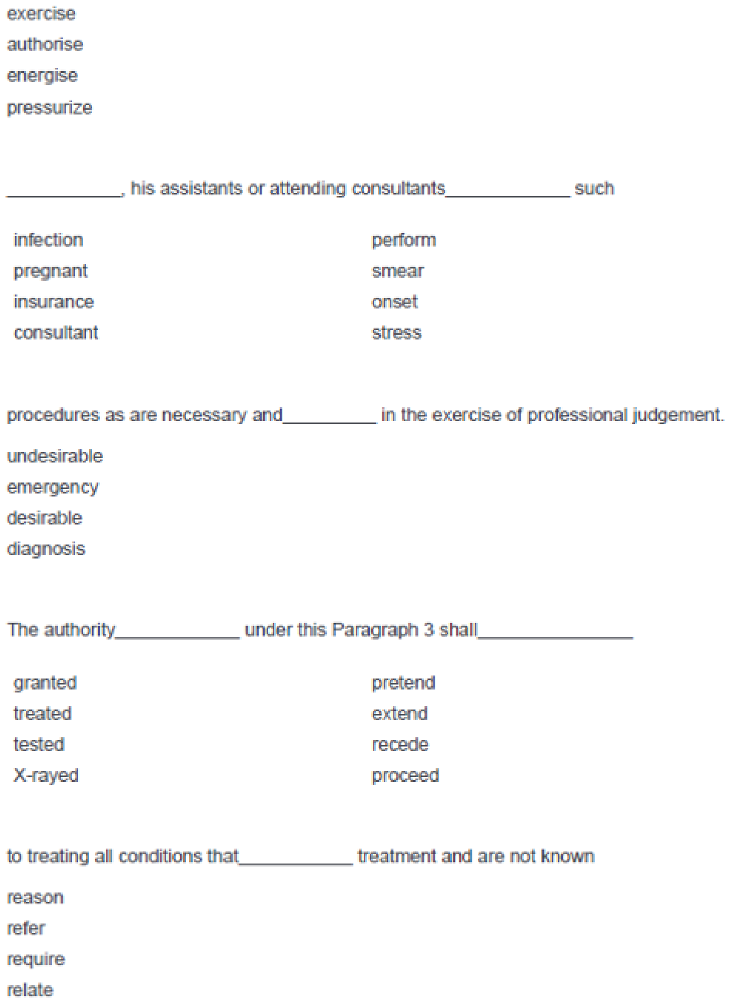
appendix
another
original
operation

different procedure(s) than those____________ forth in paragraph 2.

get
set
see
go

I, therefore,_____________ and request that the above named

Page 9



exercise
authorise
energise
pressurize

___________, his assistants or attending consultants____________ such

infection
pregnant
insurance
consultant

perform
smear
onset
stress

procedures as are necessary and_________ in the exercise of professional judgement.

undesirable
emergency
desirable
diagnosis

The authority____________ under this Paragraph 3 shall_______________

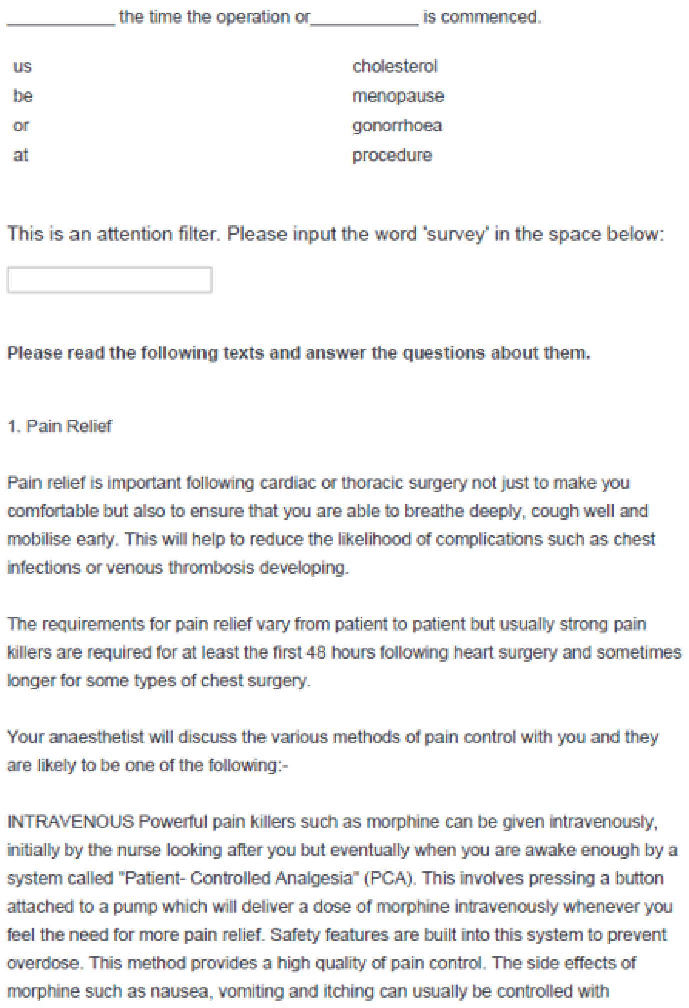
granted
treated
tested
X-rayed

pretend
extend
recede
proceed

to treating all conditions that___________ treatment and are not known

reason
refer
require
relate

Page 10



__________ the time the operation or__________ is commenced.

us
be
or
at

cholesterol
menopause
gonorrhoea
procedure

This is an attention filter. Please input the word 'survey' in the space below:

Please read the following texts and answer the questions about them.

1. Pain Relief

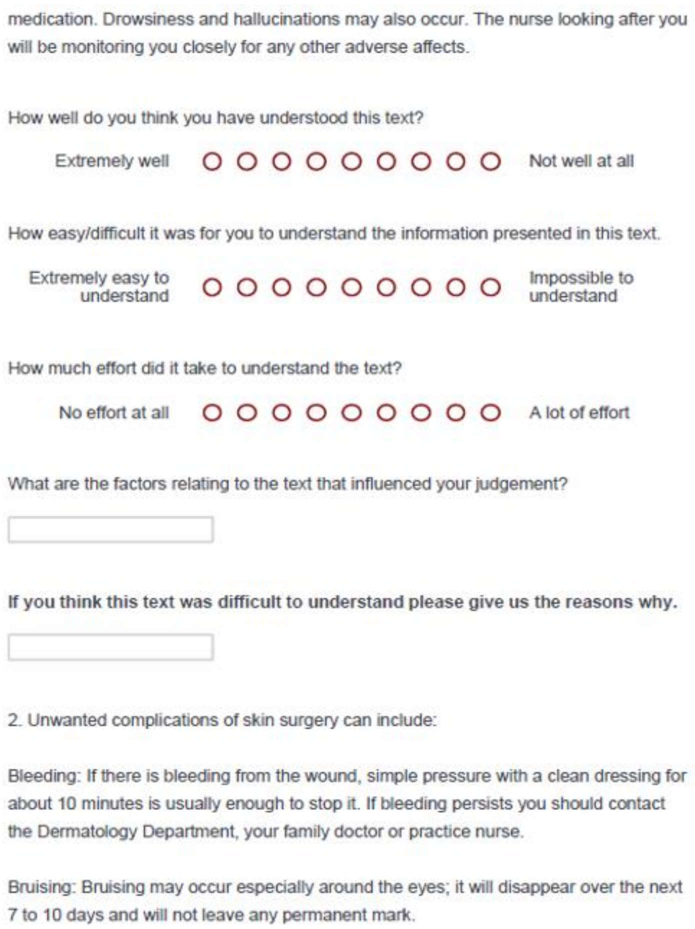
Pain relief is important following cardiac or thoracic surgery not just to make you comfortable but also to ensure that you are able to breathe deeply, cough well and mobilise early. This will help to reduce the likelihood of complications such as chest infections or venous thrombosis developing.

The requirements for pain relief vary from patient to patient but usually strong pain killers are required for at least the first 48 hours following heart surgery and sometimes longer for some types of chest surgery.

Your anaesthetist will discuss the various methods of pain control with you and they are likely to be one of the following:-

INTRAVENOUS Powerful pain killers such as morphine can be given intravenously, initially by the nurse looking after you but eventually when you are awake enough by a system called "Patient- Controlled Analgesia" (PCA). This involves pressing a button attached to a pump which will deliver a dose of morphine intravenously whenever you feel the need for more pain relief. Safety features are built into this system to prevent overdose. This method provides a high quality of pain control. The side effects of morphine such as nausea, vomiting and itching can usually be controlled with

Page 11



medication. Drowsiness and hallucinations may also occur. The nurse looking after you will be monitoring you closely for any other adverse affects.

How well do you think you have understood this text?

Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How easy/difficult it was for you to understand the information presented in this text.

Extremely easy to understand ○ ○ ○ ○ ○ ○ ○ ○ ○ Impossible to understand

How much effort did it take to understand the text?

No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

What are the factors relating to the text that influenced your judgement?

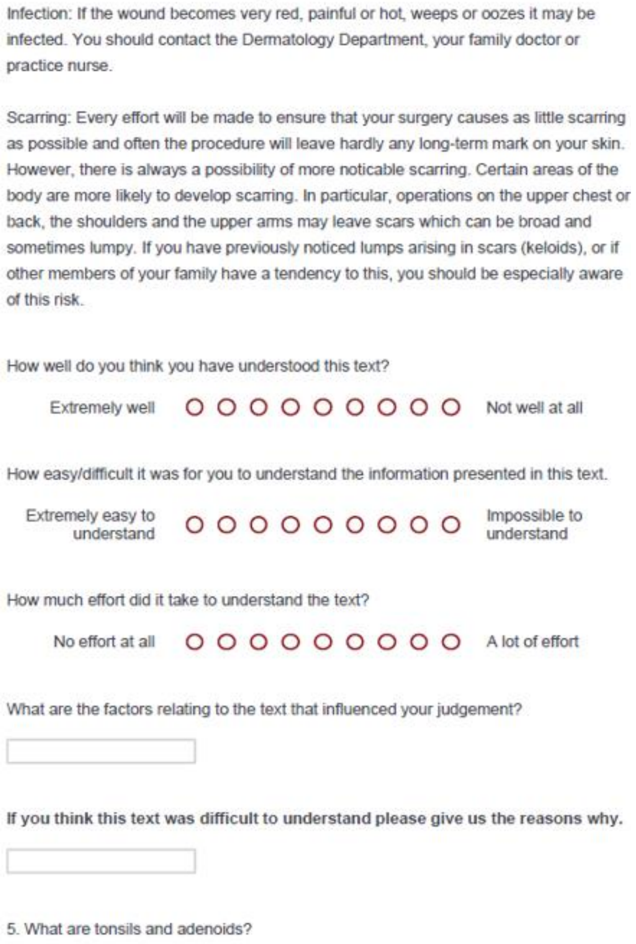
If you think this text was difficult to understand please give us the reasons why.

2. Unwanted complications of skin surgery can include:

Bleeding: If there is bleeding from the wound, simple pressure with a clean dressing for about 10 minutes is usually enough to stop it. If bleeding persists you should contact the Dermatology Department, your family doctor or practice nurse.

Bruising: Bruising may occur especially around the eyes; it will disappear over the next 7 to 10 days and will not leave any permanent mark.

Page 12



Infection: If the wound becomes very red, painful or hot, weeps or oozes it may be infected. You should contact the Dermatology Department, your family doctor or practice nurse.

Scarring: Every effort will be made to ensure that your surgery causes as little scarring as possible and often the procedure will leave hardly any long-term mark on your skin. However, there is always a possibility of more noticable scarring. Certain areas of the body are more likely to develop scarring. In particular, operations on the upper chest or back, the shoulders and the upper arms may leave scars which can be broad and sometimes lumpy. If you have previously noticed lumps arising in scars (keloids), or if other members of your family have a tendency to this, you should be especially aware of this risk.

How well do you think you have understood this text?

Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How easy/difficult it was for you to understand the information presented in this text.

Extremely easy to understand ○ ○ ○ ○ ○ ○ ○ ○ ○ Impossible to understand

How much effort did it take to understand the text?

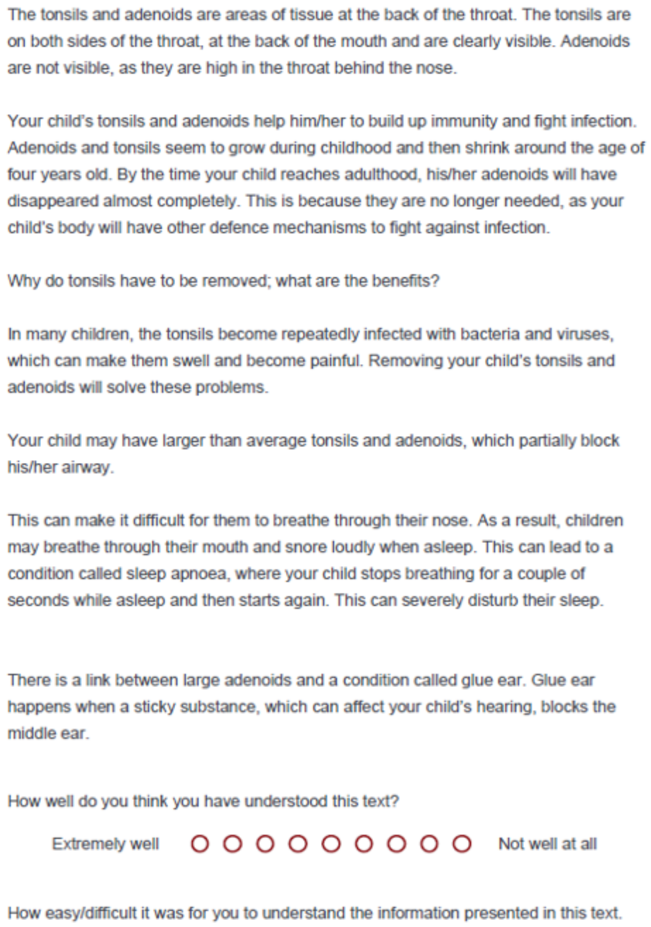
No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

What are the factors relating to the text that influenced your judgement?

If you think this text was difficult to understand please give us the reasons why.

5. What are tonsils and adenoids?

Page 13



The tonsils and adenoids are areas of tissue at the back of the throat. The tonsils are on both sides of the throat, at the back of the mouth and are clearly visible. Adenoids are not visible, as they are high in the throat behind the nose.

Your child's tonsils and adenoids help him/her to build up immunity and fight infection. Adenoids and tonsils seem to grow during childhood and then shrink around the age of four years old. By the time your child reaches adulthood, his/her adenoids will have disappeared almost completely. This is because they are no longer needed, as your child's body will have other defence mechanisms to fight against infection.

Why do tonsils have to be removed; what are the benefits?

In many children, the tonsils become repeatedly infected with bacteria and viruses, which can make them swell and become painful. Removing your child's tonsils and adenoids will solve these problems.

Your child may have larger than average tonsils and adenoids, which partially block his/her airway.

This can make it difficult for them to breathe through their nose. As a result, children may breathe through their mouth and snore loudly when asleep. This can lead to a condition called sleep apnoea, where your child stops breathing for a couple of seconds while asleep and then starts again. This can severely disturb their sleep.

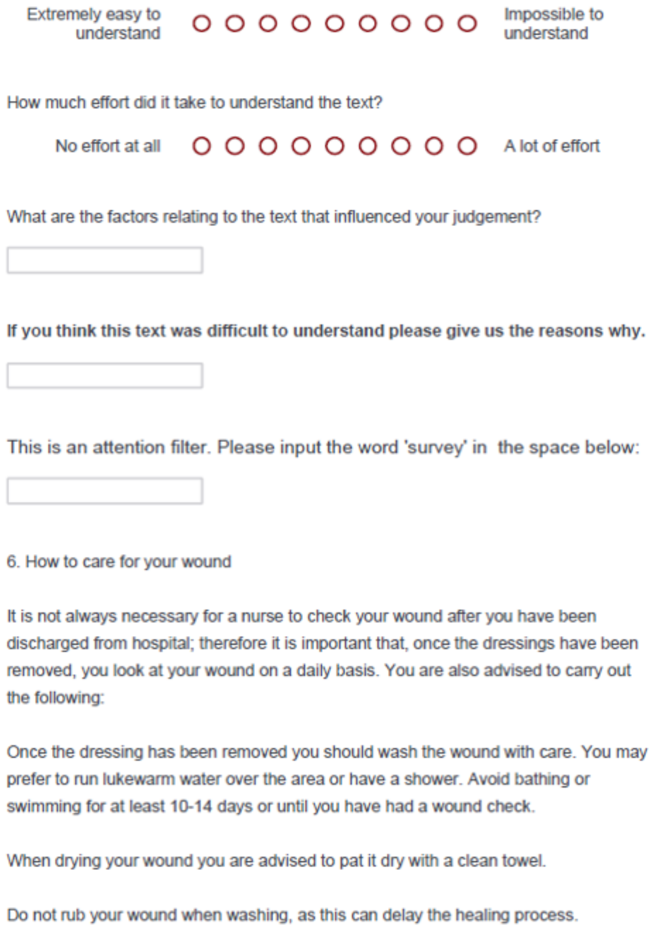
There is a link between large adenoids and a condition called glue ear. Glue ear happens when a sticky substance, which can affect your child's hearing, blocks the middle ear.

How well do you think you have understood this text?

Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How easy/difficult it was for you to understand the information presented in this text.

Page 14



Extremely easy to understand ○ ○ ○ ○ ○ ○ ○ ○ ○ Impossible to understand

How much effort did it take to understand the text?

No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

What are the factors relating to the text that influenced your judgement?

□

If you think this text was difficult to understand please give us the reasons why.

□

This is an attention filter. Please input the word 'survey' in the space below:

□

6. How to care for your wound

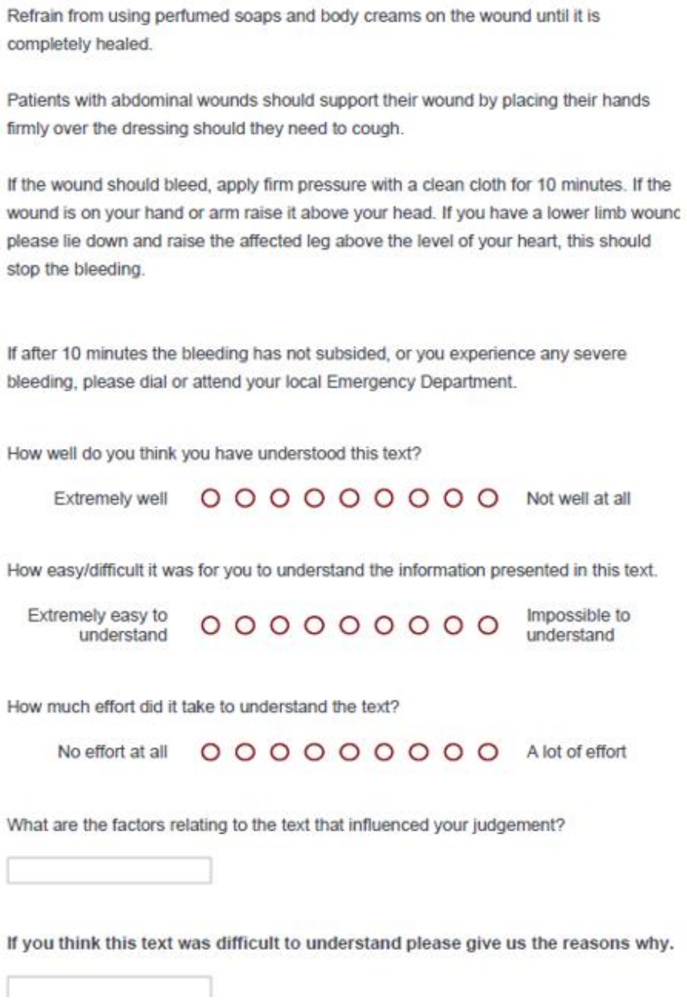
It is not always necessary for a nurse to check your wound after you have been discharged from hospital; therefore it is important that, once the dressings have been removed, you look at your wound on a daily basis. You are also advised to carry out the following:

Once the dressing has been removed you should wash the wound with care. You may prefer to run lukewarm water over the area or have a shower. Avoid bathing or swimming for at least 10-14 days or until you have had a wound check.

When drying your wound you are advised to pat it dry with a clean towel.

Do not rub your wound when washing, as this can delay the healing process.

Page 15



Refrain from using perfumed soaps and body creams on the wound until it is completely healed.

Patients with abdominal wounds should support their wound by placing their hands firmly over the dressing should they need to cough.

If the wound should bleed, apply firm pressure with a clean cloth for 10 minutes. If the wound is on your hand or arm raise it above your head. If you have a lower limb wounc please lie down and raise the affected leg above the level of your heart, this should stop the bleeding.

If after 10 minutes the bleeding has not subsided, or you experience any severe bleeding, please dial or attend your local Emergency Department.

How well do you think you have understood this text?

Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How easy/difficult it was for you to understand the information presented in this text.

Extremely easy to understand ○ ○ ○ ○ ○ ○ ○ ○ ○ Impossible to understand

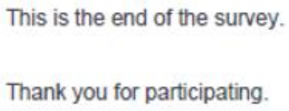
How much effort did it take to understand the text?

No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

What are the factors relating to the text that influenced your judgement?

If you think this text was difficult to understand please give us the reasons why.

