A comparative study of reading comprehension between screen and print for L1-English and L1-Chinese-speaking university students

Handan Lu

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
School of Education

April, 2020

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to thank my family and friends who have supported me throughout my time in University. In particular, I would like to thank my parents, who have always believed in my abilities and encouraged me to strive towards my goals. It is difficult to express the depth of my appreciation to my husband, Zeyu, for his support, blessings, warmth, and encouragement throughout my PhD journey.

I owe special thanks to my supervisors, Dr Richard Badger, Dr Paula Clarke, and Dr Shirley-Anne Paul (in the first three years) for their intellectual support and for giving me the motivation I needed to complete this thesis. In fact, without their encouragement, I would not have contemplated embarking upon a PhD. To them, I am eternally grateful.

This thesis is dedicated to the memory of my stillbirth child, Yumeng He.

Abstract

This study explored the differences that might exist in print and screen reading comprehension performance for the first year L1-English and L1-Chinese-speaking university students in an Anglophone academic context. It examined reading comprehension by taking into account literal and inferential dimensions of reading. It also explored the working memory and how it might contribute to those differences, according to the Multi-component Model of working memory by Baddeley and Hitch (2000). The research used a mixed method approach and combined two quantitative experimental tests (Reading Comprehension Assessment and Working Memory Capacity Test) and a qualitative interview. Results suggested that the medium had a significant effect on reading comprehension performance, with participants achieving higher scores after reading on paper than reading on screen. However, whilst higher scores were obtained in the print condition than in the screen condition for the inferential questions, no such difference was apparent for the literal comprehension questions. This pattern was found to be consistent with both L1-English and L1-Chinese speakers. Further, the capacity of working memory measured by the composite score of both complex and simple span tasks had a significant role in reading comprehension across two medium. In particular, the working memory capacity measured by complex span tasks had a meaningful contribution to inferential level comprehension across two medium; but at literal level such connection was only found with screen. This study has made important theoretical and methodological contributions to the reading research comparing screen versus print and working memory research. Pedagogical and technological implications for onscreen academic reading activities at higher education were discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

University students in Anglophone countries are increasingly reading from screens for academic purpose due to the increased use of digital libraries and internet databases. Recent advances in technology have contributed to the variety of electronic sources with screens, such as e-readers, computers and tablets (Shenoy & Aithal, 2016). With the increasing popularity of reading from screens, numerous studies have been conducted comparing reading from electronic versus print sources in terms of comprehension performance and reading processes (Clinton, 2019). This study follows this pursuit and examines differences in performance and processes between reading from print and screen for university students at higher education.

Debates concerning reading from screen vs. print have been controversial in the context of higher education. Advocates of reading from screen have emphasised the lower cost of using electronic textbooks compared with paper books for university students (Bando, Gallego, Gertler, & Romero, 2016). Also, electronic texts can be accessed from various devices such as e-readers or smartphones. This feature provides university students convenience in terms of accessibility and transportability of academic reading materials. However, critics of reading from screen argue that reading from screen is inferior to reading on paper because readers have disadvantageous performances and metacognitive awareness (Ackerman & Lauterman, 2012). There are also other concerns on reading from screen, for example, less engagement than that of reading on paper (Mangen & Kuiken, 2014), longer reading time on screen without benefits to comprehension (Daniel & Woody, 2013). However, the emerging evidence about the potential negative effects of reading from screen (e.g., Singer & Alexander, 2017; Stoop, Kreutzer, & Kircz, 2013) is based on assessments which generally assumes reading comprehension to be a global construct rather than being comprised of a more complex architecture. Thus, this study aims to unpack global reading assessments and compare the differences in the underlying constructs at different levels.

Additionally, the UK has attracted a rapidly increasing number of international students from non-English-speaking countries, particularly Chinese students, to attend universities for higher education. According to The Guardian, "The UCAS university admissions agency revealed it had received almost 20,000 undergraduate applications from students in China in 2019 (19,760, up from 15,240 in 2018), compared with 18,520 from Northern Ireland" (Sally, 2019). This context provides a rationale for research comparing reading from screen versus print by L1-Chinese-speaking students who come from different educational and language backgrounds.

In this study, first-year undergraduate students were selected for investigation because this is a critical transition period where students move from general education at school-level to disciplinary major courses at higher education. Apart from linguistic challenges in an Anglophone university, L1-Chinese-speaking students have to encounter a variety of academic reading materials that are shaped by the norms and values of the disciplinary field and culture. An investigation of L1-Chinese-speaking students who use English as a second language makes an important new contribution to research comparing reading from screen versus print.

This study brings the empirical (screen inferiority), methodological (reading comprehension assessment when comparing screen and print performances) and practical (L1- and L2-English-speaking university students for academic reading purpose) orientations and foci together. In order to achieve this aim, the remainder of this chapter will first explain the problems before conducting this study (1.2), and detail the research aims (1.3) before outlining the thesis structure (1.4).

1.2 Problem statement

> Reading in print or digital form should not be a horse race questions. One medium will not and should not be regarded as routinely better for comprehension. (Singer & Alexander, 2017, p. 29)

Three problems were identified before conducting this study. First, theoretical frameworks to understand reading comprehension at higher education and to compare comprehension performance between screen and print are underdeveloped. The most cited model of comprehension in the studies comparing screen and print is Kintsch's (1988) Construction-Integration Model of comprehension. However, this model is more often applied to reading from print and does not explicitly consider reading from digital screens. As Singer and Alexander (2017, p.2) point out, "[there is] limited understanding of how particular attributes of the context, the text or the learner might interact with the medium to enhance or inhibit comprehension". Using the theory of purposeful reading by Britt, Rouet, & Durik (2017), this study considers an authentic reading situation in an academic context where there is a set of tasks linked to reading a text. Further, this study considers the important variables that may affect reading comprehension, for example, readers' discipline and background knowledge, the length and complexity of a text that university readers often encounter (e.g. a chapter from a textbook), the reading tasks university students are most likely to undertake (e.g. locating specific elements, summary of a section).

The second problem is that, a valid and reliable experimental test that is used to compare reading comprehension performances from screen and print for university students is needed. Well-established reading comprehension tests often lack explicit information about how designers develop the test items. The danger of this is that the tests may mask the underlying differences of reading from screen versus print as they could only rely on the overall results of comprehension performance to compare the advantages and disadvantages. For example, Keenan et al. (2008) draw the following general conclusions that,

> Comprehension is a complex cognitive construct, consisting of multiple, component skills. Even though this complexity is recognized theoretically, when it comes to assessment, there is a tendency to ignore it and treat tests as if they are all measuring the same "thing." This is reflected in the fact that researchers who measure comprehension rarely give information on why they chose the particular test that they used. (Keenan et al., 2008, p. 294)

Similar issues are also identified in the research that investigates digital reading. Issues such as lacking definitions of reading comprehension for print/onscreen and task-specific information, are reported in a systematic review of digital reading research by Singer and Alexander (2017). They said,

> [As regarding print reading], of the 36 charted studies only 9 (25%) included any manner of definition of Reading on Paper and Digitally — be it explicit or implicit... [As regarding screen reading], unfortunately, the definitions of digital reading were similarly scant to those for reading [from] print. Only five articles (13.89%) included a definition of digital reading in any form… [Therefore], lack of dedicated definitions of digital reading reflects researchers' unstated perception that the distinction between reading and reading digitally has more to do with the context of the process and is not some reconceptualization of the basic construct. (Singer & Alexander, 2017, p.10).

Researcher-developed reading comprehension tests are generally configured to the purpose of the study (Kimberlin & Winterstein, 2008) but lack compelling evidence of strong reliability and validity. For example, in Noyes et al.'s study (2004), they described in detail their measures of "10 multiple-choice questions following by the administration of the NASA-TLX" (p.112), however, details regarding reliability and validity were underreported. Therefore, this study attempts to design and develop an experimental reading comprehension assessment with explicit information on the development of test items as well as detailed description of validity and reliability. This will be designed to compare the possible differences between screen and print on different levels (e.g. literal and inferential) underlying reading comprehension measurement rather than sole global result of comprehension performances.