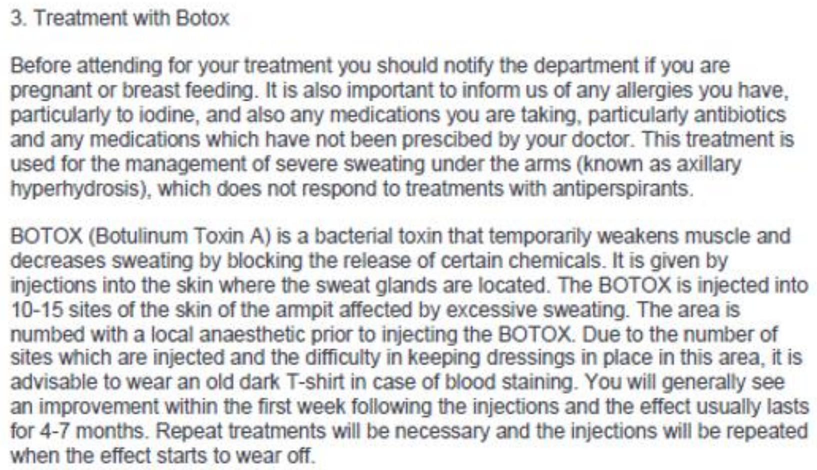
Page 16



This is the end of the survey.

Thank you for participating.

Page 17



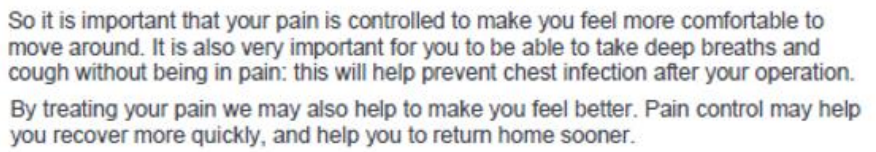
3. Treatment with Botox

Before attending for your treatment you should notify the department if you are pregnant or breast feeding. It is also important to inform us of any allergies you have, particularly to iodine, and also any medications you are taking, particularly antibiotics and any medications which have not been prescibed by your doctor. This treatment is used for the management of severe sweating under the arms (known as axillary hyperhydrosis), which does not respond to treatments with antiperspirants.

BOTOX (Botulinum Toxin A) is a bacterial toxin that temporarily weakens muscle and decreases sweating by blocking the release of certain chemicals. It is given by injections into the skin where the sweat glands are located. The BOTOX is injected into 10-15 sites of the skin of the armpit affected by excessive sweating. The area is numbed with a local anaesthetic prior to injecting the BOTOX. Due to the number of sites which are injected and the difficulty in keeping dressings in place in this area, it is advisable to wear an old dark T-shirt in case of blood staining. You will generally see an improvement within the first week following the injections and the effect usually lasts for 4-7 months. Repeat treatments will be necessary and the injections will be repeated when the effect starts to wear off.

Following treatment some patients have felt that the sweating in other parts of the body increased. Allergic or inflammatory reactions in the area are rarely observed, however, bruising may occur initially. Numbness of the skin is a recognised side effect of BOTOX but it is not normally noticeable in the armpits. Other side effects are very uncommon.

Page 18



4. As part of your treatment, your doctors may decide an operation is necessary. Even with modern advances in technique, operations can often be painful. However, with your help our staff will try to keep you comfortable.

Why is pain control important?

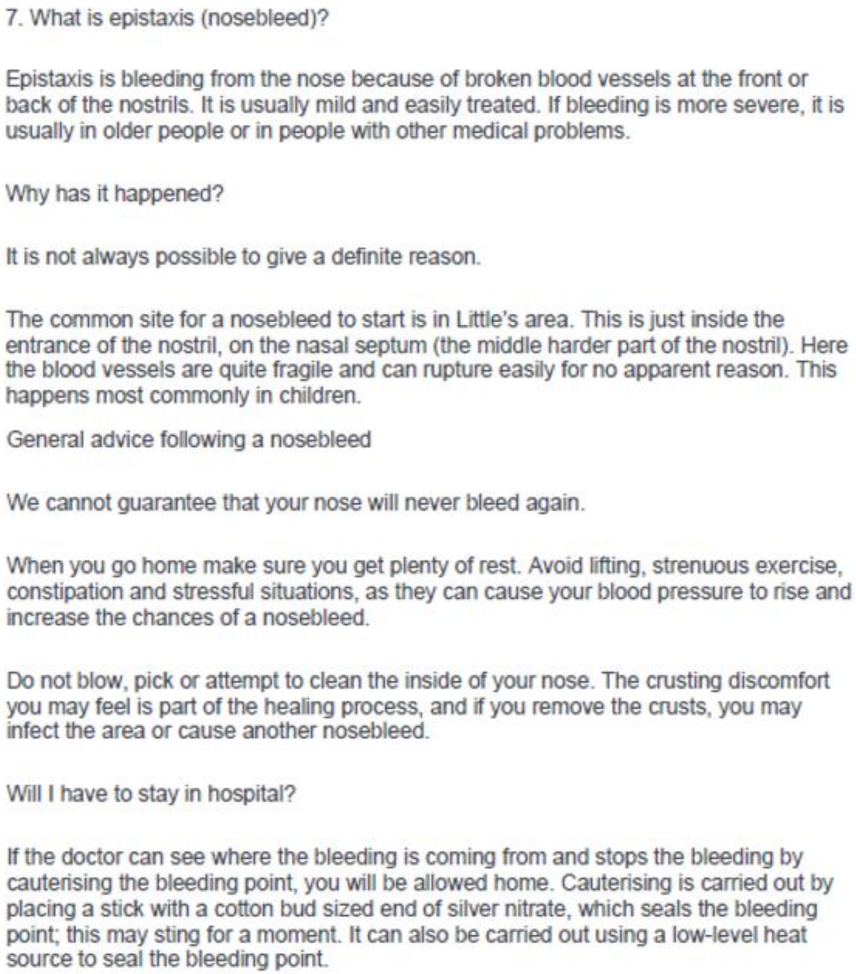
Pain is the body's way of sending a warning to the brain that something has happened and that you may need to take action. For example, if you hit your thumb with a hammer, pain would let you know to move your hand away.

After surgery it is not uncommon to have pain. This pain may be mild or could be more uncomfortable. Having pain after your operation does not necessarily mean that something is wrong, but it may affect your recovery if it stops you from being able to cough or move.

So it is important that your pain is controlled to make you feel more comfortable to move around. It is also very important for you to be able to take deep breaths and cough without being in pain: this will help prevent chest infection after your operation.

By treating your pain we may also help to make you feel better. Pain control may help you recover more quickly, and help you to return home sooner.

Page 19



7. What is epistaxis (nosebleed)?

Epistaxis is bleeding from the nose because of broken blood vessels at the front or back of the nostrils. It is usually mild and easily treated. If bleeding is more severe, it is usually in older people or in people with other medical problems.

Why has it happened?

It is not always possible to give a definite reason.

The common site for a nosebleed to start is in Little's area. This is just inside the entrance of the nostril, on the nasal septum (the middle harder part of the nostril). Here the blood vessels are quite fragile and can rupture easily for no apparent reason. This happens most commonly in children.

General advice following a nosebleed

We cannot guarantee that your nose will never bleed again.

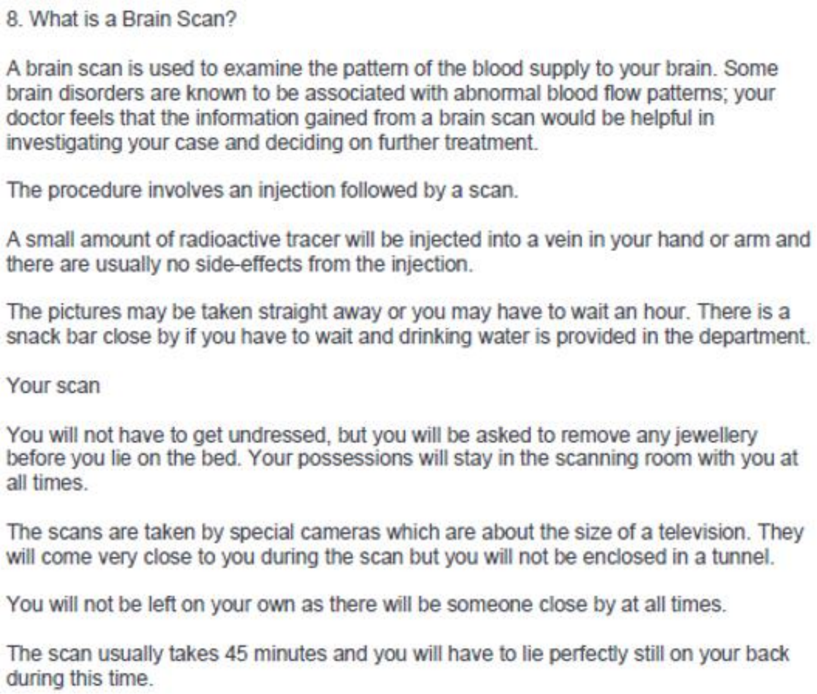
When you go home make sure you get plenty of rest. Avoid lifting, strenuous exercise, constipation and stressful situations, as they can cause your blood pressure to rise and increase the chances of a nosebleed.

Do not blow, pick or attempt to clean the inside of your nose. The crusting discomfort you may feel is part of the healing process, and if you remove the crusts, you may infect the area or cause another nosebleed.

Will I have to stay in hospital?

If the doctor can see where the bleeding is coming from and stops the bleeding by cauterising the bleeding point, you will be allowed home. Cauterising is carried out by placing a stick with a cotton bud sized end of silver nitrate, which seals the bleeding point; this may sting for a moment. It can also be carried out using a low-level heat source to seal the bleeding point.

Page 20



8. What is a Brain Scan?

A brain scan is used to examine the pattern of the blood supply to your brain. Some brain disorders are known to be associated with abnormal blood flow patterns; your doctor feels that the information gained from a brain scan would be helpful in investigating your case and deciding on further treatment.

The procedure involves an injection followed by a scan.

A small amount of radioactive tracer will be injected into a vein in your hand or arm and there are usually no side-effects from the injection.

The pictures may be taken straight away or you may have to wait an hour. There is a snack bar close by if you have to wait and drinking water is provided in the department.

Your scan

You will not have to get undressed, but you will be asked to remove any jewellery before you lie on the bed. Your possessions will stay in the scanning room with you at all times.

The scans are taken by special cameras which are about the size of a television. They will come very close to you during the scan but you will not be enclosed in a tunnel.

You will not be left on your own as there will be someone close by at all times.

The scan usually takes 45 minutes and you will have to lie perfectly still on your back during this time.

Appendix C: Chapter 6 Tables and Figures

Table 6.6. Random effects table of the final model (Perceived Understanding 3.1).

		Estimate	Est.Error	L-95%	U-95%
Standard deviation (Subject)	Intercept	8.93	2.77	3.82	14.67
	Age	15.91	6.09	5.04	28.85
	English proficiency	10.98	5.20	3.09	23.18
	UK-S-TOFHLA	22.75	8.70	8.27	42.24
	Education level	13.34	5.50	4.01	25.32
	FRE	5.85	2.02	2.10	10.00
	RDL2	6.04	2.23	2.00	10.72
Correlations	Intercept, Age	.00	.27	-.53	.52
	Intercept, English proficiency	.17	.30	-.45	.69
	Age, English proficiency	-.15	.31	-.70	.47
	Intercept, UK-S-TOFHLA	.03	.28	-.51	.57
	Age, UK-S-TOFHLA	-.28	.28	-.75	.32
	English proficiency, UK-S-TOFHLA	-.01	.32	-.61	.59
	Intercept, Education level	.06	.28	-.49	.58
	Age, Education level	-.07	.30	-.63	.51
	English proficiency, Education level	-.05	.31	-.62	.55
	UK-S-TOFHLA, Education level	-.02	.30	-.59	.56
	Intercept, FRE	.19	.28	-.39	.69
	Age, FRE	.01	.30	-.57	.58
	English proficiency, FRE	.00	.31	-.60	.59
	UK-S-TOFHLA, FRE	.15	.30	-.46	.69
	Education level, FRE	.01	.30	-.58	.58
	Intercept, RDL2	.03	.29	-.53	.60
	Age, RDL2	.02	.30	-.55	.59
	English proficiency, RDL2	-.05	.31	-.62	.55
	UK-S-TOFHLA, RDL2	.10	.30	-.51	.65
	Education level, RDL2	.12	.31	-.49	.67
	FRE, RDL2	-.25	.29	-.72	.39
Standard deviation (Subject:Text)	Intercept	3.96	.97	2.12	5.98
	Age	4.40	1.81	1.31	8.35
	English proficiency	3.38	1.52	.93	6.78
	UK-S-TOFHLA	4.87	2.15	1.37	9.67
	Education level	5.77	1.86	2.27	9.54
	FRE	5.22	2.35	1.44	10.52
	RDL2	4.83	2.13	1.37	9.54
Correlations	Intercept, Age	-.17	.28	-.69	.40
	Intercept, English proficiency	.26	.30	-.39	.75
	Age, English proficiency	-.11	.31	-.67	.52
	Intercept, UK-S-TOFHLA	.22	.29	-.39	.73
	Age, UK-S-TOFHLA	-.22	.31	-.75	.43
	English proficiency, UK-S-TOFHLA	-.11	.31	-.67	.51
	Intercept, Education level	.50	.22	-.02	.83
	Age, Education level	-.24	.29	-.73	.37
	English proficiency, Education level	.19	.30	-.44	.71
	UK-S-TOFHLA, Education level	.19	.30	-.44	.71
	Intercept, FRE	.17	.28	-.41	.67
	Age, FRE	.00	.31	-.59	.59
	English proficiency, FRE	.02	.31	-.59	.62
	UK-S-TOFHLA, FRE	.11	.31	-.51	.67
	Education level, FRE	.13	.29	-.46	.65
	Intercept, RDL2	-.01	.31	-.60	.58
	Age, RDL2	-.05	.31	-.64	.56
	English proficiency, RDL2	-.06	.31	-.64	.55
	UK-S-TOFHLA, RDL2	.03	.31	-.57	.62
	Education level, RDL2	.04	.31	-.55	.62
	FRE, RDL2	-.28	.32	-.80	.41

Table 6.8. Random effects table of the final model (Perceived Effort 3.1).

		Estimate	Est.Error	L-95%	U-95%
Standard deviation (Subject)	Intercept	13.65	3.86	6.50	21.80
	Age	15.70	6.59	4.74	30.31
	English proficiency	16.94	7.97	4.76	35.82
	UK-S-TOFHLA	34.12	12.05	14.06	60.98
	Education level	17.32	7.06	5.36	32.70
	FRE	6.87	2.54	2.29	12.17
	RDL2	9.06	2.98	3.59	15.39
Correlations	Intercept, Age	.25	.27	-.33	.72
	Intercept, English proficiency	.10	.30	-.52	.65
	Age, English proficiency	.02	.31	-.57	.61
	Intercept, UK-S-TOFHLA	.25	.25	-.27	.71
	Age, UK-S-TOFHLA	-.04	.30	-.59	.53
	English proficiency, UK-S-TOFHLA	-.02	.31	-.60	.57
	Intercept, Education level	-.06	.26	-.56	.45
	Age, Education level	-.04	.31	-.61	.55
	English proficiency, Education level	.01	.31	-.58	.60
	UK-S-TOFHLA, Education level	-.03	.30	-.59	.55
	Intercept, FRE	.01	.29	-.54	.57
	Age, FRE	.02	.31	-.57	.61
	English proficiency, FRE	.13	.31	-.49	.68
	UK-S-TOFHLA, FRE	.06	.30	-.53	.62
	Education level, FRE	-.12	.31	-.67	.50
	Intercept, RDL2	.37	.25	-.16	.78
	Age, RDL2	.07	.30	-.51	.63
	English proficiency, RDL2	.06	.31	-.56	.64
	UK-S-TOFHLA, RDL2	.29	.27	-.29	.75
	Education level, RDL2	-.10	.29	-.65	.49
	FRE, RDL2	-.23	.29	-.70	.40
Standard deviation (Subject:Text)	Intercept	5.99	1.17	3.89	8.45
	Age	4.88	2.00	1.50	9.24
	English proficiency	5.18	2.12	1.59	9.74
	UK-S-TOFHLA	7.15	2.90	2.29	13.47
	Education level	8.96	2.49	4.17	13.97
	FRE	5.59	2.49	1.57	11.12
	RDL2	6.97	2.99	2.00	13.59
Correlations	Intercept, Age	-.18	.28	-.68	.39
	Intercept, English proficiency	.45	.26	-.17	.85
	Age, English proficiency	-.14	.30	-.69	.47
	Intercept, UK-S-TOFHLA	.19	.30	-.44	.71
	Age, UK-S-TOFHLA	-.19	.31	-.72	.45
	English proficiency, UK-S-TOFHLA	-.04	.31	-.63	.56
	Intercept, Education level	.64	.16	.26	.88
	Age, Education level	-.19	.28	-.68	.38
	English proficiency, Education level	.42	.26	-.19	.83
	UK-S-TOFHLA, Education level	.14	.29	-.46	.67
	Intercept, FRE	-.05	.29	-.60	.51
	Age, FRE	-.05	.31	-.64	.56
	English proficiency, FRE	.03	.31	-.57	.61
	UK-S-TOFHLA, FRE	.03	.32	-.58	.63
	Education level, FRE	.05	.29	-.53	.60
	Intercept, RDL2	-.21	.27	-.70	.36
	Age, RDL2	.04	.31	-.56	.63
	English proficiency, RDL2	-.13	.30	-.67	.47
	UK-S-TOFHLA, RDL2	-.06	.31	-.64	.55
	Education level, RDL2	-.15	.28	-.65	.42
	FRE, RDL2	-.20	.31	-.74	.44

Table 6.9. Sensitivity Checks (Priors).

Model number	Priors	Probable effects (sign)	Highest Rhat	Chains	Notes
1	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Normal(0, 10)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
2	Predictors(Normal(0, 10)), Intercept(Normal(0, 25)), Random effects(Normal(0, 10)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	There were 17 divergent transitions after warmup.
3*	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.00	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
4	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
5	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .0001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
6	Predictors(Normal(0, 10)), Intercept(Norml(0, 10)), Random effects(Gamma(3, .00001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
7	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .000001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
8	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .01)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
9	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
10	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .0001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
11	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .00001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
12	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .000001)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
3.1**	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(2))	Age (+), UK-S-TOFHLA; health literacy (+), Education level (+)	1.00	6 chains with 8000 iterations per chain	No transitions beyond maximum treedepth.

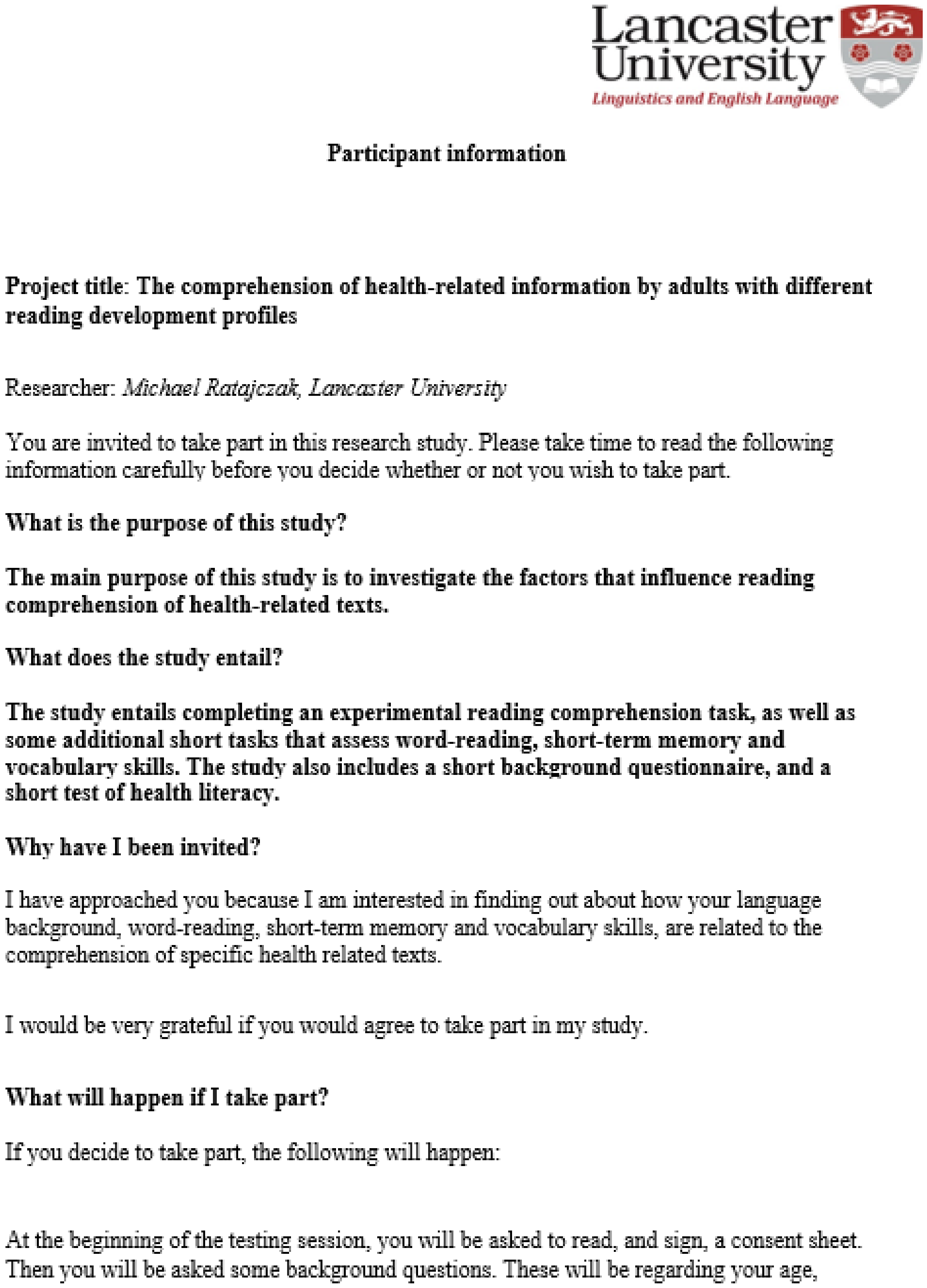
Notes. *The chosen model. **Local convergence check model (doubling the number of iterations).