The third issue is that explanations regarding individual difference factors and text processing need more clarification to account for performance outcomes between reading from print versus screen. Individual difference factors such as working memory, academic ability, vocabulary or background knowledge, level of study, age and gender have been demonstrated to play an important role in reading comprehension performance between print and digital medium (Afflerbach , 2015; Luke, Henderson, & Ferreira, 2015; Kendeou et al., 2011). Of these, working memory has been cited as one of the most potential explanations for screen inferiority (Mangen et al., 2013; Wästlund , 2007; Mayes et al, 2001). However, little research has empirically evidenced the relationship between working memory and reading comprehension from digital screen, so it is unknown to what extent comprehension performance between print and screen can be attributed to working memory. Therefore, this study not only attempts to examine whether there are any differences in reading comprehension performances between screen and print, but also explore what and how the cognitive attributes of a reader (working memory) may contribute to the possible differences.

1.3 Research aims and overall design

The aim of this thesis is to examine differences of reading comprehension from screen and print but also to explore the underlying factors that are contributing to the possible differences. In order to achieve this aim, three different domains of literature are examined: 1) exploring the theories of reading to identify a theoretical framework to understand reading comprehension and develop a valid and reliable experimental test to measure reading comprehension in higher education; 2) reviewing the studies comparing reading from screen and print to identify the factors that needs to be considered when designing a comparative reading research and implementing an experimental reading test for university-level students; 3) investigating the working memory theories to understand text processing differences from screen and print and explore what cognitive attributes of a reader that may affect comprehension performances across different medium.

The study adopted a mixed-method approach and conducted three empirical studies. A Reading Comprehension Assessment (RCA) was designed and administered to examine whether differences exist in reading comprehension performances when reading from screen and print taking into account two dimensions of reading, namely, literal and inferential. A Working Memory Capacity Test (WMCT) was employed to examine the relationship between Working Memory Capacity (WMC) and comprehension performances from screen and print. Finally, a follow-up interview was triangulated with the data collected from the RCA and WMCT. This approach provides qualitative data to compare text processing differences between screen and print but also gives access to the participants' different perceptions and their contexts that help to shape the perceptions.

1.4 Overview structure of this thesis

This section outlines the content and focus of each chapter of this thesis. The thesis is divided into 12 chapters. The reviewed literature from Chapter 2 to 4 fall into three provinces: 1) theories of reading – Chapter 2; 2) review of empirical reading research on screen and print reading – Chapter 3; 3) theoretical perspective of working memory – Chapter 4. Chapter 2 starts by reviewing the theories and models of reading in order to identify an appropriate framework to compare reading comprehension performances and processes. After reviewing the Simple View of Reading (SVR) (Gough & Tunmer,1986), the levels of reading comprehension theory (Applegate, Quinn & Applegate, 2002; Herber, 1970; Barrett, 1968), and the Construction-Integration (CI) Model (Kinstch, 1998), this chapter proposes a hybrid framework of reading comprehension assessment drawing on Barrett's taxonomy (1968) and the CI Model (Kinstch, 1998).

Chapter 3 deals with the methodological issues relating to reading research from the digital devices. It reviews the empirical studies in recent decades to develop appropriate experimental reading measurements for university-level students. The review details three lines of studies comparing reading from paper and screen: comprehension performances, cognitive processes and metacognition. It then considers important factors in conducting a self-developed reading comprehension assessment, including text length, text types, the levels of reading tasks, the academic level of readers and the L1 of readers. Reviewing previous studies suggests these factors play a critical role in relations between reading medium and reading comprehension.

In Chapter 4, theories and models of working memory are examined. Reviewing different models from cognitive psychology – Information Processing Model, the Multi-store Model of Memory by Atkinson and Shiffrin (1968) and the Levels of Processing Model (Craik & Lockhart, 1972), provides the theoretical underpinnings to understand text processing differences between onscreen and print. After discussing the measurements of working memory, this chapter ends with the rationale for utilizing the Multi-component Model of working memory (Baddeley and Hitch, 2000) as the theoretical framework to examine the relationship between working memory and reading comprehension performances from screen and print.

Chapter 5 describes the pilot study conducted prior to the main study, which aimed to validate the experimental reading test.

Chapter 6 to 9 are the empirical parts of the main study, including methodology and results from the Reading Comprehension Assessment – RCA (Chapter 6 and 7), the Working Memory Capacity Test – WMCT (Chapter 8) and a semi-structured interview (Chapter 9). Chapters 6 and 7, explain the overall design and research aims, followed by the philosophical underpinnings of the study and methods of assessing reading comprehension. Then, a hybrid framework of reading assessment is proposed and applied to the development of a complete experimental reading comprehension assessment. Detailed descriptions of the measures taken to ensure test validity and reliability are also included. Finally, this chapter ends with assumption checking of the RCA data to inform statistical analysis; the results of the statistical tests are given in Chapter 7.

Chapter 8 focuses on the WMCT , including the methodology and results. Unlike Chapters 6 to 8 which focus on the quantitative data, Chapter 9 deals with the qualitative data of this study – the responses obtained using a semi-structured interview.

Chapter 10 to 11 discuss the empirical findings in relation to the research questions. Finally, Chapter 12 provides a conclusion, including potential contributions, implications, limitations of this research and suggestions for future research.

Review of literature

Chapter 2 Theories of reading comprehension

2.1 Introduction

As the theoretical frameworks and standardized tests of reading comprehension are limited within the academic context in higher education, the literature review starts by thinking the theories and models of reading in order to identify an appropriate framework to compare reading process similarities and differences when comparing reading onscreen versus print. Section 2.2 starts with one of the most predominant models of reading at early literacy – the Simple View of Reading (SVR) proposed by Gough & Tunmer (1986). The significance of SVR model is highlighting the most important factors and skills that can help to explain reading comprehension issues. However, due to limitations of applying SVR to the higher education context, Section 2.3 continues to explore an alternative model – the levels of comprehension theory, accounting for the complexity of reading comprehension. Then, the following sub-sections review the levels of comprehension theory by discussing three different ways to understand reading comprehension – *Literal, Inferential and Evaluation* (Herber, 1970); *Literal, Low-level Inference, High-Level Inference and Response* (Applegate, Quinn & Applegate, 2002); *Barrett's Taxonomy – Literal, Reorganization, Inference, Evaluation and Appreciation* (Barrett, 1968). Of these, Barrett's Taxonomy will be highlighted because it is one of the most representative taxonomies in classroom reading instructions and assessments but also a potential framework to design a reading test for this research. Next, Section 2.4 discusses the categorizing of inferences (e.g. local and global; inter-sentence and elaborative) in that this is a core issue accounting for the debate of previous different ways to define the levels of reading and to assess comprehension performances at different levels. Section 2.5 moves to the Construction-Integration (CI) Model (Kinstch, 1998), which conceptualizes two types of inferences (text-based and knowledge-based) and two levels of comprehension (Textual model and Situational model), which seems to be the mostly used theoretical framework for the studies concerning reading onscreen versus print (Singer and Alexander, 2017). Lastly, Section 2.6 hypothesizes a hybrid framework of reading comprehension assessment by combining CI Model and Barrett's Taxonomy.

2.2 The Simple View of Reading

Reading comprehension is generally defined as the ability to extract meaning or learn from text (Rupley & Blair, 1983; Snow, 2002). This general definition embraces the conception of Simple View of Reading (SVR) (Gough & Tunmer, 1986). According to the SVR (Figure 2.1), reading comprehension results from decoding and language comprehension. Decoding is defined as word recognition and relies on students' phonological knowledge. Language comprehension represents the ability to process lexical information and to derive sentence or discourse interpretations (Hoover & Gough, 1990). Typically, linguistic comprehension is assessed with listening comprehension measures. The SVR suggests that decoding is more important at early stages of reading acquisition whereas language comprehension becomes more important in later stages as texts become increasingly complex in terms of linguistic resources. In addition, it is assumed that if either skill, decoding or language comprehension, is developed unsuccessfully, reading comprehension cannot be reached (Hoover and Gough, 1990). Therefore, reading comprehension is considered as a product of skills in decoding and language comprehension.

Figure 2.1 The Simple View of Reading Model, adapted from Gough & Tunmer (1986)

Although the SVR does not clearly indicate how to measure decoding and language comprehension, most studies in this regard have operationalized decoding skill as word-reading accuracy and fluency (Kirby & Savage, 2008). The original authors of the SVR attempted to avoid potential misunderstanding of decoding in the model by defining as "ability to rapidly derive a representation from printed input that allows access to the appropriate entry in the mental lexicon" (Hoover & Gough, 1990, p. 130; Hoover & Tunmer, 1993, p. 6). In short, decoding is equate to skilled word recognition. The components of decoding include a set of lower-level skills, such as phonological and orthographic knowledge, alphabetic understanding and word recognition fluency. It is generally viewed that these measures of decoding follow a developmentally constrained process (Tummer & Greaner, 2010). To illustrate, children are able to comprehend only once they developed accurate, automatic word reading skills and can read connected text with some degree of fluency (Dole, Duffy, Roehler, & Pearson, 1991). The accuracy and fluency of the decoding process is also referred as *automaticity* in broader literature. Regarding the target context, students at university-level are believed to have already equipped with an intermediate to advanced level of automaticity of English reading according to the A-level or IELTS reading test requirement by the university, either for L1-English-speaking natives or L1-Chinese-speaking students who have learnt and used English as a second language.

The component of language comprehension is arguably more complex. In the original version of SVR by Gough and Tunmer (1986), language comprehension included a) Vocabulary knowledge, b) Background knowledge, and c) Knowledge in text structure and sentence structure. *Vocabulary knowledge* is more than just citing the definition of a word. It requires that the reader use the word appropriately based upon a given context. This is because understanding of words in isolation is sometimes not sufficient for comprehension as the meaning of many words is often dependent upon the context in the text.

Background knowledge refers to the past experiences and/or prior understanding a reader brings to the text. It is also called schema in the literature (Anderson, 1999). The ability to incorporate the prior knowledge to create a new schema and the ability to organize understanding of the situation that can be readily applied to the text being read (Gernsbacher, Robertson, Palladino,& Werner, 2004; Kintsch & Kintsch, 2005; Zwaan, Radvansky, Hilliard, & Curiel, 1998; Anderson, 1984; Van Dijk & Kintsch, 1983;) are crucial to support the understanding of the text (Cain, Oakhill, Barnes, & Bryant, 2001; Kintsch, 1988; McNamara, 1991).

Knowledge in text structure and sentence structures has two levels of meaning: the first is to have the ability to recognize a variety of features that are common to a specific text and how a text is organized (text structures); the second is to have the ability to deal with the grammatical and syntactical complexity (sentence structures). With regards to *knowledge in text structures*, narrative and expository text are typical texts that are different in purpose (McCarthy, Graesser & McNamara, 2007) and structure (Best, Floyd, & McNamara, 2008; Fox, 2002). Narrative texts are generally written to tell a story and entertain; they usually use causal event chains, setting-conflict-resolution or story grammar structures to describe episodes (Otero, Leo´n, & Graesser, 2002; Weaver & Kintsch, 1991) and thus contain elements such as characters, settings, problems and solutions. By contrast, expository texts, often called informational texts, are designed to introduce new concepts and ideas with which the readers may not be familiar (Best et al., 2008). Expository texts have a variety of structures according to their purposes, such as Description or Listing, Sequence, Compare and Contrast, Cause and Effect, and Problem and Solution (Weaver & Kintsch, 1991); these elements are commonly organized in a hierarchy of propositions (the units of ideas) related to a central topic (Tun, 1989). With regards to *knowledge in sentence structure*, it is related to higher-level language comprehension process, such as understanding the semantic (meaning), syntactic (grammatically) and referential relationships (inferences) among connected words to construct meaning from text (Hanon & Daneman, 2001). In summary, the knowledge in *text structure and sentence structure* can substantially facilitate reading comprehension (Gersten Fuchs, Williams, & Baker, 2001). From the above, it can be seen the original SVR proposal captures the basic aspects of language comprehension, such as vocabulary, semantic and syntactic skills, which can help to explain the comprehension difficulties that may appear in reading process from screen versus print.

However, there are two limitations when applying the SVR proposal to the current context. First, the original SVR proposal does not take account of learners who come to read English as a second language. According to Frith's (1985) three-phases scheme in learning to read words – logographic, alphabetic and orthographic, the decoding and comprehension processes between English and Chinese are slightly different. Logographic refers to the use of graphic features to read words, as is done in reading Chinese or some sight word in English. Alphabetic refers to the use of grapheme-phoneme relations to read words, as is done in English. Orthographic refers to the use of spelling patterns.

Considering the L1-Chinese-speaking students in English-speaking universities, they tend to follow a strictly visual characteristics rather than letter-sound correspondences to read words. They are very likely to be already fluent in reading their first language but have weaker decoding skills in English as a second language. They are also likely to have weaker language comprehension skills in L2 than in L1. In such a situation, two possible comprehension issues may arise. For one, whether L1 decoding competence compensates for weak decoding skills in L2; for another, whether strong language comprehension in L1 compensate for weak language comprehension in L2. Indeed, some research indicate that the word-level skills may have an impact on L2 decoding and L1 language comprehension may also compensate for weaker L2 language comprehension (Deacon et al., 2007). This line of research is mostly on the basis that English has an ‘ alphabetic system’. Yet there is not much known about the languages with ‘logographic system’, such as Chinese. These issues may need investigation beyond the SVR proposal.

Second, some researchers argue that decoding with orthographic processing not only depends upon reading experience and exposure (Stanovich and West, 1989), but also upon cognitive processes (e.g., Kirby et al., 2003; Wolf and Bowers, 1999). Cognitive processes can influence decoding, but are not always seen as part of the decoding term. Most broadly, written word recognition is influenced by phonology, orthography and semantics (Plaut, 2005). The semantic component is represented by context, but also by word meaning and morphology. For example, knowing that the context of “classroom study” may speed up the recognition of white *board*, and being aware of morphology might help recognition of a morphologically complex word such as *bookcase, preview, classification*. A key additional assumption to this view is that learning to decode words depends on acquiring information about both word forms and meanings from word-learning events (Bolger, Balass, Landen, & Perfetti, 2008; Perfetti, Wlotko, & Hart, 2005). Word-learning events includes multiple word processing components such as visual attention, orthographic recognition, meaning processes, syntactic processes to be observed on a single word (Perfetti & Stafura, 2014). Therefore, the inclusion of a semantic contribution or word-learning events to decoding is different to the SVR proposal, because the semantic meaning or syntactic processes would normally be considered as part of the Language Comprehension (LC) term. Therefore, it is fair to argue that these two terms, language comprehension and decoding, should not be conceptualised as rigidly separated or completely independent (Bowey and Rutherford, 2007; Nation and Snowling, 1998). On these grounds, reading comprehension should be considered as a multi-level interactive process, rather than a product of skills. Regarding the academic reading comprehension in higher education, the reading process is far more complex than early leaning-to-read stages, requiring both word-reading levels and higher-level skills. Most possibly, the word-reading level skills – phonological awareness, alphabetic understanding, automatic grapheme-phoneme correspondence, inevitably work in conjunctions with other skills: 1) higher-level skills involved in SVR, such as vocabulary knowledge (semantics) and sentence structure (syntactic); 2) background knowledge or Schema construction; 3) knowledge about text structure.

2.3 Section summary

This section reviews the original SVR model proposed by Gough & Tunmer (1986), describing what constitute its two primary components *decoding* and *language comprehension*. The value of the SVR model for the current thesis is highlighting the most important factors and skills that help to understand reading comprehension:

1) Word-reading skills (automaticity);

2) High-level skills (e.g. semantics or syntactic);

3) Background knowledge or Schema construction;

4) Knowledge about the text structure.

Also, this section raises two potential limitations of SVR model:

1) Lacking empirical research from L2 English learners;

2) Overlooks the influences that wider language comprehension resources may have on decoding processes (e.g. the impact of semantic meaning on word identification).

For these reasons, the next section explores an alternative framework to examine reading comprehension in the target context of higher education, which usually has a focus on higher-level skills rather than basic word-level skills.