Table 6.10. Sensitivity Checks (Priors).

Model number	Priors	Probable effects (sign)	Highest Rhat	Chains	Notes
1	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Normal(0, 10)), Covariance(LKJ(2))	UK-S-TOFHLA; health literacy (-), Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
2	Predictors(Normal(0, 10)), Intercept(Normal(0, 25)), Random effects(Normal(0, 10)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
3*	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.00	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
4	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	There were 1 divergent transitions after warmup.
5	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .0001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
6	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .00001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
7	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .000001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
8	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .01)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
9	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
10	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .0001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
11	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .00001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
12	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .000001)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.01	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
3.1**	Predictors(Normal(0, 10)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(2))	Flesch Reading Ease (-)	1.00	6 chains with 8000 iterations per chain	No transitions beyond maximum treedepth.

Notes. *The chosen model. **Local convergence check model (doubling the number of iterations).

Appendix D: Chapter 7 Materials



Lancaster University
Linguistics and English Language

Participant information

Project title: The comprehension of health-related information by adults with different reading development profiles

Researcher: *Michael Ratajczak, Lancaster University*

You are invited to take part in this research study. Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the purpose of this study?

The main purpose of this study is to investigate the factors that influence reading comprehension of health-related texts.

What does the study entail?

The study entails completing an experimental reading comprehension task, as well as some additional short tasks that assess word-reading, short-term memory and vocabulary skills. The study also includes a short background questionnaire, and a short test of health literacy.

Why have I been invited?

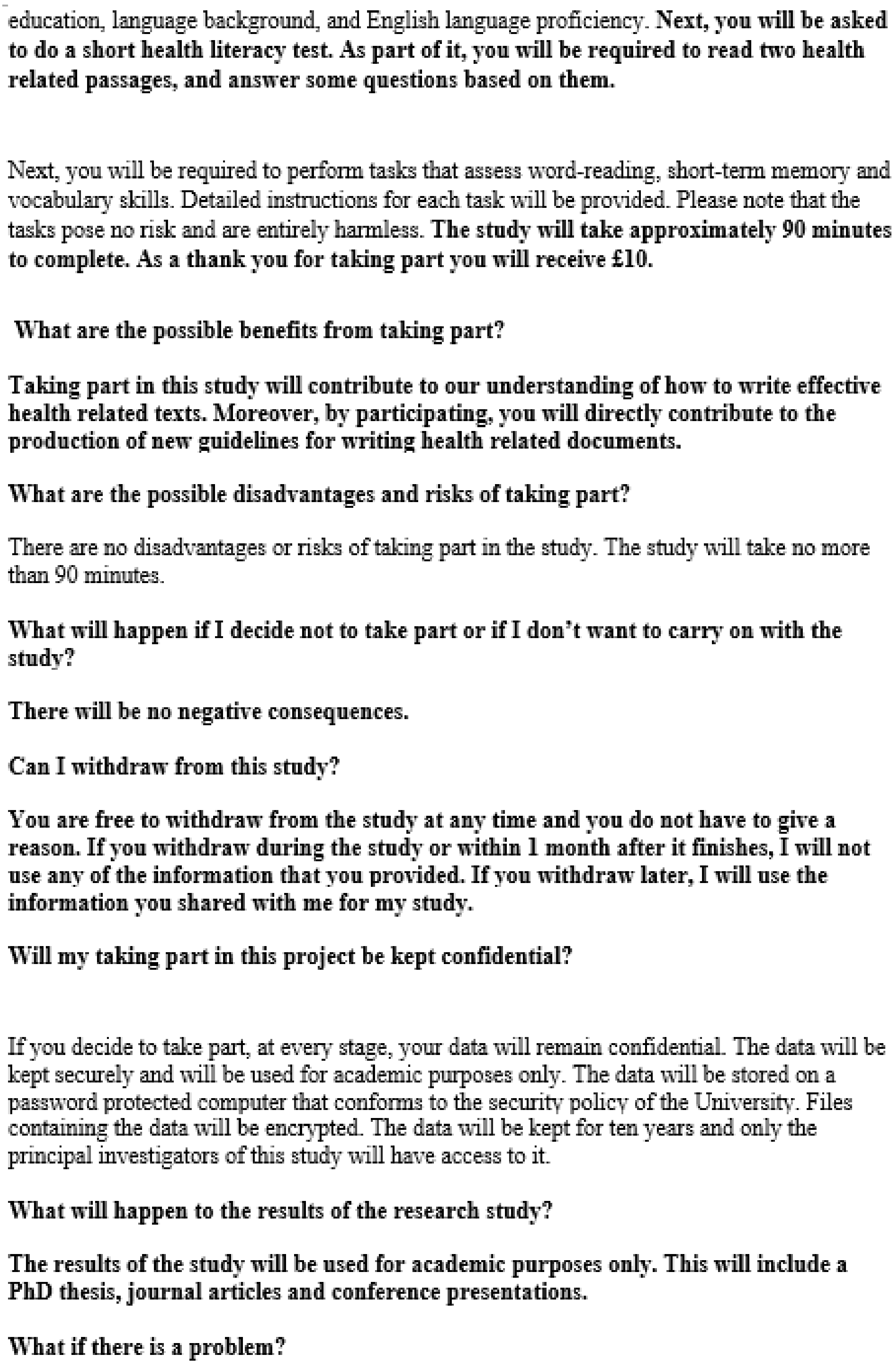
I have approached you because I am interested in finding out about how your language background, word-reading, short-term memory and vocabulary skills, are related to the comprehension of specific health related texts.

I would be very grateful if you would agree to take part in my study.

What will happen if I take part?

If you decide to take part, the following will happen:

At the beginning of the testing session, you will be asked to read, and sign, a consent sheet. Then you will be asked some background questions. These will be regarding your age,



education, language background, and English language proficiency. **Next, you will be asked to do a short health literacy test. As part of it, you will be required to read two health related passages, and answer some questions based on them.**

Next, you will be required to perform tasks that assess word-reading, short-term memory and vocabulary skills. Detailed instructions for each task will be provided. Please note that the tasks pose no risk and are entirely harmless. **The study will take approximately 90 minutes to complete. As a thank you for taking part you will receive £10.**

What are the possible benefits from taking part?

Taking part in this study will contribute to our understanding of how to write effective health related texts. Moreover, by participating, you will directly contribute to the production of new guidelines for writing health related documents.

What are the possible disadvantages and risks of taking part?

There are no disadvantages or risks of taking part in the study. The study will take no more than 90 minutes.

What will happen if I decide not to take part or if I don't want to carry on with the study?

There will be no negative consequences.

Can I withdraw from this study?

You are free to withdraw from the study at any time and you do not have to give a reason. If you withdraw during the study or within 1 month after it finishes, I will not use any of the information that you provided. If you withdraw later, I will use the information you shared with me for my study.

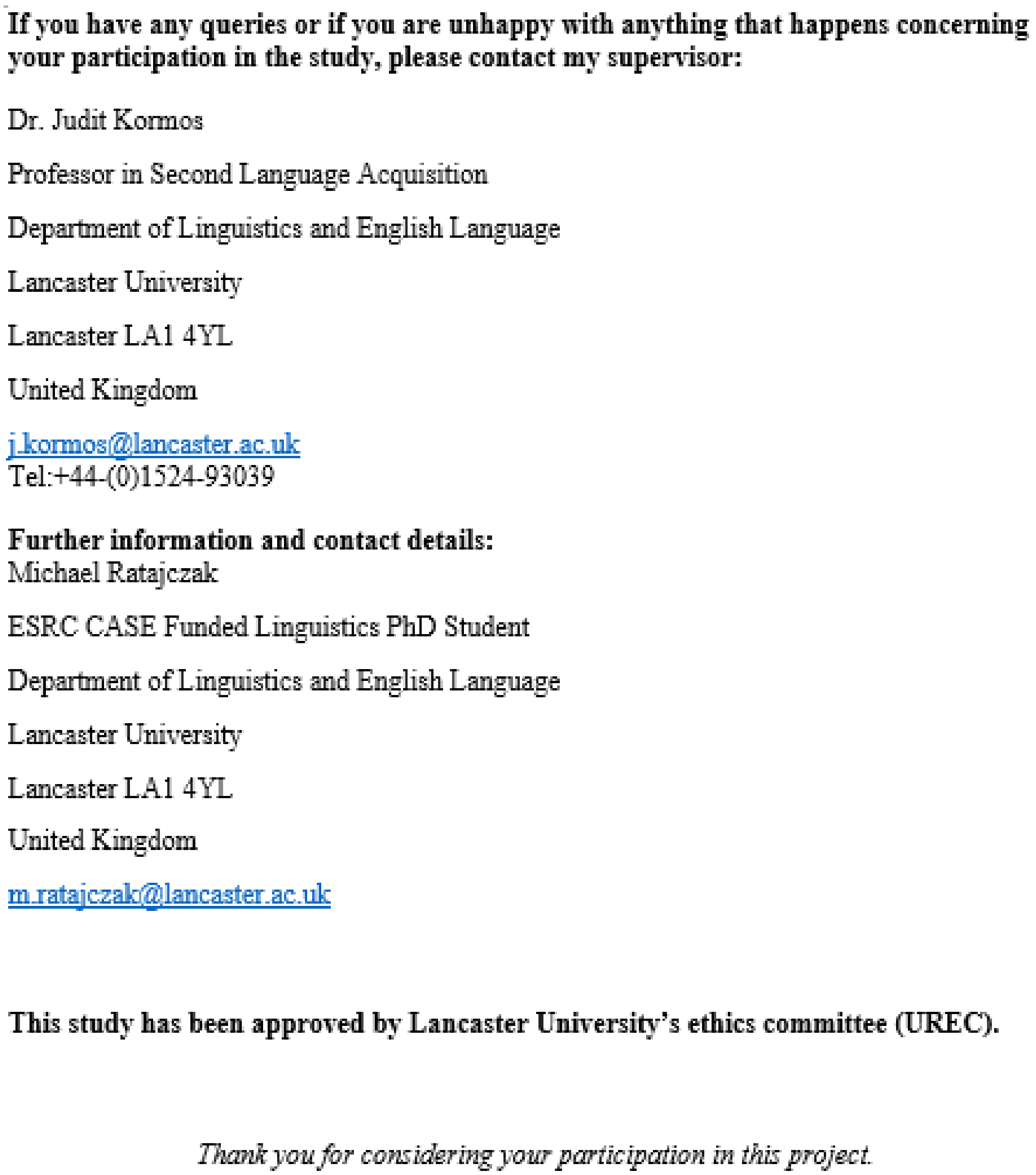
Will my taking part in this project be kept confidential?

If you decide to take part, at every stage, your data will remain confidential. The data will be kept securely and will be used for academic purposes only. The data will be stored on a password protected computer that conforms to the security policy of the University. Files containing the data will be encrypted. The data will be kept for ten years and only the principal investigators of this study will have access to it.

What will happen to the results of the research study?

The results of the study will be used for academic purposes only. This will include a PhD thesis, journal articles and conference presentations.

What if there is a problem?



If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact my supervisor:

Dr. Judit Kormos

Professor in Second Language Acquisition

Department of Linguistics and English Language

Lancaster University

Lancaster LA1 4YL

United Kingdom

j.kormos@lancaster.ac.uk
Tel:+44-(0)1524-93039

Further information and contact details:
Michael Ratajczak

ESRC CASE Funded Linguistics PhD Student

Department of Linguistics and English Language

Lancaster University

Lancaster LA1 4YL

United Kingdom

m.ratajczak@lancaster.ac.uk

This study has been approved by Lancaster University's ethics committee (UREC).

Thank you for considering your participation in this project.

Lancaster University
Linguistics and English Language

Consent Form – Study 3

Project title: The comprehension of health-related information by adults with different reading development profiles

I have read and had explained to me by *Michael Ratajczak* the information sheet relating to this project. □

I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements described in the information sheet in so far as they relate to my participation. □

I understand that my participation is entirely voluntary and that I have the right to withdraw from the project at any time. I understand that if I withdraw within a month of completion of the study, my data will be withdrawn. However, if I withdraw after this period, the information I have provided will be used in the project. □

I understand that all of the collected data will be encrypted, securely stored, anonymised, and that my identity will not be revealed at any point. □

I have received a copy of this consent form and of the accompanying information sheet. □

I agree to take part in this study. □

Name:

Signed:

Date:

Background questionnaire

What is your age?

……. Years

What is your gender?

…………………..

What is your native language?

…………………

If English is not your native language, what is your level of proficiency in English? Please circle the appropriate category.

• Beginner/Intermediate/Advanced

What is the highest level of education you have completed? Please circle the appropriate category.

• Primary school /secondary school/ vocational college / university

other, please specify: ……………………………………………………………

We want to see if written information produced by hospitals is written in an easy-to-read way. For this I would like to ask you some questions about hospital materials.

How often do you have someone (like a family member, friend, hospital/clinic worker, or caregiver) help you read hospital materials?'

(1) Always (2) Often (3) Sometimes (4) Occasionally (5) Never

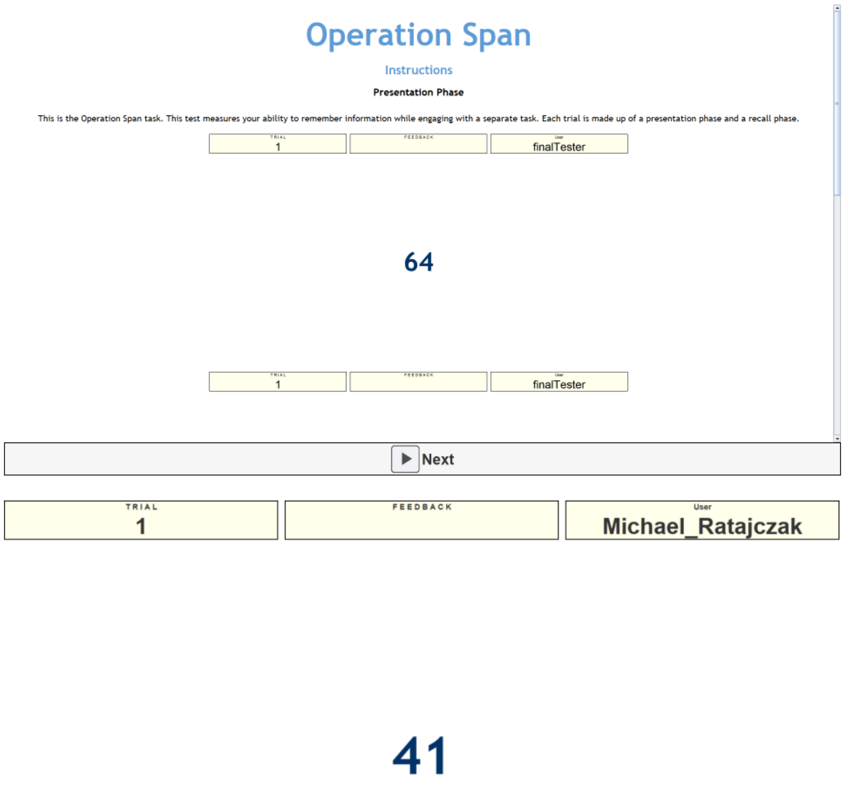
How confident are you filling out medical forms by yourself?

(1) Extremely (2) Quite a bit (3) Somewhat (4) A little bit (5) Not at all

How often do you have problems learning about your medical condition because of difficulty understanding written information?

(1) Always (2) Often (3) Sometimes (4) Occasionally (5) Never

Operation Span (WM Task)



Operation Span

Instructions

Presentation Phase

This is the Operation Span task. This test measures your ability to remember information while engaging with a separate task. Each trial is made up of a presentation phase and a recall phase.

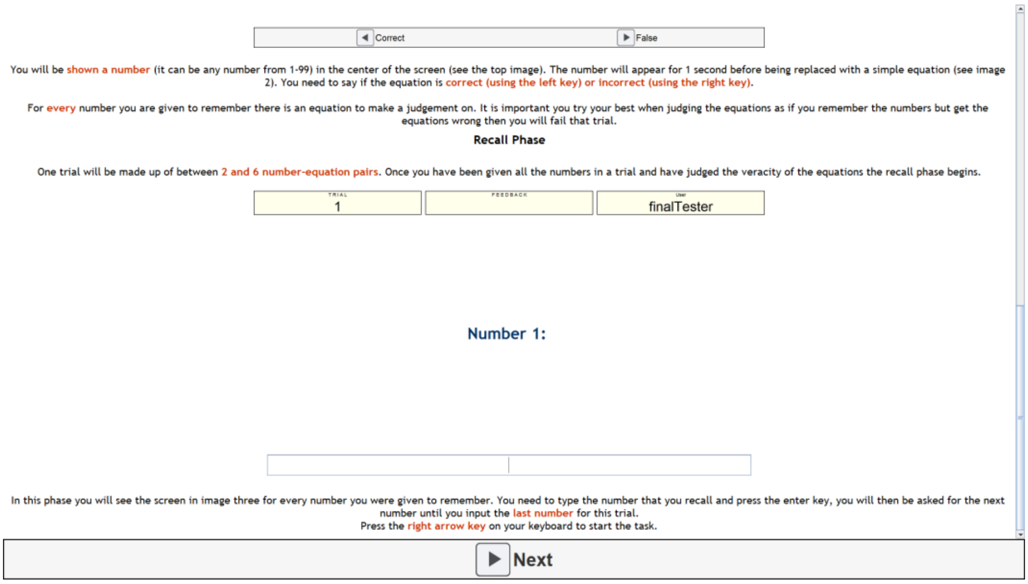
TRIAL	FEEDBACK	User
1		finalTester

64

TRIAL	FEEDBACK	User
1		finalTester

TRIAL	FEEDBACK	User
1		Michael_Ratajczak

41



You will be shown a number (it can be any number from 1-99) in the center of the screen (see the top image). The number will appear for 1 second before being replaced with a simple equation (see image 2). You need to say if the equation is correct (using the left key) or incorrect (using the right key).

For every number you are given to remember there is an equation to make a judgement on. It is important you try your best when judging the equations as if you remember the numbers but get the equations wrong then you will fail that trial.

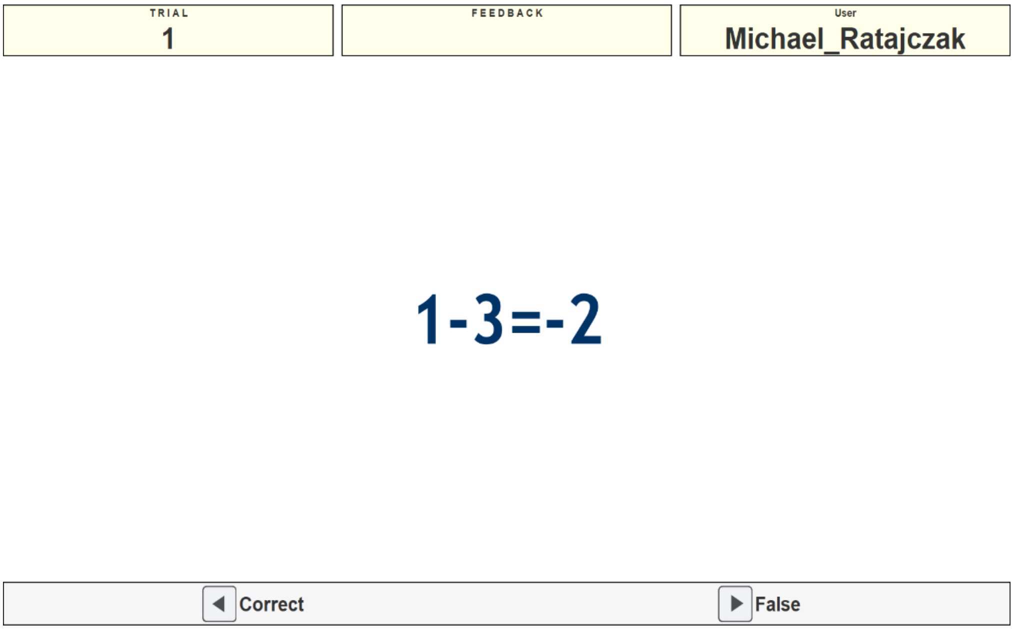
Recall Phase

One trial will be made up of between 2 and 6 number-equation pairs. Once you have been given all the numbers in a trial and have judged the veracity of the equations the recall phase begins.