2.4 The levels of reading comprehension theory: approaches, debates and concerns

> Reading is an active, fluent process which involves the reader and the reading material in building meaning. (Anderson, 2004, p. 1)

For the last decades, theories and models of reading have developed from viewing reading as product (e.g. SVR discussed in last section) towards a process (Urquhart & Weir, 2014). Anderson (1999) divides views of the reading process into three general categories: *Bottom-up, Top-down and Interactive*. *Bottom-up* is a data-driven or text-driven process from basic linguistic units to build up meaning of texts – from single phonemes to words, words to phrases, phrases to clauses, sentences and then the whole piece of discourse (Aebersold and Field, 1998). *Top-down* is a reader-driven process drawing on readers' existing knowledge of the subject and expectations to the text contents to obtain a confirmation of what they have already predicted (Grabe and Stoller, 2013). *Interactive,* however, is conceptualized in two perspectives. It not only refers to an integration of the bottom-up and top-down processes (Grabe, 2009) but also an interaction occurring between the reader and the text in general (Anderson, 1999). Therefore, during the process, a fluent reader is involved in making inferences through constant interactions with the text, integrating the activated information in mind, and finally deriving meaning from the text into a coherent discourse.

The view of reading as an interactive process is embraced by a number of researchers (Samuels, 1994; Stanovich, 1992; Eskey and Grabe, 1988; Perfetti, 1985). For example, Crystal (2007, p. 209) agrees that reading "crucially involves appreciating the sense of what is written: we read for meaning"; Sweet (2005) suggests reading depends on the skills but also on an interaction between the reader and the text. The interactive approach sees reading from a more cognitive perspective, unlike the component approach (e.g. SVR) to reading comprehension – viewing reading as a development of ordered skills from lower-level (e.g. decoding) to higher-level. Although many researchers accept the interactive view of reading (Grabe, 2013 & 2009), the issue about how particularly a reader would interact with a text is debated. A general consensus is that, as reading comprehension is comprised of complexity, a reader is expected to interact with the text at different levels (Kintsch & Rawson, 2005). For example, reading tasks of various difficulty levels may challenge readers' cognition demands and require readers to respond differently (Carnine, Silbert, Kame'enui & Tarver, 2010; Vacca et al., 2009; Lapp & Flood, 1983; Herber, 1970). With regards to the reading activities in Anglophone university contexts, the reading tasks usually involve particular reading purposes and impose different task demands for students. For example, reading to learn discipline knowledge requires understanding, elaborating concepts by integrating information from lectures and reference books' content; reading to write an essay requires summarizing the basic content of textbooks, articles and essays but also to criticize and evaluate others' ideas. To this end, the levels of reading comprehension theory may provide some insights for understanding the nature of reading comprehension required by the reading tasks in higher education context.

In brief, the theory about the levels of comprehension assumes that there are different levels of reading comprehension, each of which requiring varying degree of interaction with the text. This perception is not recent but deeply rooted in the literature. The next sections will discuss three different approaches to understand the levels of reading comprehension, from a three-level to a five-level framework, due to their prevalence and influence in the teaching practice and reading instructions for the last few decades (Carnine, Silbert, Kame'enui & Tarver, 2010; Vacca et al., 2009; Lapp & Flood, 1983; Herber, 1970). Section 2.4.1 will describe Herber's (1970) proposal of a three-level of comprehension – *Literal, Inferential, Evaluative*. Section 2.4.2 will review Applegate, Quinn, and Applegate's (2002) four-level of comprehension – *Literal, Low-level Inference, High-level Inference and Response*. Section 2.4.3 will highlight Barrett's Taxonomy (1968, cited in Clymer, 1979) with a five-level of comprehension – *Literal, Reorganization, Inferential, Evaluation and Appreciation*. Finally, Section 2.4.4 will discuss the debates and concerns about the levels of comprehension theory.

2.4.1 Literal, Inferential and Evaluative

According to Herber (1970), reading comprehension can be viewed as comprising *Literal comprehension, Inferential comprehension and Evaluative comprehension. Literal comprehension*, a stepping stone to reading, requires that a student be able to simply retrieve information that is explicitly stated in a text (Carnine et al. 2010). *Inferential* comprehension, an extension of recognition step of literal comprehension, requires readers to go beyond recognizing the facts derived from a text and interact with a text to make inferences about meanings that are not explicitly stated in the text (Rupley and Blair, 1983). *Evaluation comprehension*, a further extension of the knowledge and skills involved in literal and inferential comprehension, requires readers to juxtapose what they have read in the text with their own prior knowledge and experience, creating new meanings and/or relationships that extend beyond the scope of the text (Herber, 1970). In the same vein, Robinson, Good and Brophy (1987) use different terms to express similar ideas as Herber (1970), which are *Literal, Interpretative and Critical*.

Although different terms are used, there are *three* common assumptions related to each level of comprehension by Herber (1970) and Robinson, Good and Brophy (1987). First, *Literal comprehension* is dependent upon students' word-level processing skills, such as the ability to precisely identify individual words and understand the meaning created by the combination of words into proposition units (Perfetti et al., 2005). It is primarily made up of two components: *recall* and *recognition*. *Recall* refers to the ability to provide an idea (e.g. main idea or details). *Recognition* refers to the ability to recognize whether a piece of specific information is provided in a passage or not (Rupley and Blair, 1983). Therefore, according to Jude and Ajayi (2012), the essence of Literal comprehension is to identify the exact meaning of lexical items – reading for exact meaning at the word and sentence level, and read for information — understanding the central point by the author.

Second, unlike *Literal comprehension* with a highlight on explicitly stated information and linguistic resources, readers tend to reduce their reliance on the text at the level of *Inferential or Interpretative comprehension*. Instead, they tend to relate textual contents to their reasoning and pragmatic knowledge in order to be able to form a coherent mental representation of what the text is about. In other words, reader's dependence on the text itself decreases as inferencing becomes richer and deeper (Alptekin, 2006). This level of making inferences usually involve in *interpreting, summarising, reasoning and generalizing* (Alptekin, 2006). Thus, the essence of this level is to manipulate information in the text to search for relationships among the main idea and details and to use that information to draw conclusions about the author's intended meaning (Vacca et al., 2009).

Third, when moving beyond the first two levels towards the third *Evaluative or Critical comprehension*, readers need to analyse, critique and evaluate the information acquired from the text or imported outside the text in terms of prior knowledge and experiences (McCormick, 1992). The nature of this level is to create new meanings and relationships including divergent thinking, critical analysis and evaluation, along with affective, or personal and emotional responses when necessary (Vacca et al., 2009). The three-level-reading comprehension theory has been prevalent in a large amount of educational teaching materials and English teaching and learning classroom over the last decades (Carnine et al. 2010; Herber 1970; Lapp and Flood 1983; Vacca et al. 2009). An underlying assumption for teaching and learning assumes a continuum of reading comprehension skills within each level: a reader must first proficiently engage in tasks of literal comprehension before engaging in deeper interactions with the text such as those prompted by inferential and evaluative understanding. However, when doing academic reading tasks, whether university students follow a sequential procedure of reading skills or draw on different reading skills at the same time to reach a certain level of comprehension is expected to explore. This chapter continues to discuss this issue in Section 2.4.3.

In conclusion, this section reviews the three levels of comprehension approach – *Literal, Inferential and Evaluative* by Herber (1970). *Literal comprehension* is a stepping stone to move into more advanced comprehension tasks; *Inferential comprehension* is a logical extension of the recognition step of the first *Literal comprehension*; *Evaluative comprehension* is a further extension beyond the text. However, the dividing line between the inferential level and evaluative level is sometimes ambiguous because both levels require readers to connect existing knowledge and experience to make inferences of the text. To this end, Applegate, Quinn and Applegate (2002) delve deeper into the inferential level of comprehension and propose Low-level Inference and High-level Inference.

2.4.2 Literal, Low-level Inference, High-level Inference and Response

The four-level comprehension framework (Applegate, Quinn & Applegate, 2002) of *Literal, Low-level Inference, High-level Inference and Response* has been widely used as typical types of instruments in informal reading inventories. They are effective tools to help teachers and specialists to break the cycle of questioning (Gandy, 2013). In Applegate, Quinn and Applegate's (2002) research, they identified these four types of reading comprehension to design the open-ended constructed-questions that are used in classroom teaching and assessment. It should be noted that the reading instruments used by the researchers are also an indication about their understanding to the nature of reading. For example, if researchers ask sets of questions that are primarily for literal recall, they would, in effect, be conveying a message about their expectations about their understanding of reading. This issue will be explored in discussion chapter. For the purpose of this chapter, these four types of reading questions are considered as a holistic four-level model to explore the reading comprehension.