TRIAL	FEEDBACK	User
1		finalTester

Number 1:

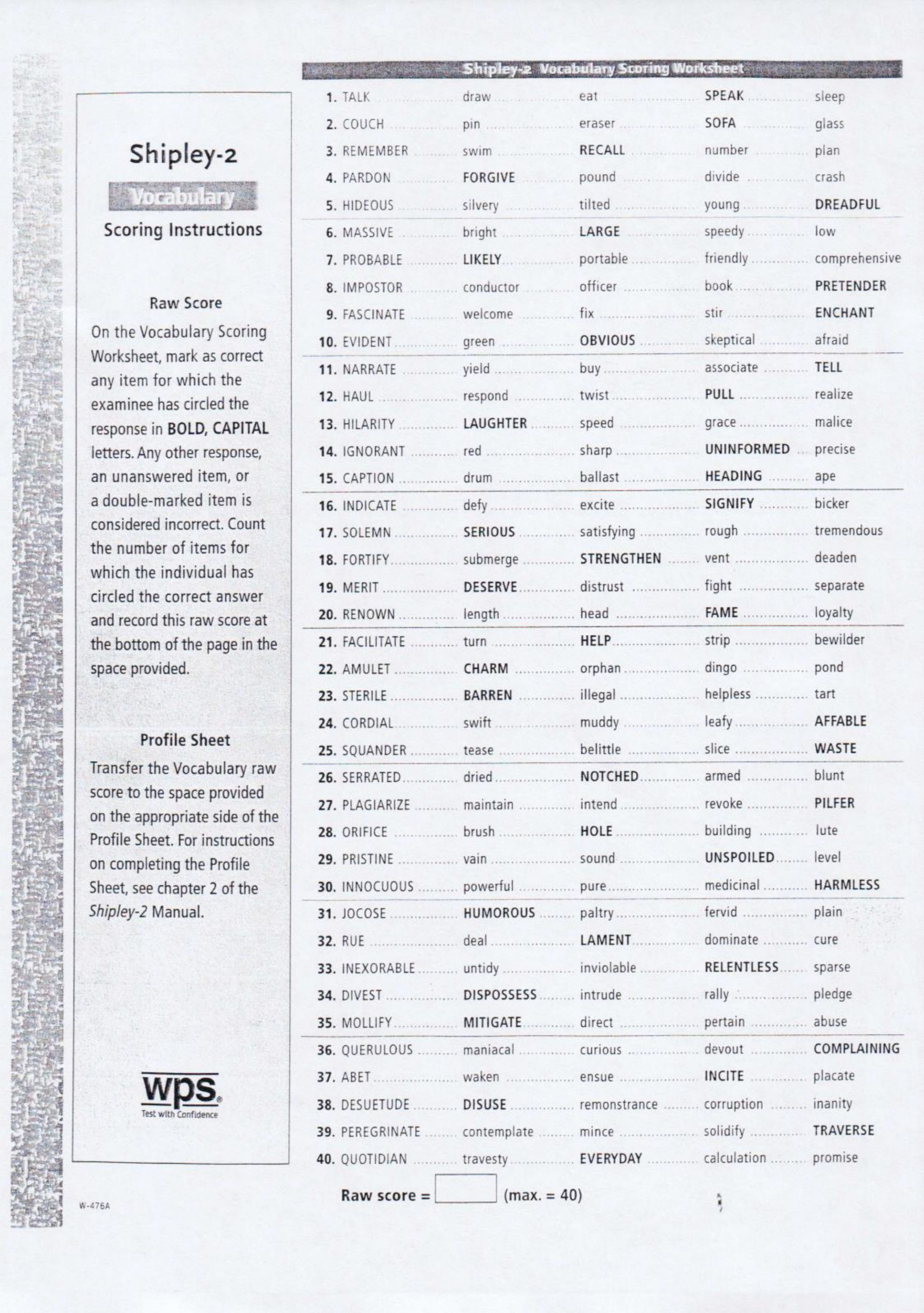
In this phase you will see the screen in image three for every number you were given to remember. You need to type the number that you recall and press the enter key, you will then be asked for the next number until you input the last number for this trial.
Press the right arrow key on your keyboard to start the task.



TRIAL	FEEDBACK	User
1		Michael_Ratajczak

$1-3=-2$

Vocabulary Test (Shipley-2 Vocabulary; correct answers are bolded)



Shipley-2

Vocabulary

Scoring Instructions

Raw Score

On the Vocabulary Scoring Worksheet, mark as correct any item for which the examinee has circled the response in **BOLD, CAPITAL** letters. Any other response, an unanswered item, or a double-marked item is considered incorrect. Count the number of items for which the individual has circled the correct answer and record this raw score at the bottom of the page in the space provided.

Profile Sheet

Transfer the Vocabulary raw score to the space provided on the appropriate side of the Profile Sheet. For instructions on completing the Profile Sheet, see chapter 2 of the *Shipley-2* Manual.

wps
Test with Confidence

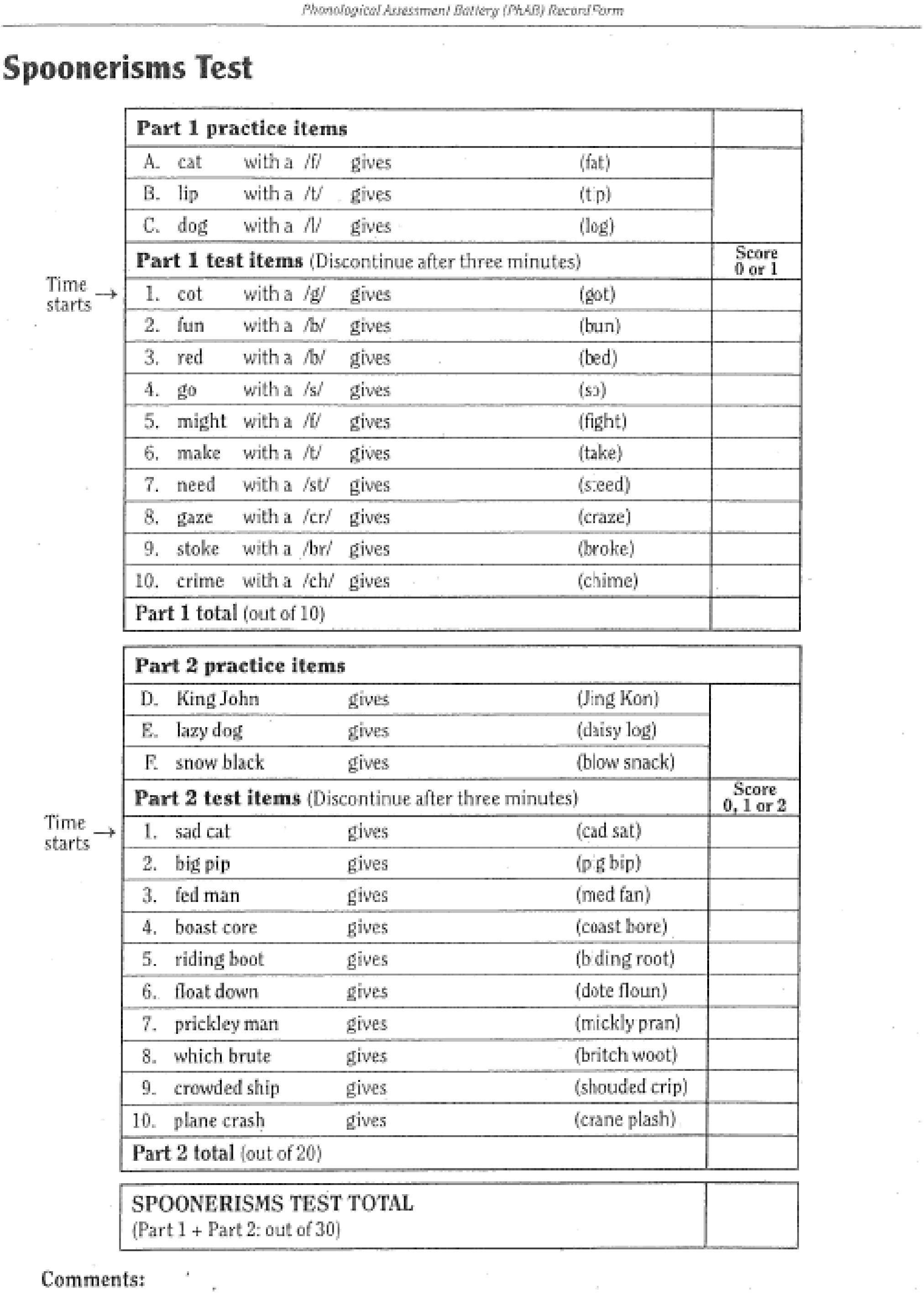
W-476A

Shipley-2 Vocabulary Scoring Worksheet

1.	TALK	draw	eat	**SPEAK**	sleep
2.	COUCH	pin	eraser	**SOFA**	glass
3.	REMEMBER	swim	**RECALL**	number	plan
4.	PARDON	**FORGIVE**	pound	divide	crash
5.	HIDEOUS	silvery	tilted	young	**DREADFUL**
6.	MASSIVE	bright	**LARGE**	speedy	low
7.	PROBABLE	**LIKELY**	portable	friendly	comprehensive
8.	IMPOSTOR	conductor	officer	book	**PRETENDER**
9.	FASCINATE	welcome	fix	stir	**ENCHANT**
10.	EVIDENT	green	**OBVIOUS**	skeptical	afraid
11.	NARRATE	yield	buy	associate	**TELL**
12.	HAUL	respond	twist	**PULL**	realize
13.	HILARITY	**LAUGHTER**	speed	grace	malice
14.	IGNORANT	red	sharp	**UNINFORMED**	precise
15.	CAPTION	drum	ballast	**HEADING**	ape
16.	INDICATE	defy	excite	**SIGNIFY**	bicker
17.	SOLEMN	**SERIOUS**	satisfying	rough	tremendous
18.	FORTIFY	submerge	**STRENGTHEN**	vent	deaden
19.	MERIT	**DESERVE**	distrust	fight	separate
20.	RENOWN	length	head	**FAME**	loyalty
21.	FACILITATE	turn	**HELP**	strip	bewilder
22.	AMULET	**CHARM**	orphan	dingo	pond
23.	STERILE	**BARREN**	illegal	helpless	tart
24.	CORDIAL	swift	muddy	leafy	**AFFABLE**
25.	SQUANDER	tease	belittle	slice	**WASTE**
26.	SERRATED	dried	**NOTCHED**	armed	blunt
27.	PLAGIARIZE	maintain	intend	revoke	**PILFER**
28.	ORIFICE	brush	**HOLE**	building	lute
29.	PRISTINE	vain	sound	**UNSPOILED**	level
30.	INNOCUOUS	powerful	pure	medicinal	**HARMLESS**
31.	JOCOSE	**HUMOROUS**	paltry	fervid	plain
32.	RUE	deal	**LAMENT**	dominate	cure
33.	INEXORABLE	untidy	inviolable	**RELENTLESS**	sparse
34.	DIVEST	**DISPOSSESS**	intrude	rally	pledge
35.	MOLLIFY	**MITIGATE**	direct	pertain	abuse
36.	QUERULOUS	maniacal	curious	devout	**COMPLAINING**
37.	ABET	waken	ensue	**INCITE**	placate
38.	DESUETUDE	**DISUSE**	remonstrance	corruption	inanity
39.	PEREGRINATE	contemplate	mince	solidify	**TRAVERSE**
40.	QUOTIDIAN	travesty	**EVERYDAY**	calculation	promise

Raw score = □ (max. = 40)

Phonological Assessment (Spoonerisms Test; Scoring Sheet)



Phonological Assessment Battery (PhAB) Record Form

Spoonerisms Test

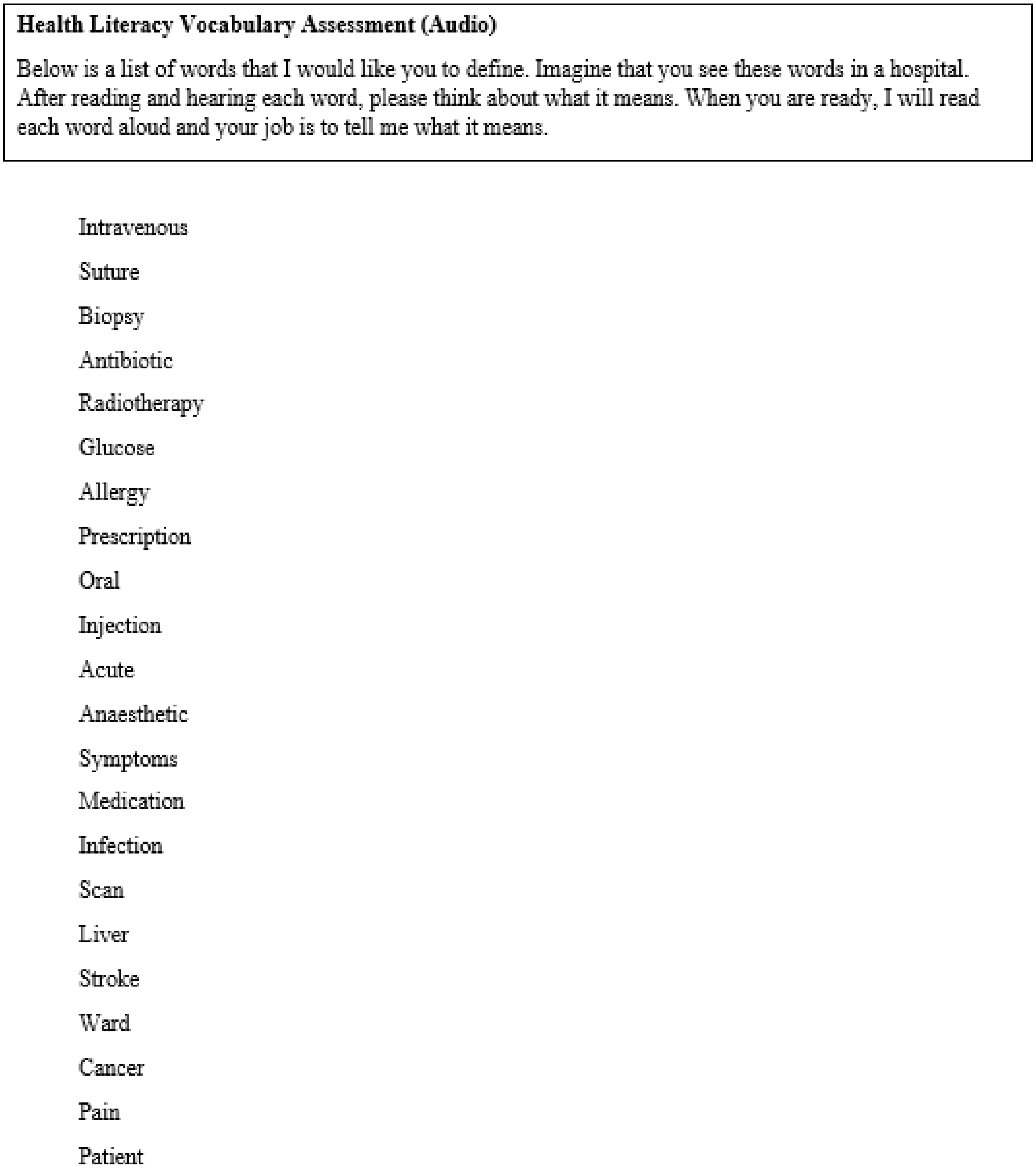
	Part 1 practice items						
	A.	cat	with a	/f/	gives	(fat)	
	B.	lip	with a	/t/	gives	(tip)	
	C.	dog	with a	/l/	gives	(log)	
	Part 1 test items (Discontinue after three minutes)						**Score 0 or 1**
Time starts →	1.	cot	with a	/g/	gives	(got)	
	2.	fun	with a	/b/	gives	(bun)	
	3.	red	with a	/b/	gives	(bed)	
	4.	go	with a	/s/	gives	(so)	
	5.	might	with a	/f/	gives	(fight)	
	6.	make	with a	/t/	gives	(take)	
	7.	need	with a	/st/	gives	(steed)	
	8.	gaze	with a	/cr/	gives	(craze)	
	9.	stoke	with a	/br/	gives	(broke)	
	10.	crime	with a	/ch/	gives	(chime)	
	Part 1 total (out of 10)						

	Part 2 practice items				
	D.	King John	gives	(Jing Kon)	
	E.	lazy dog	gives	(daisy log)	
	F.	snow black	gives	(blow snack)	
	Part 2 test items (Discontinue after three minutes)				**Score 0, 1 or 2**
Time starts →	1.	sad cat	gives	(cad sat)	
	2.	big pip	gives	(pig bip)	
	3.	fed man	gives	(med fan)	
	4.	boast core	gives	(coast bore)	
	5.	riding boot	gives	(biding root)	
	6.	float down	gives	(dote floun)	
	7.	prickley man	gives	(mickly pran)	
	8.	which brute	gives	(britch woot)	
	9.	crowded ship	gives	(shouded crip)	
	10.	plane crash	gives	(crane plash)	
	Part 2 total (out of 20)				

SPOONERISMS TEST TOTAL (Part 1 + Part 2: out of 30)	

Comments:

Health Literacy Assessment 1: Health Literacy Vocabulary Assessment



Health Literacy Vocabulary Assessment (Audio)

Below is a list of words that I would like you to define. Imagine that you see these words in a hospital. After reading and hearing each word, please think about what it means. When you are ready, I will read each word aloud and your job is to tell me what it means.

Intravenous

Suture

Biopsy

Antibiotic

Radiotherapy

Glucose

Allergy

Prescription

Oral

Injection

Acute

Anaesthetic

Symptoms

Medication

Infection

Scan

Liver

Stroke

Ward

Cancer

Pain

Patient

Health Literacy Assessment 1: Health Literacy Vocabulary Assessment (Scoring Sheet)

Each item is scored as either incorrect (0), partially correct (1) or fully correct (2). The criteria are based on definitions from Oxford's Concise Colour Medical Dictionary (Sixth Edition; 2015) and Cambridge's English Dictionary (2018). Italicized text illustrates the context in which the answers should be provided. This test is not easy, and participants should not be expected to score at ceiling.

1. Intravenous

Score a point for mentioning the following: "**into/connected/within**" (1) and "**vein**" (1).

PL: Score a point for mentioning the following or their variations: "**wprowadzajac/podajac/dostarczajac/w/do**" (1) i "**zyla** (oraz odmiany)" (1).

Example answers:

Oxford Concise Medical Dictionary

- *Adjective*. Into or within a vein

Cambridge Dictionary

- *Adjective*. Into or connected to a vein.

2. Suture

Score a point for mentioning the following or their variations: "**closure/sew up/close/keeping together**" (1) "**of a wound/a wound/a cut/**" (1). Saying "**a stich**" is not part of the definition.

PL: Score a point for mentioning the following or their variations: "**zamknac/zszyc/zeszyc**" (1) i "**rane**" (1).

Oxford Concise Medical Dictionary

- *Noun*. (in surgery) the closure of a wound or incision to facilitate the healing process, using any of various materials.
- *Noun*. The material – silk, catgut, nylon, any of various polymers, or wire – used to sew up a wound.
- *Verb*. To close a wound by suture

Cambridge Dictionary

- *Noun*. A stitch used to sew up a cut in a person's body.
- *Verb*. To sew together a cut in a person's body.

3. Biopsy

Score a point for mentioning the following: *removal of* "**tissue/skin/organ**" (1), "**for examination/for testing/to discover more about an illness**" (1).

PL: Score a point for mentioning the following or their variations: *pobranie* **"material/tkanki/skory/organu"** (1) i **"do badania/badan/przebadania/do przeprowadzenia badan/do przetestowania/aby dowiedziec sie wiecej o chorobie"** (1).

Oxford Concise Medical Dictionary

- *Noun*. The removal of a small piece of living tissues from an organ or part of the body for microscopic examination.

Cambridge Dictionary

- *Noun*. The process of removing and examining a small amount of tissue from a sick person, in order to discover more about their illness.

4. Antibiotic

Score a point for mentioning the following: "**substance/drug/antimicrobial dug/medicine/chemical**" (1) and *against/treat* "**bacteria/fungi/infections/germs**" (1).

PL: Score a point for mentioning the following or their variations: "**substancja/lekartswo**" (1) *ktore/a niszczy/zabija/zwalcza/leczy itp*. "**bakterie/drobnoustroje/zakazenie/mikroorganizmy** (oraz odmiany)" (1).

Oxford Concise Medical Dictionary

- *Noun*. A substance, produced by or derived from a microorganism, that destroys or inhibits the growth of other microorganisms. Antibiotics are used to treat infections caused by organisms that are sensitive to them, usually bacteria or fungi.

Cambridge Dictionary

- *Noun*. A medicine or chemical that can destroy harmful bacteria in the body or limit their growth.

5. Radiotherapy

Score a point for mentioning the following or their variations: **"the use of radiation/energy"** (1) **"to treat disease/treatment of disease"** (1).

PL: Score a point for mentioning the following or their variations: "**metoda leczenia chorob/leczenie chorob"** (1) *za pomoca* "**promieniowania/energi promieniowania**" (1).

Oxford Concise Medical Dictionary

- *Noun*. Therapeutic radiology: the treatment of disease with penetrating radiation, such as X-rays, beta rays, or gamma rays, which may be produced by machines or given off by radioactive isotopes.
- Many forms of cancer are destroyed by radiotherapy.

Cambridge Dictionary

- *Noun*. The use of controlled amounts of radiation (= a form of energy) aimed at a particular part of the body, to treat disease.

6. Glucose

Score a point for mentioning the following or their variations: "**a simple sugar/a type of sugar/sugar**" (1) "**source of energy/supplies energy**" (1).

PL: Score a point for mentioning the following or their variations: "**cukier prosty/rodzaj cukru/cukier**" (1) "**dostarcza energi/zrodlo energi**" (1).