In Applegate, Quinn and Applegate's (2002) definitions, each level of question items represents a level of comprehension and has its own characteristics. The followings are the details (adapted from Applegate, Quinn & Applegate, 2002):

1. Literal

Definition: Readers are required to recall what they have read.

Characteristics: The answers to these items are stated explicitly (verbatim) in the text.

Example of question items:

The text states that Mary, a character in the story, is in fourth grade.

Questions: The literal question asks, "What grade was Mary in?"

2. Low-level inference

Definition: Readers are required to draw a conclusion on the basis of the text and use their background experiences to some extent. However, low-level inferences require very little in the way of drawing conclusions.

Characteristics: The answers to low-level inferences are not stated verbatim in the text but may be so close to literal as to be obvious. This include:

• Those that involve the recognition of information in different words from those used in the original text.

• Those that require the reader to identify relationships that exist between ideas in the text. Such items as these are not literal only because the writer has not made the relationship explicit by using a grammatical marker (e.g., because).

• Those that deal with details largely irrelevant to the central message of the text; or

• Those that require that the reader draw solely on background knowledge or speculate about the action of characters without the benefit of information in the text.

Example of question items:

The text states that “Mr. Wilson’s car would not start. Mr. Wilson was late for work.”

Questions: “Why was Mr. Wilson late for work?”

3. High-level inference

Definition: Readers are required to link experience with the text and to draw a logical conclusion.

Characteristics: The answers to these items are NOT stated explicitly and require more complex thinking than low-level inferences. This include:

• Devise an alternative solution to a specific problem described in the text.

• Describe a plausible motivation that explains a character’s actions.

• Provide a plausible explanation for a situation, problem, or action.

• Predict a past or future action based on characteristics or qualities developed in the text.

• Describe a character or action based on the events in a story.

Example of question items: The text describes two characters and several circumstances in their lives.

Questions: “Why do you think that the two characters in the story became friends?”

4. Response

Definition: Readers are required to express and defend an idea related to the actions of characters, the outcome of events or to discuss and react to the underlying meaning of the passage as a whole.

Characteristics: The answers to these items are NOT stated explicitly and usually direct toward a specific element or problem in the passage, broader ideas or underlying themes. Response items require a reader to discuss and react to the underlying meaning of the passage. This include:

• Describe the lesson(s) a character may have learned from experience.

• Judge the efficacy of the action or decisions of a character and defend the judgment.

• Devise and defend alternative solutions to a complex problem described in a story.

• Respond positively or negatively to a character based on a logical assessment of the actions or traits of that character.

Example of question items: The story describes characteristics of two young children on a field trip.

Questions: “If you were a teacher, which of the two children would you rather have in your class and why?”

It can be seen that this four-level framework captures the levels of comprehension in order to develop comprehension questions. By contrast to Herber’s (1970) three-level model of Literal, Inferential and Evaluative, Applegate, Quinn and Applegate (2002) provide more practical and instructional criteria and examples for teachers and material developers in classroom teaching and assessment. However, the central ideas of *Literal* and Evaluative or Response share several similarities. The *Literal* level focuses on explicit textual understanding and readers' ability in word recognition with little background knowledge involved in. The Evaluative/Response level emphasizes the global understanding of the text and readers' ability in creating new meanings by linking existing experience. Moreover, the Inferential level of comprehension in Herber (1970) is further classified into Low-level and High-level Inferences by Applegate, Quinn and Applegate (2002). The difference between Low-level and High-level Inference lie in the extent to which the background knowledge is used. At Low-level Inference, for example, readers can identify causal relationship between two clauses without the explicit cohesive markers (e.g. because, as a result, due to); at High-level Inference, readers have to draw out a logical conclusion, such as predicating future actions/consequences, by relevant life experience. However, it is ambiguous to define the cut-line between the level of *literal comprehension* and *lower-level inference* in a higher-education context. Indeed, in a school-level context, literal comprehension question can be defined as those *answers are explicitly stated*; lower-level inference question can be defined as those answers are not explicitly stated but still obvious such as identifying existing relationship without the help of cohesive markers. For a university student with advanced English proficiency, as reading process is probably much more automatic and fluent than school-level students, one can easily understand the connection between clauses or sentences even without the cohesive markers. Therefore, it can be argued that the *lower-level inference* tends to be equivalent to *literal comprehension* for students with advanced level of English or proficient in English. The ambiguous division between literal comprehension and inferential comprehension leads to further exploration of the level of reading comprehension and thus next section reviews the third approach – a five-level taxonomy which was originally proposed by Barrett (1968).

2.4.3 Literal, Reorganization, Inference, Evaluation and Appreciation

Initially, Barrett (1968) developed a Taxonomy of Cognitive and Affective Dimensions of Reading Comprehension to eliminate a misconception that reading skills are assumed to cover an amount of unmanageable and undefinable skills (cited in Clymer, 1968). In order to demonstrate a controllable teaching process of reading, Barrett (1988) drew on Bloom (1956), Letton (1958) and Sanders (1966) to develop their taxonomy. Barrett's Taxonomy (Figure 2.2) proposes a five-level pyramid framework of reading from *Literal, Reorganization, Inference* to *Evaluation and Appreciation.* The framework includes the levels of comprehension and details what constructs should be included within each level. The descriptions of subskills are directly related to the different levels of comprehension. In this taxonomy, the subskills are ordered hierarchically from the lowest to the highest level of reading (Pearson, 2009), indicating a bottom-up teaching process of subskills. For example, the categories of literal comprehension and reorganization deal with the information that are explicitly presented in the text, which is considered to be the basic level; the remaining levels are more advanced, dealing with the implicit information that are implied in the written text. Returning to the issue that is raised in Section 2.4.1, although Barrett's Taxonomy supports a sequential process of reading skills from a basic level to a more advanced level, it is arguable that in the actual reading process whether students follow this sequential procedure of reading skills or draw on different reading skills at the same time to reach a certain level of comprehension. However, the current purpose of reviewing Barrett's Taxonomy is to identify a framework to develop a reading comprehension assessment for the target context. Thus, this thesis will turn back to this issue in Chapter 10 but look at its contents first.

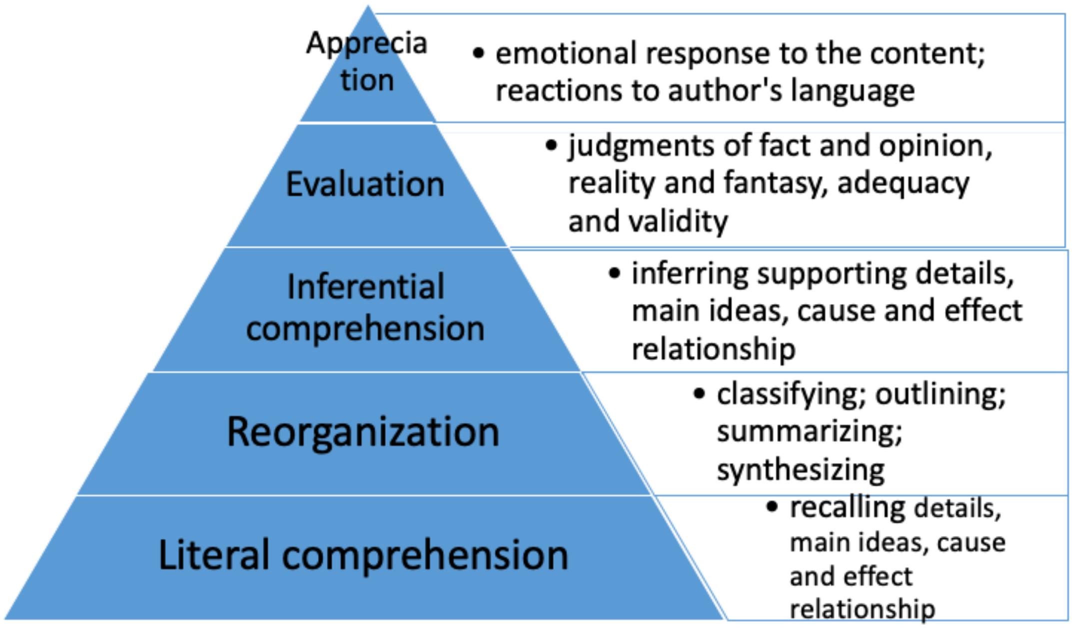


Figure 2.2 The Barrett's Taxonomy of Reading Comprehension (1968)

Essentially, Barrett's taxonomy distinguishes between the five levels of comprehension by the reading tasks. The reading tasks that test '*Literal Comprehension*' are those where answers are 'textually explicit'; tasks that involve '*Reorganization*' are those where the answers are in the text but not quite as obvious as 'Literal comprehension'. The tasks that test *Inferential, Evaluation and Appreciation Comprehension* are those where students need to use a combination of the content from the text along with their personal background knowledge of a particular subject to come up with an answer (Pearson and Johnson, 1979). Table 2.1 is a summary of Barrett's Taxonomy showing how each level of comprehension is conceptualized and what sub-skills can be included. Arguably, the taxonomy attempts to distinguish between reading tasks and questions by requiring students to 'read the lines' (Literal comprehension and Reorganization), 'read between the lines' (Inferential Comprehension), and 'read beyond the lines' (Evaluation and Appreciation).