Oxford Concise Medical Dictionary

- *Noun.* A simple sugar containing six carbon atoms (a hexose). Glucose is an important source of energy in the body and the sole source of energy for the brain.

Cambridge Dictionary

- *Noun.* A type of sugar that is found in plants, especially fruit, and supplies an important part of the energy that animals need.

7. Allergy

Score a point for mentioning the following and their variations "**sensitivity/hypersensitivity**" (1), *to* "**certain substances/antigens/allergens**" (1).

PL: Score a point for mentioning the following or their variations: "**wrazliwosc/nadwrazliwosc/reakcja**" (1) *organizmu na* "**pewne czynniki/substancje/alergeny** (oraz odmiany)" (1).

Oxford Concise Medical Dictionary

- *Noun.* A disorder in which the body becomes hypersensitive to particular antigens (called allergens), which provoke characteristic symptoms whenever they are subsequently inhaled, ingested, injected, or otherwise contacted.

Cambridge Dictionary

- *Noun.* A condition that makes a person become sick or develop skin or breathing problems because they have eaten certain foods or been near certain substances.

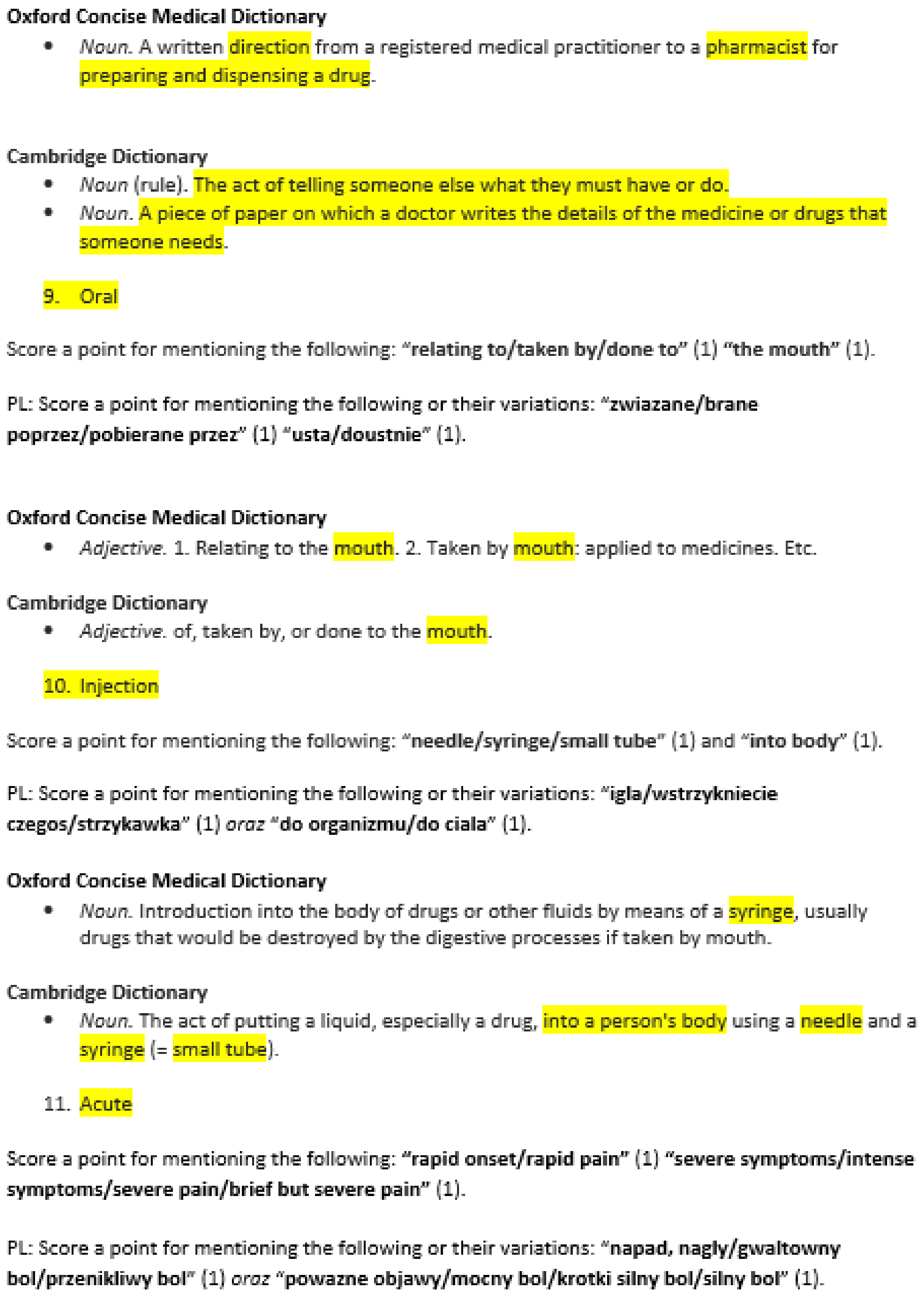
8. Prescription

Score a point for mentioning either one of or the following or their variations:

- "**a written direction/instruction**" (1) "**what drugs to give to a person/patient**" (1)
- or "**a piece of paper on which a doctor writes the details of the medicine or drugs**" (1) "**that someone needs**" (1)
- or *in the context of a doctor telling a patient* "**the act of telling someone else**" (1) "**what they must have or do**" (1)

PL: Score a point for mentioning the following or their variations:

- "**zlecenie lekarskie/pisemne zlecenie lekarskie/instrukcja/przepis**" (1) *na ktorego podstawie* "**sporzadzane sa leki/wydawane sa leki/**" (1).
- "**zlecenie lekarskie/pisemne zlecenie lekarskie/instrukcja**" (1) *ktore precyzuje jakie* "**lekarstwa/leki/antybiotyki**" (1) *pacjent potrzebuje.*
- *Kiedy* "**lekarz**" (1) "**mowi pacjentowi co ma miec albo robic**" (1).



Oxford Concise Medical Dictionary

- *Noun*. A written direction from a registered medical practitioner to a pharmacist for preparing and dispensing a drug.

Cambridge Dictionary

- *Noun* (rule). The act of telling someone else what they must have or do.
- *Noun*. A piece of paper on which a doctor writes the details of the medicine or drugs that someone needs.

9. Oral

Score a point for mentioning the following: **"relating to/taken by/done to"** (1) **"the mouth"** (1).

PL: Score a point for mentioning the following or their variations: **"zwiazane/brane poprzez/pobierane przez"** (1) **"usta/doustnie"** (1).

Oxford Concise Medical Dictionary

- *Adjective*. 1. Relating to the mouth. 2. Taken by mouth: applied to medicines. Etc.

Cambridge Dictionary

- *Adjective*. of, taken by, or done to the mouth.

10. Injection

Score a point for mentioning the following: **"needle/syringe/small tube"** (1) and **"into body"** (1).

PL: Score a point for mentioning the following or their variations: **"igla/wstrzykniecie czegos/strzykawka"** (1) *oraz* **"do organizmu/do ciala"** (1).

Oxford Concise Medical Dictionary

- *Noun*. Introduction into the body of drugs or other fluids by means of a syringe, usually drugs that would be destroyed by the digestive processes if taken by mouth.

Cambridge Dictionary

- *Noun*. The act of putting a liquid, especially a drug, into a person's body using a needle and a syringe (= small tube).

11. Acute

Score a point for mentioning the following: **"rapid onset/rapid pain"** (1) **"severe symptoms/intense symptoms/severe pain/brief but severe pain"** (1).

PL: Score a point for mentioning the following or their variations: **"napad, nagly/gwaltowny bol/przenikliwy bol"** (1) *oraz* **"powazne objawy/mocny bol/krotki silny bol/silny bol"** (1).

Oxford Concise Medical Dictionary

- *Adjective*. 1. Describing a disease of rapid onset, severe symptoms, and brief duration. 2. Describing any intense symptom, such as severe pain.

Cambridge Dictionary

- *Noun*. 1. If a bad situation is acute, it causes severe problems or damage. 2. An acute pain or illness is one that quickly becomes very severe.

12. Anaesthetic

Score a point for mentioning the following or their variations: **“a substance”** (1) *that makes you or someone else* **“unable to feel pain/reduces or abolishes sensation”** (1).

PL: Score a point for mentioning the following or their variations: “**substancja/lek/cos**” (1) *co sprawia* **“nieodczuwanie bolu/zmnieszyc wrazliwosci na cierpienia** (lub bol,)**, usmierzyc bol,zniesc bol”** (1).

Oxford Concise Medical Dictionary

- *Noun*. An agent that reduces or abolishes sensation, affecting either the whole body (general anaesthetic) or a particular area or region (local anaesthetic). 2. *Adjective*. Reducing or abolishing sensation.

Cambridge Dictionary

- *Noun*. A substance that makes you unable to feel pain.

13. Symptom(s)

Score a point for mentioning the following: **“an indication/sign/feeling/problem”** (1) *of/caused by* **“disease (or disorder)/illness/another problem”** (1).

PL: Score a point for mentioning the following or their variations: “**oznaka/objaw/sygnaly/uczucie/problem**” (1) *spowodowane*“**choroby/problem/stanu pacjenta**” (1).

Oxford Concise Medical Dictionary

- *Noun*. An indication of a disease or disorder noticed by the patient himself.

Cambridge Dictionary

- *Noun*. 1. Any feeling of illness or physical or mental change that is caused by a particular disease. 2. Any single problem that is caused by and shows a more serious and general problem.

14. Medication

Score a point for mentioning the following: **“a substance/drugs/medicine”** (1) **“used for treatment/to improve patient’s condition (or illness)**” (1).

PL: Score a point for mentioning the following or their variations: **"substancja/antybiotyk/tabletka/lek"** (1) *ktora/y* **"uzywany w leczeniu/leczy choroby/pamaga pacjentowi/srodek zaradczy/pomagajacy"** (1).

Oxford Concise Medical Dictionary

- *Noun*. 1. A substance administered by mouth, applied to the body, or introduced into the body for the purpose of treatment. 2. Treatment of a patient using drugs.

Cambridge Dictionary

- *Noun*. A medicine, or a set of medicines or drugs, used to improve a particular condition or illness.

15. Infection

Score a point for mentioning the following: **"a disease/illness/inflammation/invasion of the body"** (1) **"organisms/pathogens/bacteria/fungi/viruses/germs"** (1).

PL: Score a point for mentioning the following or their variations: **"zapalenie/choroba wywolana poprzez/inwazja organizmu** (lub: ciala)**"** (1) **"wniknie cie drobnoustrojow** (lub: bakteri/virusow/grzybow/patogenow**) do organizmu** (lub: ciala/skory)**"** (1).

Oxford Concise Medical Dictionary

- *Noun*. Invasion of the body by harmful organisms (pathogens), such as bacteria, fungi, protozoa, rickettsia, or viruses.

Cambridge Dictionary

- *Noun*. A disease in a part of your body that is caused by bacteria or a virus.

16. Scan

Score a point for mentioning the following: **"an examination of the body or part of the body"** (1) **"ultrasonography/computerized tomography/MRI/scintigraphy/image/machine"** (1).

PL: Score a point for mentioning the following or their variations: **"badanie ciala lub czesci ciala"** (1) *poprzez* **"ultrasonografie (USG)/tomografie/rezonans magnetyczny/scyntygrafie/zdjecie/maszyne"** (1).

Oxford Concise Medical Dictionary

- *Noun*. An examination of the body or part of the body using ultrasonography, computerized tomography, MRI, or scintigraphy. 2. The image obtained from such an examination. 3. *Verb*. To examine the body using any of these techniques.

Cambridge Dictionary

- *Verb.* to look at something carefully, with the eyes or with a machine, in order to get information.
- *Noun.* a medical examination in which an image of the inside of the body is made using a special machine.

17. Liver

Score correct for mentioning the following: "**gland/organ**" (1) "**cleans the blood/produces bile/metabolises food/regulates blood sugar/detoxifies**" (1).

PL: Score a point for mentioning the following or their variations: "**gruczol/organ/narzad**" (1) "**czysci krew/przemienia materie/metabolizuje materie/wydziela zolc/reguluje cukier we krwi/odtruwa**" (1).

Oxford Concise Medical Dictionary

- *Noun.* The largest gland of the body. Situated in the top right portion of the abdominal cavity. The liver has a number of important functions. It synthesises bile. It is an important site of metabolism of carbohydrates, proteins and fats. It regulates the amount of blood sugar. It has an important role in the detoxification of poisonous substances.

Cambridge Dictionary

- *Noun.* A large organ in the body that cleans the blood and produces bile, or this organ from an animal used as meat.

18. Stroke

Score a point for mentioning the following or their variations: "**interruption/change in the blood supply/flow** (also accept "**bleeding**")" (1) "**to the brain/in the brain (or to a part of the brain)**" (1).

PL: Score a point for mentioning the following or their variations: "**zaburzenie czynnosci/wylew krwi/zatrzymanie doplywu krwi/krwawienie**" (1) "**mozgu/w mozgu/czesci mozgu**" (1).

Oxford Concise Medical Dictionary

- *Noun.* A sudden attack of weakness usually affecting one side of the body. It is the consequence of an interruption to the flow of blood to the brain.

Cambridge Dictionary

- *Noun.* A sudden change in the blood supply to a part of the brain, sometimes causing a loss of the ability to move particular parts of the body.

19. Ward

Score a point for mentioning the following or their variations: "**large room in a hospital**" (1) "**with beds for patients**" (1).

Extension: Alternatively, score a point for mentioning the following or their variations: **"department/unit/area"** (1) **"of a hospital"** (1).

PL: Score a point for mentioning the following or their variations: **"sala/pomieszczenie/pokoj (**w szpitalu)" (1) **"chorych/z lozkami dla pacjentow"** (1).

PL Extension: Alternatively, score a point for mentioning the following or their variations: **"wyodrebniona czesc szpitala/departament/oddzial/wydzial"** (1) **"spelniajaca okreslone funkcje/szpitalny"** (1).

Oxford Concise Medical Dictionary

- *No definition.*

Cambridge Dictionary

- *Noun.* One of the **parts** or **large rooms** into which a **hospital** is **divided**, usually with beds for patients.

20. Cancer

Score a point for mentioning the following and their variations: **"disease/tumour"** (1) and **"division of cells/multiplication of cells/cells in the body grow in an uncontrolled** (also accept "**not normal/abnormal"**) **way**" (1).

PL: Score a point for mentioning the following or their variations: "**choroba/nowotwor/guz**" (1) I *nienormalne/niekontrolowane* "**podzial/dzielenie/mnozenie komorek/rozrost komorek/proliferacja komorek (w nienormalny sposob)**" (1).

Oxford Concise Medical Dictionary

- *Noun.* Any malignant tumour, including carcinoma, lymphoma, leukaemia, and sarcoma. It arises from the abnormal, purposeless, and uncontrolled division of cells that the invade and destroy surrounding tissues.

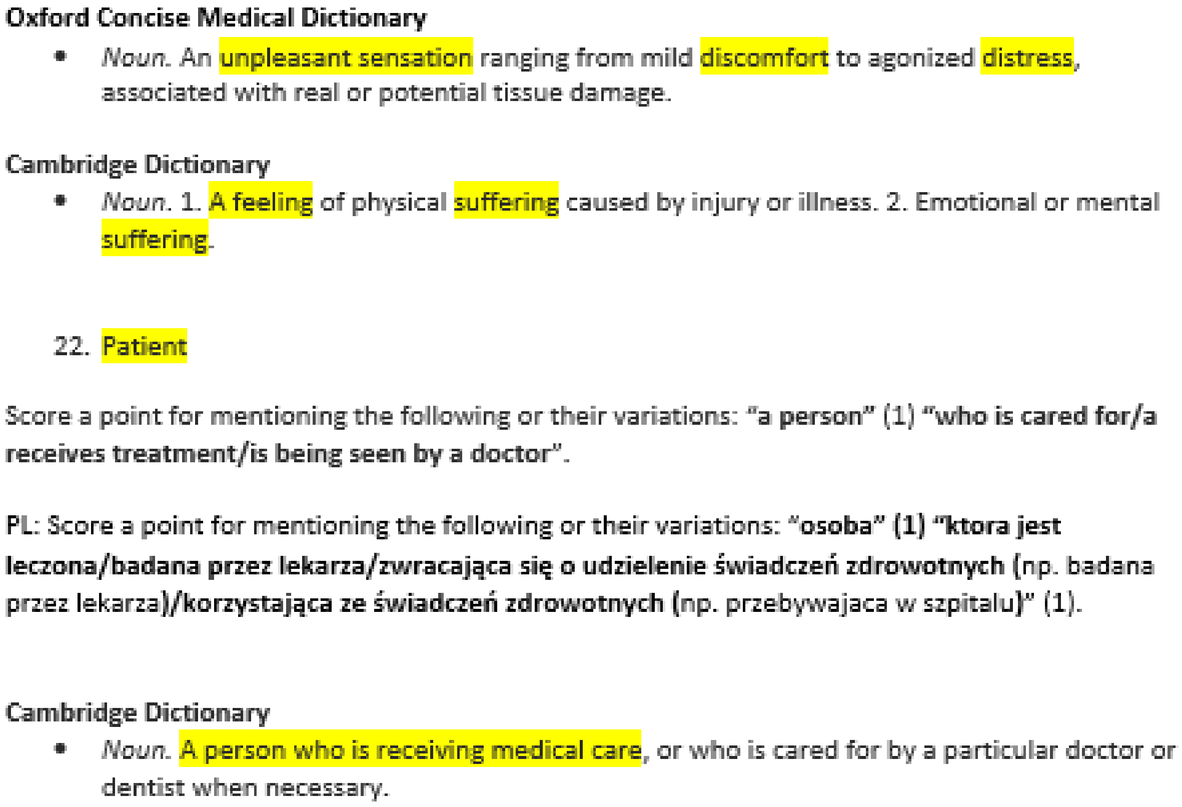
Cambridge Dictionary

- *Noun.* A serious disease that is caused when cells in the body grow in a way that is uncontrolled and not normal, killing normal cells and often causing death.

21. Pain

Score a point for mentioning the following: **"unpleasant sensation/discomfort/distress/suffering/when it hurts"** (1) **"caused by injury/illness/tissue damage"** (1).

PL: Score a point for mentioning the following or their variations: **"nieprzyjemne/przykre wrazenie/negatywne wrazenie/niewygoda/cierpienie"** (1) *powstajace pod wplywem/przez* **"zranienia/bodzcow/uszkodzenie tkanki/chorobe"** (1).



Oxford Concise Medical Dictionary

- *Noun*. An unpleasant sensation ranging from mild discomfort to agonized distress, associated with real or potential tissue damage.

Cambridge Dictionary

- *Noun*. 1. A feeling of physical suffering caused by injury or illness. 2. Emotional or mental suffering.

22. Patient

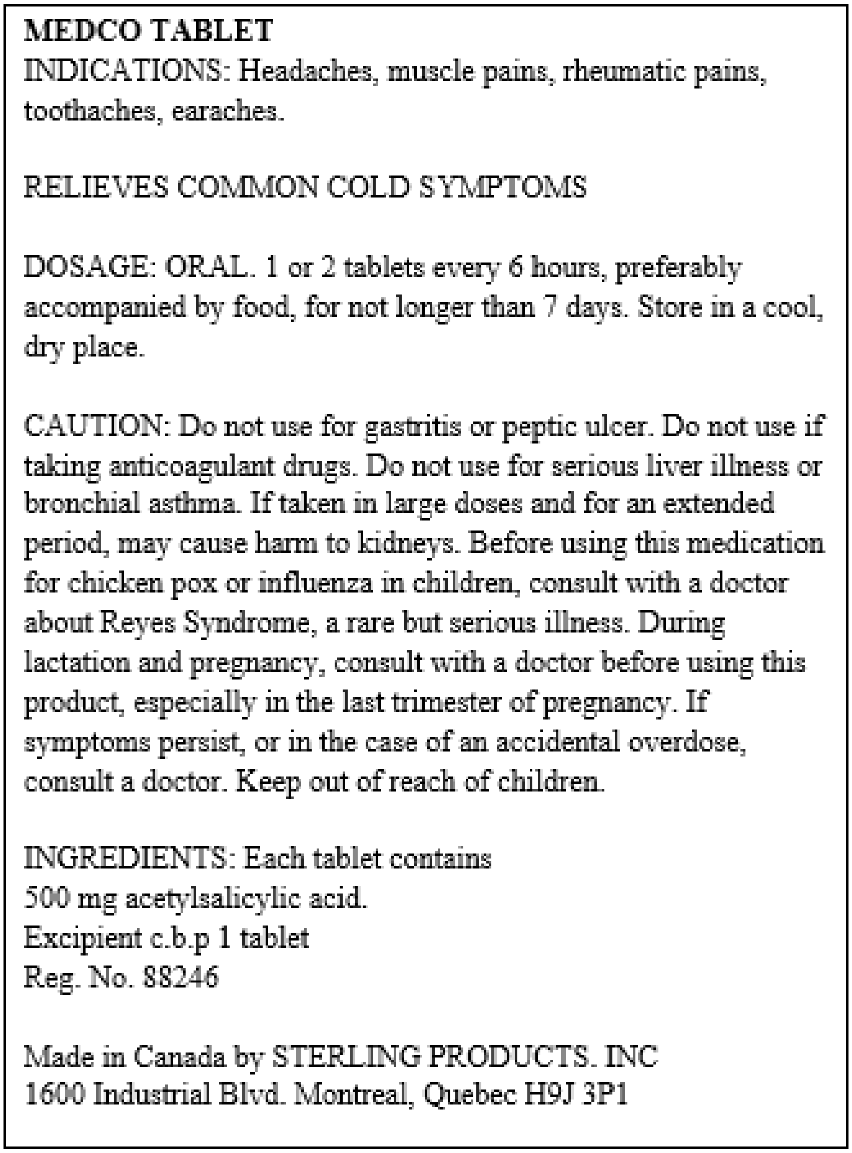
Score a point for mentioning the following or their variations: **"a person"** (1) **"who is cared for/a receives treatment/is being seen by a doctor**".