Table 2.1 Descriptions of Barrett's Taxonomy (Adapted from Barrett's Taxonomy)

Levels of comprehension and subskills	Definitions and examples
1.Literal Comprehension Recognition or recall of - details - main ideas - a sequence - comparison - cause and effect relationships - character traits	To locate or identify any kind of explicitly stated fact or detail; For example, names of characters, places, likeness and differences, reasons for actions
2. Reorganization - classifying - outlining - summarizing - synthesizing	To organize, sort into categories, paraphrase or consolidate explicitly stated information or ideas in a reading selection
3. Inferential Comprehension - main ideas - supporting details - sequence - comparisons - cause and effect relationships - character traits - predicting outcomes	To use conjecture, personal intuition, experience, background knowledge, or clues in a reading selection as a basis of forming hypotheses and inferring details or ideas; For example, the significance of a theme, the motivation or nature of a character which are *not explicitly stated* in the reading selection.

- interpreting figurative language	
4. Evaluation - reality or fantasy - fact or opinion - adequacy or validity - appropriateness - worth, desirability and acceptability	To make evaluative judgement; use external criteria provided by other sources/ authorities or internal criteria; use students' own values, experiences, or background knowledge of the subject; For example, on qualities of accuracy, acceptability, desirability, worth or probability by comparing information or ideas presented in a reading selection.
5. Appreciation -Emotional response to content -Identification with characters -Reactions to author's language use -Imagery	To show emotional and aesthetic/ literary sensitivity to the reading selection; For example, showing a reaction to the worth of its psychological and artistic elements (including literary techniques, forms, styles, and structuring)

The advantage of Barrett Taxonomy is to give detailed descriptions of reading comprehension questions and subskills, relating them to different levels of comprehension. By contrast to the previous approaches from Herber (1970) and Applegate, Quinn and Applegate's (2002) framework, Barrett's Taxonomy is more detailed in that each level contains concrete subskills. For example, Barrett classifies *Recognition or recall of sequences and cause and effect relationships* as *Literal Comprehension*; by contrast, Applegate, Quinn and Applegate's (2002) tend to define *Recognition cause and effect* as Low-level Inference. Yet Herber (1970) hardly mentions the subskills within Literal Comprehension. To this end, Barrett's Taxonomy is probably more appropriate than Heber's three level (1970) or Applegate, Quinn and Applegate's (2002) four-level framework for the present study. For the purpose of identifying a theoretical framework to design an appropriate reading comprehension assessment, the division in Barrett's Taxonomy between each level is clearly defined and the subskills within each level is provided.

However, there are some repeated subskills within certain levels in Barrett's Taxonomy. For example, "cause and effect" appears in both *Reorganization* and *Inferential* level. Some may argue that the *Cause and effect* is supposed to be classified as *Inferential* because such relationship requires logical reasoning; some may claim it should be *Reorganization* if there are some explicit cohesive markers stated in the text; some others propose that it could be *Low-level Inference* as only a small amount of background knowledge involved, as discussed in last section. This ambiguity may be rooted in wider debate about the *inferences* which has been controversial for a long time in the literature. Different definitions of inferences result in different classifications to the levels of comprehension; in turn, different classifications reflect the understanding to the nature of reading and shape reading comprehension instruments differently in a number of empirical research. Thus, debates and concerns about *inferences* will be highlighted and discussed in Section 2.5.

2.4.4 Debates and concerns

The last sections have discussed three different approaches to examine the levels of reading comprehension. Two main concerns have been identified. First, the relationship between each level is unclear. Reading practitioners tend to assume that one skill is a prerequisite for another skill in the next level of comprehension. This is because an amount of research (e.g. McCormick, 1992; Snider, 1988), have suggested the relative degrees of difficulty of literal and inferential comprehension questions or literal, inferential and evaluative comprehension questions. For example, in McCormick's (1992) work with fifth grade students identified as struggling readers, she observed statistically significant differences in the amount of literal and inferential questions they were able to answer correctly. Specifically, while students in her study were able to answer, on average, 70% of literal questions correctly, they were only able to answer 61% of inferential questions correctly, implying that these questions were more difficult for them. On the other hand, opponents of this view (Lapp &

Flood,1983; Vacca, Vacca, & Gove, 1988) have pointed out an intrinsic flaw – oversimplification of the comprehension process by assuming a sequential and hierarchal order between the levels and skills. This is because the relative degrees of difficulty can only suggest that different level of comprehension tasks would place different cognitive complexity and processing demands on the reader. It does not necessarily mean the real reading process has to follow such sequential procedures. The central issues remains as to whether the relationship between each level is hierarchical or whether processes occur in conjunction with and support one another (Basaraba, Yovanoff, Alonzo, & Tindal, 2013). This issue will be returned to in Chapter 10 (Section 10.5) when considering the findings of the present study.

Second, there is a lack of empirical support for the levels of comprehension theory. The conceptualization and definition of different levels of reading (e.g. literal or inferential) and reading subskills (e.g. recognition, classification, predicting) are primarily found in teaching and learning practice, such as textbooks and assessment materials. Nevertheless, the existences of the levels of reading comprehension and reading subskills are not widely supported in empirical research. As indicated by Alderson (2000), “The notions of skills and subskills in reading is enormously pervasive and influential, despite the lack of clear empirical justification” (p. 10).

Earlier comprehension studies have utilized a variety of techniques such as factor analysis (Pettit, 1970; Davis, 1944 & 1968) in attempt to find evidences for comprehension subskills but come to reach inconsistent results. Some early work by Davis (1944) and Alshan (1964) using a factor analysis technique to provided more detailed information regarding the skills measured by the Cooperative Reading Comprehension Tests. Alshan (1964) attempted to determine 1) in a typical test of reading comprehension, the items define a number of distinct factors of reading ability and 2) those items written to measure one specific skill consistently had loadings on the same factor. The five orthogonal factors obtained in this sophisticated statistical treatment found that only one factor, reading comprehension, was apparent; and those items written to measure a specific skill did not seem to be identifiable as distinct skills. Opposite to this conclusion, Pettit (1970) tested the effects of reading for given purposes on literal and inferential comprehension. Setting of purpose for reading was provided for the experimental groups by instructions to read to find answers to questions calling for stated or implied meanings. Although she found that direct instruction did not significantly affect achievement on measures of either literal or inferential reading comprehension, she did find achievement to considerably vary within these dimensions of reading. In other words, the distinction between literal and inferential dimensions did exist. Still, these studies are quite outdated. Thus, this thesis aims to contribute to this line of research by replicating factor-analysis technique to justify the division between the levels of comprehension and the existence of specific subskills in a reading comprehension test for university students.

2.4.5 Section summary

This section reviews three different approaches to look at the levels of reading comprehension theory, from a three-level approach (Herber, 1970) – *Literal, Inferential, Evaluative,* a four-level approach – *Literal, Low-level Inference, High-level Inference and Response (*Applegate et. al., 2002) to a five-level approach (Barrett, 1968). These approaches were reviewed in an attempt to identify a theoretical framework to examine reading comprehension processes and develop an appropriate reading comprehension assessment to compare the differences between reading from screen and paper. Although each approach provides some insights for developing a reading test, Barrett's Taxonomy stands out because it clearly defines the division of five levels but also provides specific subskills within each level. These subskills can be developed as reading constructs when designing the test question items. However, there are two concerns applying Barrett's Taxonomy into the present study.