PL: Score a point for mentioning the following or their variations: **"osoba" (1) "ktora jest leczona/badana przez lekarza/zwracająca się o udzielenie świadczeń zdrowotnych** (np. badana przez lekarza)**/korzystająca ze świadczeń zdrowotnych** (np. przebywajaca w szpitalu)" (1).

Cambridge Dictionary

- *Noun*. A person who is receiving medical care, or who is cared for by a particular doctor or dentist when necessary.

Health Literacy Assessment 2: MEDCO Medicine Label



MEDCO TABLET
INDICATIONS: Headaches, muscle pains, rheumatic pains, toothaches, earaches.

RELIEVES COMMON COLD SYMPTOMS

DOSAGE: ORAL. 1 or 2 tablets every 6 hours, preferably accompanied by food, for not longer than 7 days. Store in a cool, dry place.

CAUTION: Do not use for gastritis or peptic ulcer. Do not use if taking anticoagulant drugs. Do not use for serious liver illness or bronchial asthma. If taken in large doses and for an extended period, may cause harm to kidneys. Before using this medication for chicken pox or influenza in children, consult with a doctor about Reyes Syndrome, a rare but serious illness. During lactation and pregnancy, consult with a doctor before using this product, especially in the last trimester of pregnancy. If symptoms persist, or in the case of an accidental overdose, consult a doctor. Keep out of reach of children.

INGREDIENTS: Each tablet contains
500 mg acetylsalicylic acid.
Excipient c.b.p 1 tablet
Reg. No. 88246

Made in Canada by STERLING PRODUCTS. INC
1600 Industrial Blvd. Montreal, Quebec H9J 3P1

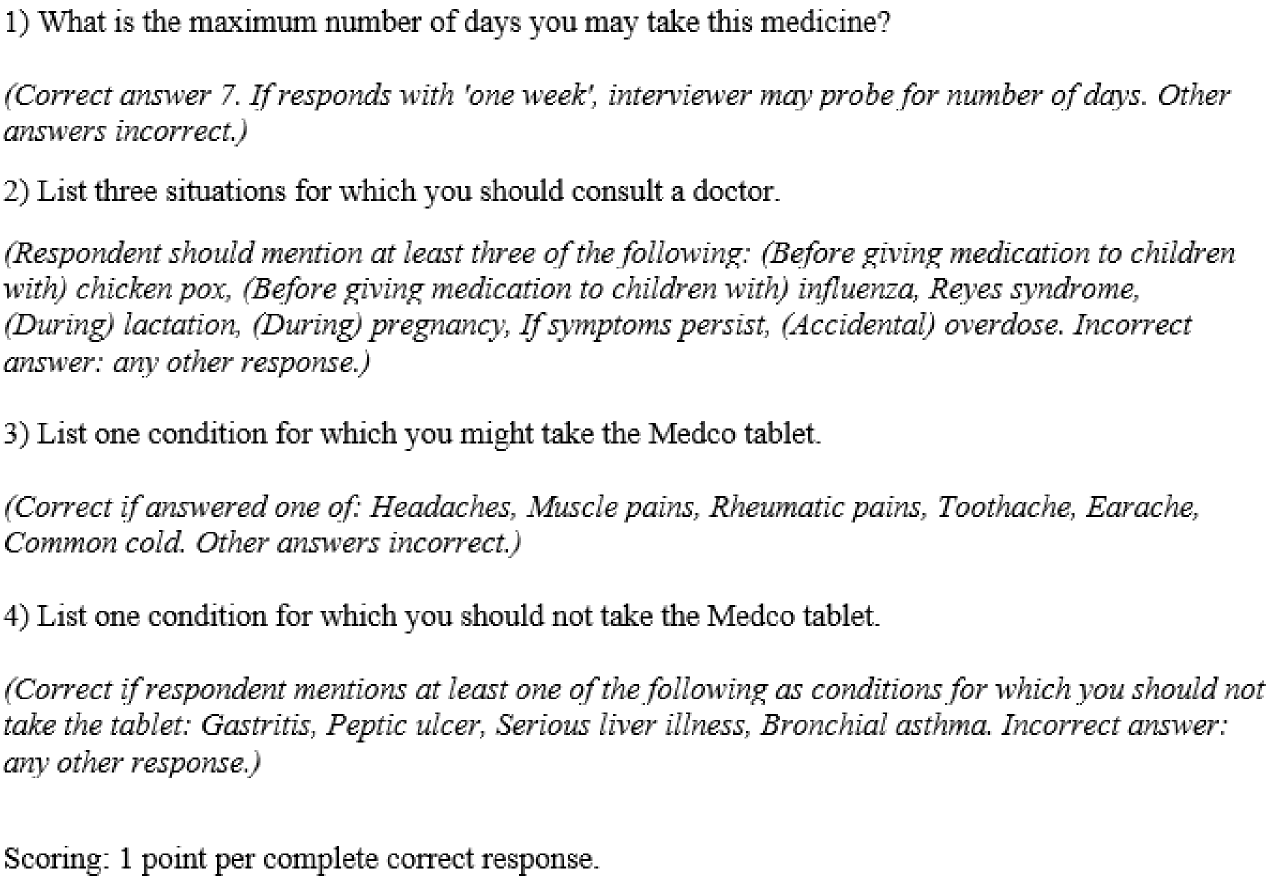
1) What is the maximum number of days you may take this medicine?

2) List three situations for which you should consult a doctor.

3) List one condition for which you might take the Medco tablet.

4) List one condition for which you should not take the Medco tablet.

Health Literacy Assessment 2: MEDCO Medicine Label (Scoring Sheet)



1) What is the maximum number of days you may take this medicine?

(Correct answer 7. If responds with 'one week', interviewer may probe for number of days. Other answers incorrect.)

2) List three situations for which you should consult a doctor.

(Respondent should mention at least three of the following: (Before giving medication to children with) chicken pox, (Before giving medication to children with) influenza, Reyes syndrome, (During) lactation, (During) pregnancy, If symptoms persist, (Accidental) overdose. Incorrect answer: any other response.)

3) List one condition for which you might take the Medco tablet.

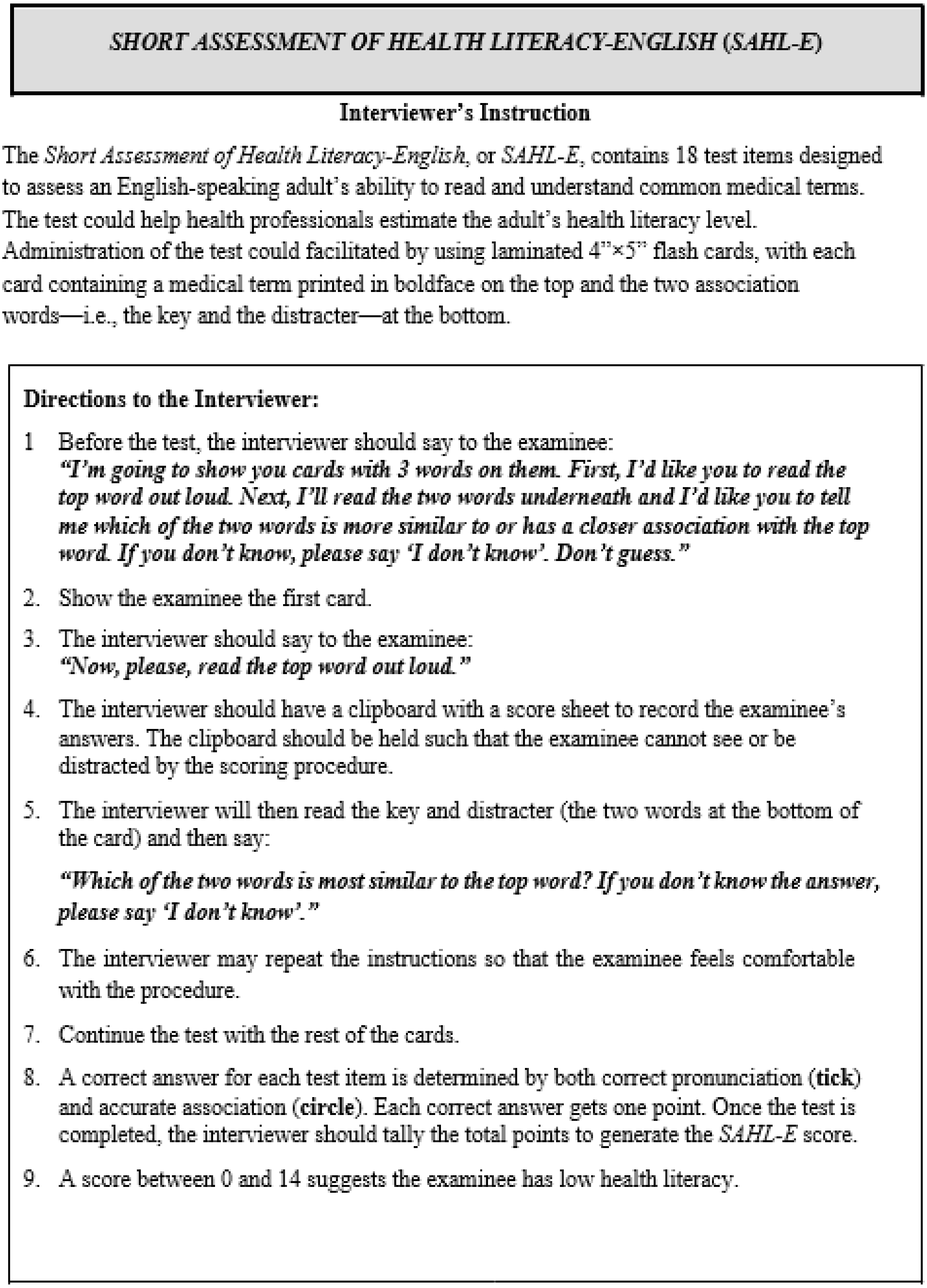
(Correct if answered one of: Headaches, Muscle pains, Rheumatic pains, Toothache, Earache, Common cold. Other answers incorrect.)

4) List one condition for which you should not take the Medco tablet.

(Correct if respondent mentions at least one of the following as conditions for which you should not take the tablet: Gastritis, Peptic ulcer, Serious liver illness, Bronchial asthma. Incorrect answer: any other response.)

Scoring: 1 point per complete correct response.

Health Literacy Assessment 3: Instruction for Administering *SAHL-E*



SHORT ASSESSMENT OF HEALTH LITERACY-ENGLISH (SAHL-E)

Interviewer's Instruction

The *Short Assessment of Health Literacy-English*, or *SAHL-E*, contains 18 test items designed to assess an English-speaking adult's ability to read and understand common medical terms. The test could help health professionals estimate the adult's health literacy level. Administration of the test could facilitated by using laminated 4"×5" flash cards, with each card containing a medical term printed in boldface on the top and the two association words—i.e., the key and the distracter—at the bottom.

Directions to the Interviewer:

1 Before the test, the interviewer should say to the examinee:
"I'm going to show you cards with 3 words on them. First, I'd like you to read the top word out loud. Next, I'll read the two words underneath and I'd like you to tell me which of the two words is more similar to or has a closer association with the top word. If you don't know, please say 'I don't know'. Don't guess."

2. Show the examinee the first card.

3. The interviewer should say to the examinee:
"Now, please, read the top word out loud."

4. The interviewer should have a clipboard with a score sheet to record the examinee's answers. The clipboard should be held such that the examinee cannot see or be distracted by the scoring procedure.

5. The interviewer will then read the key and distracter (the two words at the bottom of the card) and then say:

   "Which of the two words is most similar to the top word? If you don't know the answer, please say 'I don't know'."

6. The interviewer may repeat the instructions so that the examinee feels comfortable with the procedure.

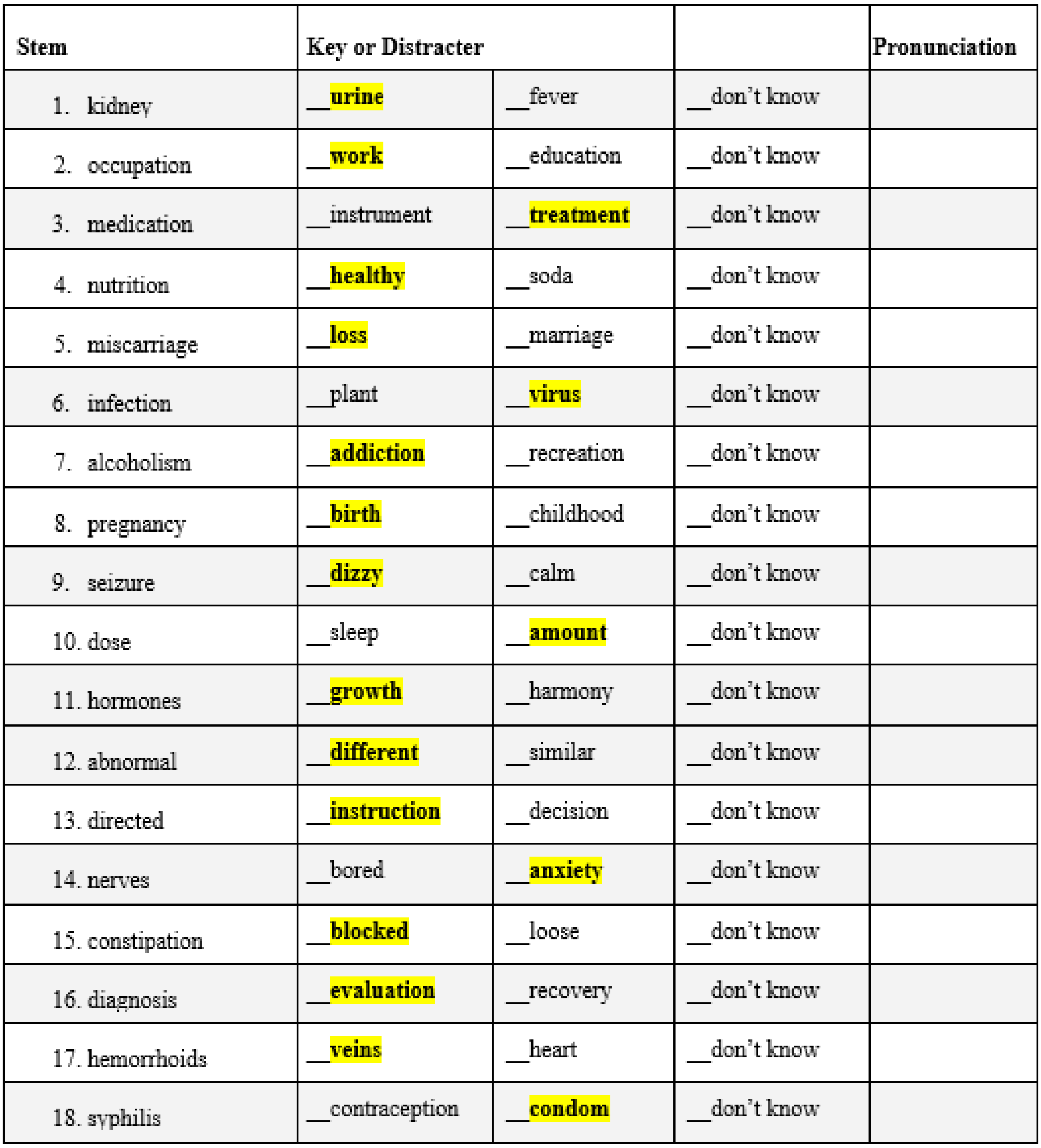
7. Continue the test with the rest of the cards.

8. A correct answer for each test item is determined by both correct pronunciation (**tick**) and accurate association (**circle**). Each correct answer gets one point. Once the test is completed, the interviewer should tally the total points to generate the *SAHL-E* score.

9. A score between 0 and 14 suggests the examinee has low health literacy.

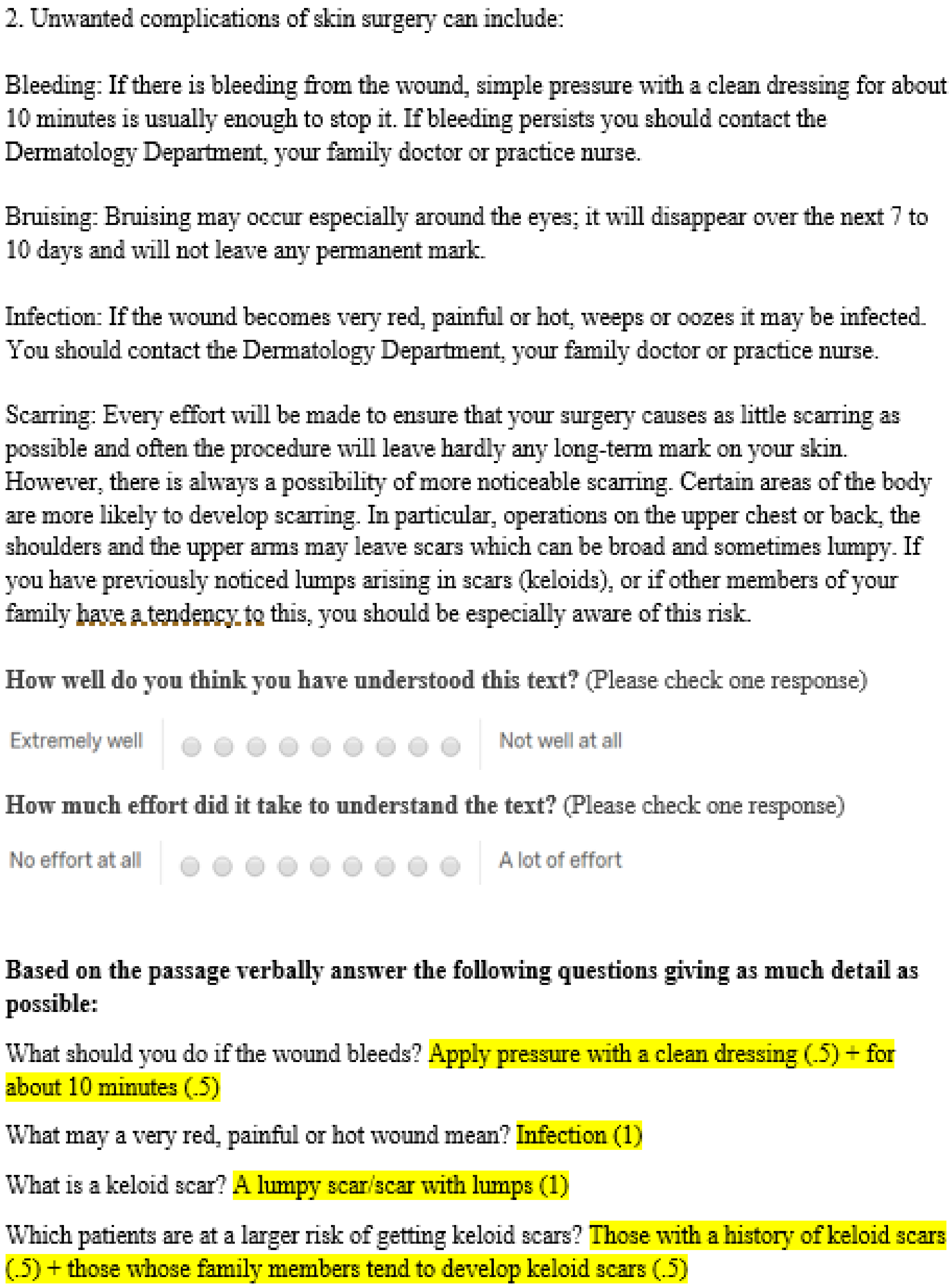
The 18 items of *SAHL-E*, ordered according to item difficulty (keys and distracters are listed in the same random order as in the field interview)

Correct answers are bolded and highlighted in yellow



Stem	**Key or Distracter**			**Pronunciation**
1. kidney	__**urine**	__fever	__don't know	
2. occupation	__**work**	__education	__don't know	
3. medication	__instrument	__**treatment**	__don't know	
4. nutrition	__**healthy**	__soda	__don't know	
5. miscarriage	__**loss**	__marriage	__don't know	
6. infection	__plant	__**virus**	__don't know	
7. alcoholism	__**addiction**	__recreation	__don't know	
8. pregnancy	__**birth**	__childhood	__don't know	
9. seizure	__**dizzy**	__calm	__don't know	
10. dose	__sleep	__**amount**	__don't know	
11. hormones	__**growth**	__harmony	__don't know	
12. abnormal	__**different**	__similar	__don't know	
13. directed	__**instruction**	__decision	__don't know	
14. nerves	__bored	__**anxiety**	__don't know	
15. constipation	__**blocked**	__loose	__don't know	
16. diagnosis	__**evaluation**	__recovery	__don't know	
17. hemorrhoids	__**veins**	__heart	__don't know	
18. syphilis	__contraception	__**condom**	__don't know	

Reading Comprehension Test (Scoring Sheet)



2. Unwanted complications of skin surgery can include:

Bleeding: If there is bleeding from the wound, simple pressure with a clean dressing for about 10 minutes is usually enough to stop it. If bleeding persists you should contact the Dermatology Department, your family doctor or practice nurse.

Bruising: Bruising may occur especially around the eyes; it will disappear over the next 7 to 10 days and will not leave any permanent mark.

Infection: If the wound becomes very red, painful or hot, weeps or oozes it may be infected. You should contact the Dermatology Department, your family doctor or practice nurse.

Scarring: Every effort will be made to ensure that your surgery causes as little scarring as possible and often the procedure will leave hardly any long-term mark on your skin. However, there is always a possibility of more noticeable scarring. Certain areas of the body are more likely to develop scarring. In particular, operations on the upper chest or back, the shoulders and the upper arms may leave scars which can be broad and sometimes lumpy. If you have previously noticed lumps arising in scars (keloids), or if other members of your family have a tendency to this, you should be especially aware of this risk.

How well do you think you have understood this text? (Please check one response)

Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How much effort did it take to understand the text? (Please check one response)

No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

Based on the passage verbally answer the following questions giving as much detail as possible:

What should you do if the wound bleeds? Apply pressure with a clean dressing (.5) + for about 10 minutes (.5)

What may a very red, painful or hot wound mean? Infection (1)

What is a keloid scar? A lumpy scar/scar with lumps (1)

Which patients are at a larger risk of getting keloid scars? Those with a history of keloid scars (.5) + those whose family members tend to develop keloid scars (.5)

If the wound becomes infected then you should:

a) Contact the Pathology Department
b) Apply a dressing to the wound
c) Contact your general practitioner (GP)

Bruising around the eyes:

a) Disappears after a week
b) Disappears after a couple of days
c) Disappears after ten minutes

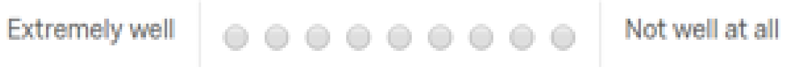
3. Treatment with Botox

Before attending for your treatment you should notify the department if you are pregnant or breast feeding. It is also important to inform us of any allergies you have, particularly to iodine, and also any medications you are taking, particularly antibiotics and any medications which have not been prescribed by your doctor. This treatment is used for the management of severe sweating under the arms (known as axillary hyperhydrosis), which does not respond to treatments with antiperspirants.

BOTOX (Botulinum Toxin A) is a bacterial toxin that temporarily weakens muscle and decreases sweating by blocking the release of certain chemicals. It is given by injections into the skin where the sweat glands are located. The BOTOX is injected into 10-15 sites of the skin of the armpit affected by excessive sweating. The area is numbed with a local anaesthetic prior to injecting the BOTOX. Due to the number of sites which are injected and the difficulty in keeping dressings in place in this area, it is advisable to wear an old dark T-shirt in case of blood staining. You will generally see an improvement within the first week following the injections and the effect usually lasts for 4-7 months. Repeat treatments will be necessary and the injections will be repeated when the effect starts to wear off.

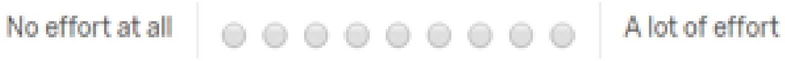
Following treatment some patients have felt that the sweating in other parts of the body increased. Allergic or inflammatory reactions in the area are rarely observed, however, bruising may occur initially. Numbness of the skin is a recognised side effect of BOTOX but it is not normally noticeable in the armpits. Other side effects are very uncommon.

How well do you think you have understood this text? (Please check one response)



Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How much effort did it take to understand the text? (Please check one response)



No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

Based on the passage verbally answer the following questions giving as much detail as possible:

What is this treatment used for? The management of severe sweating under the arms/axillary hyperhidrosis (1)

How is the BOTOX given to the patient? It is injected (.5) + into the skin where the sweat glands are located/ 10-15 sites of skin of the armpit (.5)

Why is it recommended to wear an old T-Shirt? Due to the possibility of blood staining (1)

How soon can you expect to see an improvement? Within the first week (.5) + following the injections (.5)

If you are taking any antibiotics prior to the treatment you should:

a) Tell the medical staff about them
b) Take them just before the treatment
c) Take them as prescribed

How long does the effect of the treatment last for?

a) Over a year
b) Around half a year
c) Less than two months

5. What are tonsils and adenoids?

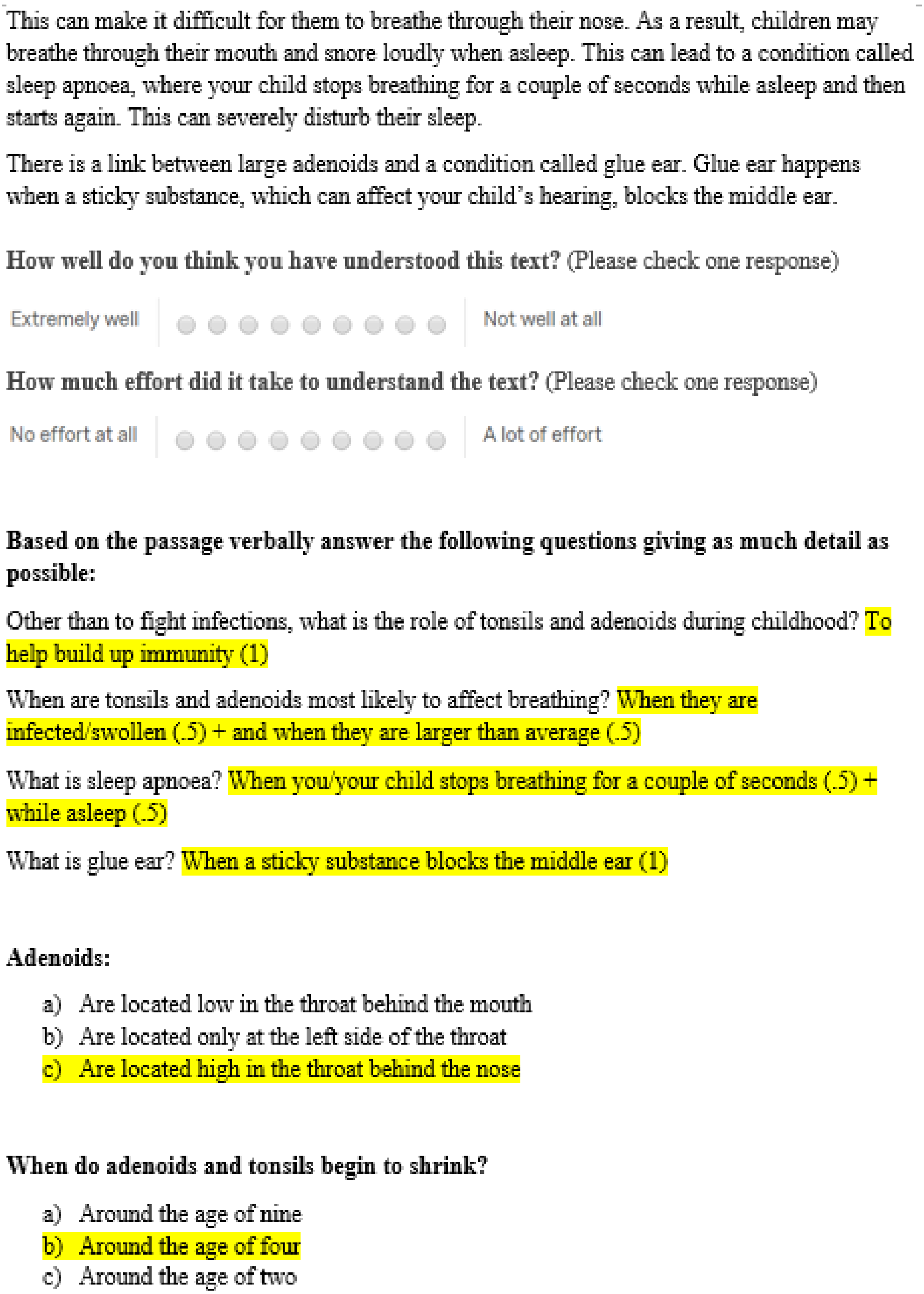
The tonsils and adenoids are areas of tissue at the back of the throat. The tonsils are on both sides of the throat, at the back of the mouth and are clearly visible. Adenoids are not visible, as they are high in the throat behind the nose.

Your child's tonsils and adenoids help him/her to build up immunity and fight infection. Adenoids and tonsils seem to grow during childhood and then shrink around the age of four years old. By the time your child reaches adulthood, his/her adenoids will have disappeared almost completely. This is because they are no longer needed, as your child's body will have other defence mechanisms to fight against infection.

Why do tonsils have to be removed; what are the benefits?

In many children, the tonsils become repeatedly infected with bacteria and viruses, which can make them swell and become painful. Removing your child's tonsils and adenoids will solve these problems.

Your child may have larger than average tonsils and adenoids, which partially block his/her airway.



This can make it difficult for them to breathe through their nose. As a result, children may breathe through their mouth and snore loudly when asleep. This can lead to a condition called sleep apnoea, where your child stops breathing for a couple of seconds while asleep and then starts again. This can severely disturb their sleep.

There is a link between large adenoids and a condition called glue ear. Glue ear happens when a sticky substance, which can affect your child's hearing, blocks the middle ear.

How well do you think you have understood this text? (Please check one response)

Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How much effort did it take to understand the text? (Please check one response)

No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

Based on the passage verbally answer the following questions giving as much detail as possible:

Other than to fight infections, what is the role of tonsils and adenoids during childhood? To help build up immunity (1)

When are tonsils and adenoids most likely to affect breathing? When they are infected/swollen (.5) + and when they are larger than average (.5)

What is sleep apnoea? When you/your child stops breathing for a couple of seconds (.5) + while asleep (.5)

What is glue ear? When a sticky substance blocks the middle ear (1)

Adenoids:

a) Are located low in the throat behind the mouth
b) Are located only at the left side of the throat
c) Are located high in the throat behind the nose

When do adenoids and tonsils begin to shrink?

a) Around the age of nine
b) Around the age of four
c) Around the age of two

7. What is epistaxis (nosebleed)?

Epistaxis is bleeding from the nose because of broken blood vessels at the front or back of the nostrils. It is usually mild and easily treated. If bleeding is more severe, it is usually in older people or in people with other medical problems.

Why has it happened?

It is not always possible to give a definite reason.

The common site for a nosebleed to start is in Little's area. This is just inside the entrance of the nostril, on the nasal septum (the middle harder part of the nostril). Here the blood vessels are quite fragile and can rupture easily for no apparent reason. This happens most commonly in children.

General advice following a nosebleed

We cannot guarantee that your nose will never bleed again.

When you go home make sure you get plenty of rest. Avoid lifting, strenuous exercise, constipation and stressful situations, as they can cause your blood pressure to rise and increase the chances of a nosebleed.

Do not blow, pick or attempt to clean the inside of your nose. The crusting discomfort you may feel is part of the healing process, and if you remove the crusts, you may infect the area or cause another nosebleed.

Will I have to stay in hospital?

If the doctor can see where the bleeding is coming from and stops the bleeding by cauterising the bleeding point, you will be allowed home. Cauterising is carried out by placing a stick with a cotton bud sized end of silver nitrate, which seals the bleeding point; this may sting for a moment. It can also be carried out using a low-level heat source to seal the bleeding point.

How well do you think you have understood this text? (Please check one response)

Extremely well ○ ○ ○ ○ ○ ○ ○ ○ ○ Not well at all

How much effort did it take to understand the text? (Please check one response)

No effort at all ○ ○ ○ ○ ○ ○ ○ ○ ○ A lot of effort

Based on the passage verbally answer the following questions giving as much detail as possible:

Name one group of people, other than the elderly, who are more likely to experience a severe nosebleed. People with other medical problems (1)

Why are nosebleeds likely to start in Little's area? Because the blood vessels there are quite fragile (.5) + and can rupture easily (.5) for no apparent reason

Why should the patient avoid raising their blood pressure following a nosebleed? To prevent another nosebleed (1)

How can nose bleeding be stopped by a doctor? Cauterising (.5) + sealing the bleeding point with silver nitrate/sealing the bleeding point with heat (.5)

Nosebleeds that start in Little's area are most common among:

a) Older adults
b) Teenagers
c) Children

Nosebleeds are:

a) Usually mild and difficult to treat
b) Often intense and not easy to treat
c) Typically light and not difficult to treat

Appendix E: Chapter 7 Tables and Figures

Table 7.2. Frequency values of medical terms used in the HLVA.

Word	**Cognate (English-Polish)**	**BNC* Frequency**	**BNC** Frequency**	**logBNC** (Zipf)**	**SUBTLEX-UK Frequency**	**logSUBTLEX-UK (Zipf)**	**BNC** Frequency (high vs. low)**	**SUBTLEX-UK Frequency (high vs. low)**
Intravenous	No	367	370	3.57	87	2.64	low	low
Suture	No	91	93	2.97	98	2.69	low	low
Biopsy	Yes	806	806	3.9	124	2.79	low	low
Antibiotic	Yes	246	246	3.39	168	2.92	low	low
Radiotherapy	Yes	210	210	3.32	211	3.02	low	low
Glucose	Yes	597	601	3.78	313	3.19	low	low
Allergy	Yes	227	225	3.35	398	3.30	low	low
Prescription	No	655	670	3.82	905	3.65	low	low
Oral	No	2323	2329	4.36	535	3.42	high	low
Injection	No	1003	1020	4.01	883	3.64	high	low
Acute	No	2273	2294	4.36	928	3.66	high	low
Anaesthetic	No	319	321	3.51	931	3.67	low	low
Symptoms	No	3110	3128	4.49	1587	3.90	high	low
Medication	No	486	487	3.69	1639	3.91	low	low
Infection	Yes	2699	2718	4.43	2281	4.05	high	high
Scan	Yes	663	667	3.82	2479	4.09	low	high
Liver	No	1633	1656	4.22	2720	4.13	high	high
Stroke	No	1591	1644	4.21	3114	4.19	high	high
Ward	No	3436	3547	4.55	3902	4.29	high	high
Cancer	No	4199	4216	4.62	7569	4.57	high	high
Pain	No	7002	7338	4.86	14165	4.85	high	high
Patient	Yes	8168	8232	4.91	10649	4.72	high	high
Mean	N/A	1913.82	1946.27	4.01	2531.18	3.7	high	low

Notes. * As reported by the BNC Consortium (2007); ** as reported by van Heuven et al. (2014). High vs. low frequencies were determined in accordance with criteria reported by van Heuven et al. (2014), where Log-Zipf-frequencies below 4 were rated as relatively low-frequency words, whereas those above 4 as relatively high-frequency words. The Zipf scale is a relatively new measure of word frequencies, where a Zipf value of 1 corresponds to words with frequencies of 1 per 100 million words, a Zipf value of 2 corresponds to words with

Table 7.11. Interaction effects of the optimal model (Model 19.1).

Coefficients	Estimate	Est.Error	L-95%	U-95%	Estimate OR	L-95% OR	U-95% OR	Probable (sign)
Education level:RDL2	.00	2.29	-4.48	4.49	1.00	.01	89.12	
Education level:FRE	.00	2.26	-4.48	4.45	1.00	.01	85.63	
Education level:Temporal connectives	-.01	2.26	-4.47	4.41	.99	.01	82.27	
Education level:All connectives	.00	2.32	-4.56	4.56	1.00	.01	95.58	
Education level:Stem overlap	-.02	2.23	-4.34	4.35	.98	.01	77.48	
Education level:Hypernymy noun	.01	2.28	-4.49	4.43	1.01	.01	83.93	
Education level:Hypernymy verb	-.01	2.25	-4.40	4.43	.99	.01	83.93	
Education level:Deep cohesion	.01	2.22	-4.35	4.31	1.01	.01	74.44	
Education level:Referential cohesion	-.01	2.35	-4.59	4.57	.99	.01	96.54	
Education level:Causal connectives	-.03	2.28	-4.51	4.45	.97	.01	85.63	
Education level:CELEX frequency	-.01	2.30	-4.48	4.48	.99	.01	88.23	
Education level:Sentence length	.03	2.30	-4.46	4.60	1.03	.01	99.48	
Education level:Passive voice	-.02	2.25	-4.41	4.43	.98	.01	83.93	
Education level:Syntax similarity	.01	2.30	-4.51	4.52	1.01	.01	91.84	
Education level:Causal cohesion	.00	2.30	-4.56	4.48	1.00	.01	88.23	
Education level:Logical connectives	.01	2.32	-4.54	4.56	1.01	.01	95.58	
Education level:Gerunds	.00	2.32	-4.52	4.48	1.00	.01	88.23	
Education level:BNC frequency	-.01	2.34	-4.53	4.56	.99	.01	95.58	
English proficiency:RDL2	-.13	2.30	-4.65	4.34	.88	.01	76.71	
English proficiency:FRE	-.13	2.27	-4.58	4.35	.88	.01	77.48	
English proficiency:Temporal connectives	.06	2.24	-4.32	4.45	1.06	.01	85.63	
English proficiency:All connectives	.00	2.34	-4.57	4.61	1.00	.01	100.48	
English proficiency:Stem overlap	.02	2.25	-4.40	4.42	1.02	.01	83.10	
English proficiency:Hypernymy noun	.10	2.33	-4.45	4.67	1.11	.01	106.70	
English proficiency:Hypernymy verb	-.09	2.26	-4.46	4.31	.91	.01	74.44	
English proficiency:Deep cohesion	.00	2.23	-4.33	4.39	1.00	.01	80.64	
English proficiency:Referential cohesion	-.07	2.28	-4.56	4.41	.93	.01	82.27	
English proficiency:Causal connectives	-.14	2.32	-4.69	4.41	.87	.01	82.27	
English proficiency:CELEX frequency	-.12	2.30	-4.62	4.39	.89	.01	80.64	

English proficiency:Sentence length	.14	2.25	-4.32	4.56	1.15	.01	95.58
English proficiency:Passive voice	.07	2.27	-4.34	4.47	1.07	.01	87.36
English proficiency:Syntax similarity	-.02	2.31	-4.47	4.50	.98	.01	90.02
English proficiency:Causal cohesion	.00	2.28	-4.37	4.48	1.00	.01	88.23
English proficiency:Logical connectives	.14	2.28	-4.37	4.63	1.15	.01	102.51
English proficiency:Gerunds	.06	2.31	-4.49	4.59	1.06	.01	98.49
English proficiency:BNC frequency	.05	2.33	-4.51	4.61	1.05	.01	100.48
HLVA:RDL2	.02	2.35	-4.59	4.59	1.02	.01	98.49
HLVA:FRE	-.01	2.26	-4.47	4.45	.99	.01	85.63
HLVA:Temporal connectives	.05	2.23	-4.27	4.46	1.05	.01	86.49
HLVA:All connectives	-.02	2.34	-4.63	4.59	.98	.01	98.49
HLVA:Stem overlap	.08	2.25	-4.33	4.52	1.08	.01	91.84
HLVA:Hypernymy noun	-.05	2.30	-4.58	4.48	.95	.01	88.23
HLVA:Hypernymy verb	-.03	2.25	-4.47	4.39	.97	.01	80.64
HLVA:Deep cohesion	-.05	2.25	-4.43	4.37	.95	.01	79.04
HLVA:Referential cohesion	.05	2.32	-4.52	4.55	1.05	.01	94.63
HLVA:Causal connectives	.03	2.28	-4.39	4.49	1.03	.01	89.12
HLVA:CELEX frequency	.03	2.31	-4.50	4.56	1.03	.01	95.58
HLVA:Sentence length	.01	2.26	-4.39	4.43	1.01	.01	83.93
HLVA:Passive voice	.04	2.26	-4.41	4.51	1.04	.01	90.92
HLVA:Syntax similarity	.04	2.30	-4.50	4.61	1.04	.01	100.48
HLVA:Causal cohesion	-.02	2.31	-4.54	4.49	.98	.01	89.12
HLVA:Logical connectives	-.06	2.34	-4.65	4.55	.94	.01	94.63
HLVA:Gerunds	.01	2.32	-4.54	4.54	1.01	.01	93.69
HLVA:BNC frequency	.01	2.30	-4.48	4.50	1.01	.01	90.02
Age:RDL2	-.02	2.31	-4.56	4.53	.98	.01	92.76
Age:FRE	-.04	2.27	-4.44	4.44	.96	.01	84.77
Age:Temporal connectives	.04	2.22	-4.30	4.42	1.04	.01	83.10
Age:All connectives	-.07	2.31	-4.55	4.43	.93	.01	83.93
Age:Stem overlap	-.04	2.25	-4.47	4.38	.96	.01	79.84
Age:Hypernymy noun	.05	2.33	-4.52	4.67	1.05	.01	106.70
Age:Hypernymy verb	.03	2.27	-4.41	4.44	1.03	.01	84.77