First, it is unclear whether the relationship between multiple levels and subskills is sequential or supported. Barrett's Taxonomy assumes a hierarchical relationship from basic literal comprehension to a more advanced inferential and evaluation comprehension, meaning, the subskills required by literal comprehension are prerequisite steps before mastering higher level subskills required by inferential level comprehension. Yet an opposite view questions such sequential relationship and argues reading processes may occur in conjunction with and support one another (Basaraba et al., 2013).

Second, the notions of multiple levels and subskills are in need of empirical justification. Although this notion has been sometimes taken for granted due to its prevalence in classroom teaching and learning, few studies have evidenced its empirical existence (Alderson, 2000). Therefore, for the purpose of designing a reading test, the levels of reading comprehension theory, in particular, Barrett's Taxonomy will be utilized as a fundamental framework due to its practicality. In addition to test design, this present study also attempts to respond to the two issues identified and contribute to theories of the nature of reading.

2.5 Inferences: Types, debates and concerns

The previous sections have discussed three different approaches to understand the levels of comprehension, Herber (1970) – *Literal, Inferential, Evaluative*, Applegate, Quinn and Applegate (2002) – *Literal, Low-level Inference, High-level Inference and Response* and Barrett's Taxonomy (1968) – *Literal, Reorganization, Inference, Evaluation and Appreciation*. The divergent categorizations can be attributed to different definitions and understanding about *inferences*. For example, *Low-level Inference* in Applegate, Quinn and Applegate's (2002) definition includes *identifying causal links* with the hint of cohesive markers whereas in Barrett's Taxonomy (1968) it is considered as *Reorganization*. Another example is *understanding the implied opinions of the author* is deemed as *High-level Inference* whereas in Barrett's Taxonomy it is regarded as *Evaluation.* The role of *making inferences* is considered to be at ''heart of the comprehension process'' (Dole et al., 1991). However, researchers have not reached agreement with regards categorizing and labelling various types of inferences. This may give rise to difficulties for the present study in designing the reading test items, for example, how the inferential questions are defined and what types of inferential questions should be included. To inform the methodological approaches to be used in the present study, Section 2.5.1 explores the distinction between different categorizations and Section 2.5.2 discusses two representative works on the inference taxonomies and their historical reasons as well as future directions.

2.5.1 What is the distinction?

A number of studies on reading comprehension has particularly examined the role of inferences because they are at the "heart of the comprehension process" (Doleet et al.,1991). In simple terms, making inferences means using two or more pieces of information from a text in order to arrive at third piece of information that is implicit (Kispal, 2008). For example, inferences can be as simple as associating the pronoun 'this' or 'that' with a previously mentioned item. Inferences can also be as complex as understanding a piece of subtle information from the choice of a particular word by the writer. Different researchers have identified different types of inferences (Bowyer-Crane and Snowling, 2005; Cain and Oakhill, 1999; Graesser et. al, 1994). It seems that there is no general consensus in the literature about the number of types of inferences and what they should be called. Also, this could be the reason contributing to the inconsistency in categorizing the levels of reading comprehension (e.g. *Reorganization* or *Low-Level Inference*; *Higher-Level Inference* or *Evaluative*).

Researchers have distinguished various types and categories of inferences, ranging from thirteen (Graesser et al., 1994), nine in Pressley and Afflerback (1995), to the most recent two (Cain and Oakhill, 1999; Bowyer-Crane and Snowling, 2005). Even amongst those researchers who have identified two types of inferences, there is an assortment of labelling. The earliest work of identifying two types of inferences include Cain and Oakhill (1999). They made a distinction between *text-connecting/inter-sentence inferences and gap-filling* inferences. *Text-connecting/inter-sentence inferences* are necessary to establish cohesion between sentences and involve integration of textual information; *gap-filling inferences*, by contrast, take advantages of information beyond the text from readers' knowledge pool. The more current terms used by Bowyer-Crane and Snowling (2005) are *coherence* versus *elaborative inferencing* which are appropriately equal to *text-connecting* and *gap-filling inferences. Coherence inferences* maintain a coherent text as the cognitive activity keeps going until the necessary information found to make the inferences; *elaborative inferences*, by contrast, are influenced by accessibility of knowledge.

Bowyer-Crane and Snowling (2005) have extended and refined the distinctions between coherence versus elaborative inferencing by adding knowledge-based and evaluative. As explained by the researchers themselves, adding the two additional inference types is because they are deemed essential to the understanding of a text even though they are dependent on life experience and existing knowledge. *Knowledge-based inferences* rely on the 'mediating idea' from the reader's own world-knowledge without which the text is disjointed. *Evaluative inferences* relate to the emotional outcome of the text, which is very similar to the ideas of *Appreciation* comprehension in Barrett's Taxonomy.

In some other studies, the dividing line has been defined in different terms, for example, local and global (Graesser et al., 1994; Gygax et al., 2004). *Local inferences* means creating a coherent representation at the level of sentences and paragraphs; *global inferences* means creating a coherent representation covering the whole text by drawing on local pieces of information. The details are below.

Local inferences include:

1. Coherence inferences - a coherent representation at the level of sentences and paragraphs.

2. Case structure role assignments, e.g. Dan stood his bike against the tree. The reader needs to realise that the tree is assigned to a location role.

3. Antecedent causal inferences, e.g. He rushed off, leaving his bike unchained. The reader would need to infer that Dan was in a hurry and left his bicycle vulnerable to theft.

Global inferences include:

1. On-line inferences: inferences drawn automatically during reading.

2. Off-line inferences: inferences drawn strategically after reading.

Graesser et al. (1994) were primarily interested in the distinction between on-line/off-line inferences, both of which were considered as *global inferences*. The inferences that are carried out automatically during reading are called *On-line* while the inferences that arise only when prompted are called *Off-line*. McKoon and Ratcliff (1992) express similar ideas by using the term '*automatic*' and '*strategic*' which are equal to *On-line* and *Off-line*. The following Table 2.2 summarizes the distinctions between different types of inference that have motioned previously.

Table 2.2 Distinctions between different types of inferences

Author	Distinctions	
Barnes *et al. (*1996)	coherence	elaborative
Bowyer-Crane and Snowling (2005)	coherence	elaborative/ knowledge-based/ evaluative
Cain and Oakhill (1998)	inter-sentence/ text- connecting	gap-filling
Graesser *et al*. (1994)	text-connecting/ on-line	knowledge-based/ off-line
Gygax *et al*. (2004)	local	global
McKoon and Ratcliff (1992)	automatic	strategic
Pressley and Afflerbach (1995)	unconscious	conscious

2.5.2 How many inferences are there and why?

The last section has reviewed some most frequently used types of inferences in the literature. Some of this work focuses on the differences and distinctions between two or three types of inferences, as discussed previously. However, the works of Graesser (1994) and Pressley and Afflerbach (1995) do not focus on the distinctions, rather, they detail a cataloguing of as many inferences as they are capable to find. In the work of Graesser et al. (1994), they produce an in-depth and comprehensive taxonomy of inferences from narrative texts; Singer (1997) subsequently made a contribution to their taxonomy from expository texts. In the course of their work, Graesser et al. (1994) has coined thirteen different forms of inferences. With regards to Pressley and Afflerbach (1995), they employed 'think-aloud' method which involved cognitive processes during reading; subsequently, they listed nine forms of inferences which readers carried out consciously. The following Table 2.3 presents a summary of their taxonomies.