Age:Deep cohesion	-.04	2.24	-4.40	4.30	.96	.01	73.70
Age:Referential cohesion	-.03	2.33	-4.60	4.54	.97	.01	93.69
Age:Causal connectives	.01	2.31	-4.51	4.51	1.01	.01	90.92
Age:CELEX frequency	.04	2.27	-4.42	4.51	1.04	.01	90.92
Age:Sentence length	.00	2.30	-4.54	4.51	1.00	.01	90.92
Age:Passive voice	.06	2.27	-4.44	4.53	1.06	.01	92.76
Age:Syntax similarity	-.06	2.30	-4.52	4.45	.94	.01	85.63
Age:Causal cohesion	-.06	2.27	-4.53	4.38	.94	.01	79.84
Age:Logical connectives	.00	2.31	-4.53	4.58	1.00	.01	97.51
Age:Gerunds	.08	2.32	-4.46	4.66	1.08	.01	105.64
Age:BNC frequency	.07	2.34	-4.47	4.65	1.07	.01	104.58
WM:RDL2	-.05	2.33	-4.62	4.53	.95	.01	92.76
WM:FRE	-.06	2.28	-4.51	4.40	.94	.01	81.45
WM:Temporal connectives	.04	2.24	-4.33	4.39	1.04	.01	80.64
WM:All connectives	.00	2.28	-4.46	4.47	1.00	.01	87.36
WM:Stem overlap	.04	2.26	-4.33	4.46	1.04	.01	86.49
WM:Hypernymy noun	-.02	2.30	-4.51	4.49	.98	.01	89.12
WM:Hypernymy verb	-.04	2.25	-4.43	4.36	.96	.01	78.26
WM:Deep cohesion	-.02	2.24	-4.37	4.39	.98	.01	80.64
WM:Referential cohesion	.01	2.31	-4.56	4.56	1.01	.01	95.58
WM:Causal connectives	-.02	2.27	-4.51	4.40	.98	.01	81.45
WM:CELEX frequency	-.02	2.30	-4.52	4.48	.98	.01	88.23
WM:Sentence length	.05	2.29	-4.48	4.56	1.05	.01	95.58
WM:Passive voice	.08	2.27	-4.36	4.52	1.08	.01	91.84
WM:Syntax similarity	.01	2.30	-4.53	4.55	1.01	.01	94.63
WM:Causal cohesion	-.05	2.27	-4.52	4.43	.95	.01	83.93
WM:Logical connectives	.01	2.29	-4.48	4.51	1.01	.01	90.92
WM:Gerunds	.02	2.31	-4.48	4.55	1.02	.01	94.63
WM:BNC frequency	.02	2.32	-4.55	4.56	1.02	.01	95.58
Phonology:RDL2	.11	2.30	-4.41	4.61	1.12	.01	100.48
Phonology:FRE	.09	2.29	-4.40	4.60	1.09	.01	99.48
Phonology:Temporal connectives	-.03	2.25	-4.40	4.35	.97	.01	77.48

Phonology:All connectives	.00	2.30	-4.50	4.46	1.00	.01	86.49
Phonology:Stem overlap	-.04	2.26	-4.42	4.39	.96	.01	80.64
Phonology:Hypernymy noun	-.05	2.29	-4.58	4.41	.95	.01	82.27
Phonology:Hypernymy verb	.05	2.25	-4.33	4.43	1.05	.01	83.93
Phonology:Deep cohesion	-.01	2.23	-4.35	4.29	.99	.01	72.97
Phonology:Referential cohesion	.06	2.31	-4.48	4.59	1.06	.01	98.49
Phonology:Causal connectives	.06	2.30	-4.46	4.56	1.06	.01	95.58
Phonology:CELEX frequency	.06	2.30	-4.42	4.57	1.06	.01	96.54
Phonology:Sentence length	-.12	2.27	-4.55	4.31	.89	.01	74.44
Phonology:Passive voice	-.05	2.27	-4.55	4.38	.95	.01	79.84
Phonology:Syntax similarity	.00	2.29	-4.47	4.49	1.00	.01	89.12
Phonology:Causal cohesion	-.01	2.25	-4.40	4.43	.99	.01	83.93
Phonology:Logical connectives	-.07	2.25	-4.47	4.39	.93	.01	80.64
Phonology:Gerunds	-.02	2.31	-4.55	4.48	.98	.01	88.23
Phonology:BNC frequency	-.03	2.32	-4.61	4.52	.97	.01	91.84
MEDCO:RDL2	.01	2.31	-4.57	4.62	1.01	.01	101.49
MEDCO:FRE	-.02	2.25	-4.38	4.38	.98	.01	79.84
MEDCO:Temporal connectives	.02	2.23	-4.34	4.36	1.02	.01	78.26
MEDCO:All connectives	.00	2.29	-4.47	4.51	1.00	.01	90.92
MEDCO:Stem overlap	.01	2.24	-4.34	4.37	1.01	.01	79.04
MEDCO:Hypernymy noun	-.02	2.32	-4.54	4.54	.98	.01	93.69
MEDCO:Hypernymy verb	-.01	2.29	-4.47	4.48	.99	.01	88.23
MEDCO:Deep cohesion	-.01	2.24	-4.37	4.37	.99	.01	79.04
MEDCO:Referential cohesion	.01	2.32	-4.53	4.54	1.01	.01	93.69
MEDCO:Causal connectives	.02	2.27	-4.39	4.41	1.02	.01	82.27
MEDCO:CELEX frequency	.01	2.27	-4.47	4.47	1.01	.01	87.36
MEDCO:Sentence length	.00	2.27	-4.43	4.36	1.00	.01	78.26
MEDCO:Passive voice	.02	2.25	-4.40	4.48	1.02	.01	88.23
MEDCO:Syntax similarity	.00	2.31	-4.51	4.52	1.00	.01	91.84
MEDCO:Causal cohesion	-.03	2.30	-4.51	4.42	.97	.01	83.10
MEDCO:Logical connectives	-.02	2.28	-4.49	4.48	.98	.01	88.23
MEDCO:Gerunds	-.01	2.31	-4.54	4.52	.99	.01	91.84

MEDCO:BNC frequency	.00	2.31	-4.55	4.50	1.00	.01	90.02
Metacomprehension:RDL2	-.04	2.31	-4.52	4.52	.96	.01	91.84
Metacomprehension:FRE	-.03	2.26	-4.43	4.42	.97	.01	83.10
Metacomprehension:Temporal connectives	-.01	2.21	-4.31	4.27	.99	.01	71.52
Metacomprehension:All connectives	-.02	2.29	-4.49	4.46	.98	.01	86.49
Metacomprehension:Stem overlap	-.01	2.23	-4.40	4.36	.99	.01	78.26
Metacomprehension:Hypernymy noun	.03	2.31	-4.52	4.56	1.03	.01	95.58
Metacomprehension:Hypernymy verb	-.04	2.25	-4.40	4.38	.96	.01	79.84
Metacomprehension:Deep cohesion	.01	2.22	-4.33	4.35	1.01	.01	77.48
Metacomprehension:Referential cohesion	-.04	2.34	-4.62	4.54	.96	.01	93.69
Metacomprehension:Causal connectives	-.05	2.32	-4.62	4.46	.95	.01	86.49
Metacomprehension:CELEX frequency	-.04	2.28	-4.52	4.45	.96	.01	85.63
Metacomprehension:Sentence length	.04	2.26	-4.42	4.47	1.04	.01	87.36
Metacomprehension:Passive voice	.01	2.26	-4.37	4.39	1.01	.01	80.64
Metacomprehension:Syntax similarity	-.02	2.28	-4.46	4.49	.98	.01	89.12
Metacomprehension:Causal cohesion	.01	2.26	-4.43	4.43	1.01	.01	83.93
Metacomprehension:Logical connectives	.03	2.28	-4.46	4.52	1.03	.01	91.84
Metacomprehension:Gerunds	.03	2.29	-4.42	4.51	1.03	.01	90.92
Metacomprehension:BNC frequency	.02	2.29	-4.47	4.46	1.02	.01	86.49
THLQ1:RDL2	.05	2.29	-4.43	4.54	1.05	.01	93.69
THLQ1:FRE	.02	2.25	-4.39	4.39	1.02	.01	80.64
THLQ1:Temporal connectives	-.06	2.24	-4.45	4.31	.94	.01	74.44
THLQ1:All connectives	.06	2.32	-4.48	4.62	1.06	.01	101.49
THLQ1:Stem overlap	-.02	2.23	-4.40	4.33	.98	.01	75.94
THLQ1:Hypernymy noun	-.01	2.29	-4.47	4.48	.99	.01	88.23
THLQ1:Hypernymy verb	-.02	2.25	-4.42	4.39	.98	.01	80.64
THLQ1:Deep cohesion	.06	2.23	-4.32	4.43	1.06	.01	83.93
THLQ1:Referential cohesion	.03	2.27	-4.44	4.51	1.03	.01	90.92
THLQ1:Causal connectives	-.04	2.29	-4.53	4.44	.96	.01	84.77
THLQ1:CELEX frequency	-.03	2.29	-4.50	4.44	.97	.01	84.77
THLQ1:Sentence length	.00	2.28	-4.44	4.46	1.00	.01	86.49
THLQ1:Passive voice	-.07	2.23	-4.44	4.31	.93	.01	74.44

THLQ1:Syntax similarity	.02	2.24	-4.33	4.43	1.02	.01	83.93
THLQ1:Causal cohesion	.07	2.28	-4.39	4.47	1.07	.01	87.36
THLQ1:Logical connectives	.02	2.30	-4.47	4.52	1.02	.01	91.84
THLQ1:Gerunds	-.06	2.32	-4.58	4.49	.94	.01	89.12
THLQ1:BNC frequency	-.08	2.30	-4.58	4.50	.92	.01	90.02

Note 1 . OR refers to Odds Ratio. *Note 2* . English proficiency variable constitutes self-assessed English language proficiency, English language vocabulary, and a vocabulary-based assessment of health literacy (Chapter 7, section 7.2.3.iv). *Note 3.* Metacomprehension variable constitutes of self-rated perceived understanding of, and perceived effort required to understand, health-related texts (Chapter 7, section 7.2.3.iv). *Note 4* . HLVA is health literacy vocabulary assessment; WM is working memory; MEDCO is a medicine-label-based health literacy assessment; THLQ1 is a screening question used to rapidly assess health literacy; RDL2 is Coh-Metrix L2 Readability Index (Crossley et al., 2008); and FRE is Flesch Reading Ease (Flesch, 1948).

Table 7.12. Random effects table of the optimal model (Model 19.1).

		Estimate	Est.Error	L-95%	U-95%
Standard deviation (Question)	sd(Intercept)	2.28	.38	1.67	3.15
	sd(Education level)	.30	.16	.03	.64
	sd(English proficiency)	1.25	.32	.70	1.96
	sd(HLVA)	.53	.24	.07	1.04
	sd(Age)	.34	.19	.03	.73
	sd(WM)	.26	.16	.01	.62
	sd(Phonology)	.41	.21	.04	.85
	sd(MEDCO)	.66	.18	.33	1.06
	sd(Metacomprehension)	.28	.20	.01	.76
	sd(THLQ1)	.34	.18	.03	.73
Correlations	cor(Intercept,Education level)	-.08	.27	-.59	.45
	cor(Intercept,English proficiency)	-.11	.21	-.51	.31
	cor(Education level,English proficiency)	-.18	.27	-.66	.36
	cor(Intercept,HLVA)	.05	.26	-.45	.53
	cor(Education level,HLVA)	-.16	.29	-.67	.42
	cor(English proficiency,HLVA)	.14	.26	-.38	.64
	cor(Intercept,Age)	.03	.26	-.48	.53
	cor(Education level,Age)	.08	.29	-.49	.61
	cor(English proficiency,Age)	-.28	.27	-.74	.31
	cor(HLVA,Age)	.03	.28	-.52	.57
	cor(Intercept,WM)	-.01	.28	-.55	.53
	cor(Education level,WM)	-.01	.29	-.57	.55
	cor(English proficiency,WM)	.13	.28	-.45	.64
	cor(HLVA,WM)	-.02	.29	-.58	.54
	cor(Age,WM)	-.02	.30	-.58	.55
	cor(Intercept,Phonology)	.16	.27	-.38	.64
	cor(Education level,Phonology)	-.10	.29	-.62	.48
	cor(English proficiency,Phonology)	-.04	.27	-.54	.49
	cor(HLVA,Phonology)	-.06	.28	-.59	.50
	cor(Age,Phonology)	.01	.29	-.54	.56
	cor(WM,Phonology)	-.09	.30	-.63	.50
	cor(Intercept,MEDCO)	-.19	.23	-.59	.27
	cor(Education level,MEDCO)	.06	.27	-.48	.57
	cor(English proficiency,MEDCO)	.07	.24	-.40	.53
	cor(HLVA,MEDCO)	.12	.26	-.40	.60
	cor(Age,MEDCO)	-.06	.27	-.58	.47
	cor(WM,MEDCO)	.12	.28	-.44	.63
	cor(Phonology,MEDCO)	-.24	.27	-.71	.32
	cor(Intercept,Metacomprehension)	-.14	.29	-.66	.46
	cor(Education level,Metacomprehension)	-.04	.30	-.60	.54
	cor(English proficiency,Metacomprehension)	.09	.29	-.49	.62
	cor(HLVA,Metacomprehension)	.00	.30	-.58	.56
	cor(Age,Metacomprehension)	-.11	.30	-.65	.49
	cor(WM,Metacomprehension)	.02	.30	-.56	.58
	cor(Phonology,Metacomprehension)	.02	.30	-.55	.58
	cor(MEDCO,Metacomprehension)	-.04	.29	-.58	.52
	cor(Intercept,THLQ1)	-.13	.28	-.63	.43
	cor(Education level,THLQ1)	.00	.29	-.56	.55
	cor(English proficiency,THLQ1)	.05	.28	-.49	.57
	cor(HLVA,THLQ1)	.12	.29	-.46	.65
	cor(Age,THLQ1)	.02	.29	-.54	.57
	cor(WM,THLQ1)	.02	.29	-.54	.58
	cor(Phonology,THLQ1)	-.17	.29	-.68	.41
	cor(MEDCO,THLQ1)	.29	.27	-.30	.74
	cor(Metacomprehension,THLQ1)	-.03	.30	-.59	.54

Standard deviation (Subject)	sd(Intercept)	.18	.12	.01	.44
	sd(Education level)	.82	.21	.36	1.22
	sd(English proficiency)	.35	.25	.02	.93
	sd(HLVA)	.28	.21	.01	.76
	sd(Age)	.33	.22	.01	.81
	sd(WM)	.37	.24	.02	.88
	sd(Phonology)	.29	.21	.01	.78
	sd(MEDCO)	.42	.25	.02	.92
	sd(Metacomprehension)	.49	.31	.03	1.14
	sd(THLQ1)	.32	.21	.01	.79
Correlations	cor(Intercept,Education level)	.09	.29	-.49	.62
	cor(Intercept,English proficiency)	.09	.31	-.52	.65
	cor(Education level,English proficiency)	.01	.29	-.56	.57
	cor(Intercept,HLVA)	.03	.30	-.56	.59
	cor(Education level,HLVA)	.00	.30	-.57	.59
	cor(English proficiency,HLVA)	-.04	.30	-.61	.55
	cor(Intercept,Age)	.06	.30	-.53	.62
	cor(Education level,Age)	.03	.29	-.54	.58
	cor(English proficiency,Age)	-.01	.30	-.57	.56
	cor(HLVA,Age)	-.04	.30	-.60	.55
	cor(Intercept,WM)	-.03	.29	-.59	.54
	cor(Education level,WM)	-.08	.28	-.60	.50
	cor(English proficiency,WM)	-.04	.30	-.61	.55
	cor(HLVA,WM)	-.01	.30	-.58	.57
	cor(Age,WM)	.03	.30	-.56	.59
	cor(Intercept,Phonology)	.04	.30	-.55	.61
	cor(Education level,Phonology)	-.01	.29	-.57	.56
	cor(English proficiency,Phonology)	-.04	.30	-.60	.55
	cor(HLVA,Phonology)	-.03	.30	-.60	.56
	cor(Age,Phonology)	.01	.30	-.56	.59
	cor(WM,Phonology)	-.08	.31	-.64	.53
	cor(Intercept,MEDCO)	-.01	.30	-.58	.57
	cor(Education level,MEDCO)	.08	.29	-.50	.62
	cor(English proficiency,MEDCO)	-.03	.30	-.59	.55
	cor(HLVA,MEDCO)	-.02	.30	-.58	.56
	cor(Age,MEDCO)	.13	.30	-.49	.68
	cor(WM,MEDCO)	-.12	.31	-.67	.50
	cor(Phonology,MEDCO)	.00	.30	-.57	.57
	cor(Intercept,Metacomprehension)	.09	.30	-.50	.63
	cor(Education level,Metacomprehension)	.06	.28	-.51	.59
	cor(English proficiency,Metacomprehension)	-.07	.31	-.64	.54
	cor(HLVA,Metacomprehension)	-.07	.30	-.63	.53
	cor(Age,Metacomprehension)	-.03	.30	-.59	.56
	cor(WM,Metacomprehension)	-.04	.30	-.60	.55
	cor(Phonology,Metacomprehension)	-.04	.30	-.60	.54
	cor(MEDCO,Metacomprehension)	-.08	.30	-.63	.51
	cor(Intercept,THLQ1)	-.01	.30	-.57	.56
	cor(Education level,THLQ1)	-.15	.29	-.67	.46
	cor(English proficiency,THLQ1)	-.02	.30	-.59	.56
	cor(HLVA,THLQ1)	-.04	.30	-.60	.55
	cor(Age,THLQ1)	-.01	.30	-.59	.56
	cor(WM,THLQ1)	.02	.30	-.56	.58
	cor(Phonology,THLQ1)	-.02	.30	-.59	.56
	cor(MEDCO,THLQ1)	-.08	.30	-.63	.53
	cor(Metacomprehension,THLQ1)	-.08	.30	-.64	.52

Table 7.13. Sensitivity Checks (Priors).

Model number	Priors	Probable effects (sign)	Highest Rhat	LOOIC	Chains	Notes
1	Predictors(Cauchy(0, 2.5)), Intercept(Cauchy(0, 10)), Random effects(Gamma(2, .01)), Covariance(LKJ(1))	English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4038.09	6 chains with 4000 iterations per chain	There were 10554 divergent transitions after warmup.
2	Predictors(Cauchy(0, 2.5)), Intercept(Cauchy(0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(1))	English language proficiency (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.01	4048.93	6 chains with 4000 iterations per chain	There were 12000 divergent transitions after warmup.
3	Predictors(Cauchy(0, 2.5)), Intercept(Cauchy(0, 10)), Random effects(Gamma(1.5, .01)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4033.33	6 chains with 4000 iterations per chain	There were 11706 divergent transitions after warmup.
4	Predictors(Cauchy(0, 2.5)), Intercept(Cauchy(0, 10)), Random effects(Cauchy(0, 2.5)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.04	4029.73	6 chains with 4000 iterations per chain	There were 10000 divergent transitions after warmup.
5	Predictors(Cauchy(0, .75)), Intercept(Cauchy(0, 10)), Random effects(Gamma(2, .01)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.01	4036.07	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
6	Predictors(Cauchy(0, .75)), Intercept(Cauchy(0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(1))	English language proficiency (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.01	4047.74	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.

7	Predictors(Cauchy(0, .75)), Intercept(Cauchy(0, 10)), Random effects(Gamma(1.5, .01)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4032.18	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
8	Predictors(Cauchy(0, .75)), Intercept(Cauchy(0, 10)), Random effects(Cauchy(0, .75)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4027.70	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
9	Predictors(Normal(0, 2.5)), Intercept(Normal(0, 10)), Random effects(Gamma(2, .01)), Covariance(LKJ(1))	English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4038.04	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
10	Predictors(Normal(0, 2.5)), Intercept(Normal(0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(1))	English language proficiency (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4047.59	6 chains with 4000 iterations per chain	There were 1 divergent transitions after warmup.
11	Predictors(Normal(0, 2.5)), Intercept(Normal(0, 10)), Random effects(Gamma(1.5, .01)), Covariance(LKJ(1))	English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.01	4033.97	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
12	Predictors(Normal(0, 2.5)), Intercept(Normal(0, 10)), Random effects(Cauchy(0, 2.5)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4030.31	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.

13	Predictors(Normal(0, 2.5)), Intercept(Normal(0, 10)), Random effects(Cauchy(0, .75)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4028.51	6 chains with 4000 iterations per chain	There were 1 divergent transitions after warmup.
14	Predictors(Student_t(7, 0, 2.5)), Intercept(Student_t(7, 0, 10)), Random effects(Gamma(2, .01)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4037.47	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
15	Predictors(Student_t(7, 0, 2.5)), Intercept(Student_t(7, 0, 10)), Random effects(Gamma(3, .01)), Covariance(LKJ(1))	English language proficiency (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4048.56	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
16	Predictors(Student_t(7, 0, 2.5)), Intercept(Student_t(7, 0, 10)), Random effects(Gamma(1.5, .01)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4032.70	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
17	Predictors(Student_t(7, 0, 2.5)), Intercept(Student_t(7, 0, 10)), Random effects(Cauchy(0, .75)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4029.26	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
18	Predictors(Student_t(7, 0, 2.5)), Intercept(Student_t(7, 0, 10)), Random effects(Cauchy(0, 2.5)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.01	4032.10	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.

19*	Predictors(Normal(0, 2.5)), Intercept(Normal(.5, .5)), Random effects(Cauchy(0, 2.5)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4030.70	6 chains with 4000 iterations per chain	No transitions beyond maximum treedepth.
19.1**	Predictors(Normal(0, 2.5)), Intercept(Normal(.5, .5)), Random effects(Cauchy(0, 2.5)), Covariance(LKJ(1))	Education level (+), English language proficiency (+), HLVA; health literacy (+), Age (-), MEDCO; health literacy (+), THLQ1; screening question (+)	1.00	4030.70	6 chains with 8000 iterations per chain	There were 3 divergent transitions after warmup.

Notes. *The chosen model. **Local convergence check model (doubling the number of iterations).