Table 2.3 Taxonomy of inference by Grasser (1994) and Pressley and Afflerbach (1995)

Graesser et al. (1995)		Pressley and Afflerbach (1995)	
1	Referential	1	Referential
2	Case structure role assignment	2	Filling in deleted information
3	Antecedent causal	3	Inferring meanings of words
4	Superordinate goal	4	Inferring connotations of words/ sentences
5	Thematic	5	Relating text to prior knowledge
6	Character emotion	6	Inferences about the author
7	Causal consequence	7	Characters or state of world as depicted in text
8	Instantiation noun category	8	Confirming/ disconfirming previous inferences
9	Instrument	9	Drawing conclusion
10	Superordinate goal action		
11	State		
12	Readers' emotion		
13	Author's intent		

According to Table 2.3, it can be seen that there are some overlaps in the two list of taxonomies. For example, both mention the inferences about the character and the author. However, Graesser et al. (1994) focus on the inferences about the character, theme and instrument whereas Pressley and Afflerbach (1995) emphasize the process during making inferences such as confirming, conclusion and relating. The difference is presumably because the work of Graesser et al. (1994) is on a basis of narrative texts while the work of Pressley and Afflerbach (1995) is as a result of 'think-aloud' methods. More importantly, the divergence reflects different views of looking at inferences. Such different views may have historical reasons – the fierce debate between ' constructionist' view embraced by Graesser et al. (1994) and 'minimalist' view followed by McKoon and Ratcliff (1992). The constructionist view assumes the reader is engaged in a constant search of meaning to build a situation model of the text at local and global level in order to achieve coherence. The minimalist view is "*readers do not automatically construct inferences to fully represent the situation described by the text … there is only a minimal automatic processing of inferences during reading*" (McKoon & Ratcliff, 1992 p. 440). That means inferences are not required to establish local coherence but are encoded by readily available world knowledge (Long et al., 1996, p. 162). Therefore, there are overlapping between two taxonomies in that no consensus about the types of inferences can be reached.

Although it has been difficult to identify the specific skills and abilities that a reader should possess and practice to become advanced in making inferences, researchers have directed their focus to cognitive process of inferences. The direction has moved from pedagogic perspective to psychological perspective to examine the issues about inferences. Subsequently, working memory, an important concept which derives from psychology, has gained considerable attention in the research of reading comprehension. This concept is also closely related with the levels of comprehension theory. In essence, the idea underlying the higher-order of understanding of a text, regardless of different terms or names in the literature (e.g. Evaluative, Higher-level Inferences, Response or Appreciation), is that understanding a text is not merely dependent on information that is presented in the text explicitly stated or appearing across multiple clauses in the text. Rather, readers are required to hold presented information in their minds and simultaneously access information or experience from their storied knowledge pool so as to analyse or evaluate what they have just read. The ability to hold information in mind is related to the capacity called working memory (Baddeley, 2000), and the storied knowledge pool is referred as long-term memory. In this sense, working memory plays a central role in each level of comprehension. As working memory is a complex but essential concept, Chapter 4 will expand this topic in details.

Next section moves to another of theory of reading comprehension – Construction-Integration Model (1998). Research has attempted to apply this model in explaining and tracing out the cognitive process of what is happening in the mind during an inference and how it is related to reading comprehension.

2.6 Construction-Integration Model

The previous review of literature on the levels of comprehension theory (Section 2.3) suggests that reading comprehension is consisted of multiple levels and a set of reading skills within different levels and accepts the continuum of these levels (Barrett, 1968). However, Kinstch (1998) proposes a CI Model viewing the different levels of comprehension from the interaction between a reader and a text. This section will review the basic conceptualizations and assumptions underlying this model.

Kinstch (1988 & 1998) offer some important ideas in the CI Model to explore reading comprehension. First, it is assumed that there are two phases during reading – *Construction and Integration*. Second, there are two basic forms of structural relation of a text – *Micro structure and Macro-structure*. Third, there are two levels of comprehension – *Textual model comprehension* and *Situational model of comprehension*. The following will explain these terms in detail not only because they have been frequently used in the literature to reveal the nature of reading but also they are of vital importance to explain the reading comprehension issues in the current research.

2.6.1 Two phases: Construction and Integration

The CI Model suggests that reading comprehension can be achieved through two phases – a Construction phase and an Integration phase (Kinstch, 1998). Figure 2.3 indicates the process.

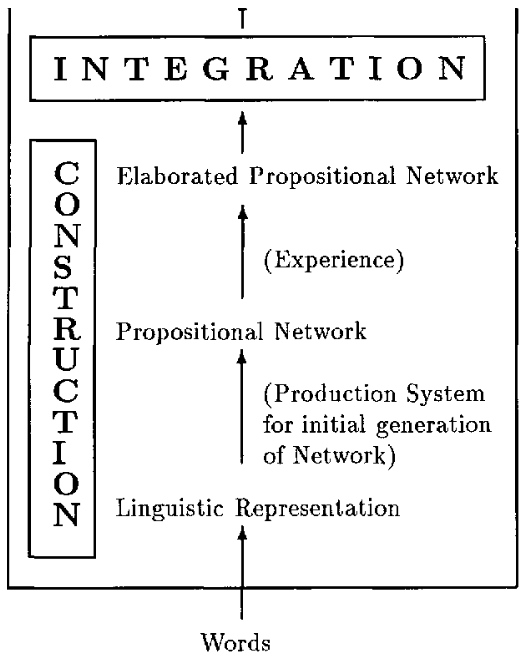


Figure 2.3 The Construction and Integration Model (Wharton and Kinstch, 1991)

Reading starts with "a construction phase, in which an approximate but incoherent mental model is constructed locally from the textual input and the readers' goals and knowledge; an integration phase that is essentially a constraint-satisfaction process that rejects inappropriate local constructions in favour of those that fit together into a coherent whole" (Kintsch, 1998, p. 119). When a sentence is being read, readers experience a construction phase. During the *Construction phase*, there is a quick activation of linguistic, semantic and topic-related knowledge from the surface of the text. This is followed by the integration stage when this activation is consolidated into a coherent meaning representation (Graesser et al., 2000). To some extent, the *Construction phase* can be seen as a bottom-up process encompassing word recognition and decoding, syntactic parsing and proposition formation (Grabe, 2009), which largely manipulates the explicitly stated information. Meanwhile, the *Integration phase* can be seen as a restructuring top-down processing guided by readers' own interpretations, which often requires making inferences about underlying implicit information.

2.6.2 Two forms of structural relations – Microstructure and Macrostructure

The CI Model assumes that the meaning of a text can be represented by a network of propositions (units of ideas), and the network of propositions establishes two kinds of basic forms of structural relation of a text: Microstructure and Macrostructure. The *Microstructure*, which is also referred as local-structure of a text, contains propositions that include information from a sentence. On the other hand, the *Macrostructure*, which is referred as global structure of a text, contains “a hierarchically ordered set of propositions” (Kinstch, 1998, p.50) which is usually derived from the entire section of a text. For example, the Macro-propositions involves the recognition of global topics and their interrelationships, which are frequently conventionalized as *schema*. Kinstch (1998), stresses the importance of Macrostructure for reading longer texts because it may dominate the comprehension process.

2.6.3 Two levels of comprehension – Textual model comprehension and Situation model comprehension

Corresponding to the two phases of reading process – moving from constructing information to integrating meanings of a text, Kinstch (1998) made a distinction between *Textual model comprehension* and *Situation model comprehension*. The Textual model comprehension, also referred as text-based understanding, is related to the units of ideas (propositions) at both Microstructure and Macrostructure which are directly attained from the text (Van Dijk & Kinstch,1983). To put it in another way, Microstructure and Macrostructure together constitute Textual model comprehension (Textbase understanding). The Situation model comprehension, is a representation of the textual information which is integrated from existing knowledge. It is on a basis of textual information and personal experience that readers create meanings in the Situation model comprehension. As cited,

> The text base [textual model] represents the meaning of the text, as it is actually expressed by the text. But if a reader only comprehends what is explicitly expressed in a text, comprehension will be shallow, sufficient perhaps to reproduce the text, but not for deeper understanding. For that, the text content must be used to construct a situation model; that is, a

> mental model of the situation described by the text. Generally, this requires the integration of information provided by the text with relevant prior knowledge and the goals of the reader. (Kinstch and Rawson, 2005, p. 211)

Moreover, the CI Model emphasizes the importance of making inferences in Situation model comprehension. This is because inferences fill coherence gaps to build a representation of a text during the Integration process. Kintsch (1998) conceptualized inferences primarily as *Text-based inferences* and *Knowledge-based inferences*, in line with other works of inferences (discussed in Section 2.4), for example, *inter-sentence* and *gap-filling* (Cain & Oakhill, 1998), *text-connecting* and *knowledge-based*, *on-line* and *off-line* (Graesser, 1994). With *Text-based inferences*, the previous explicit statement is re-active and related to the current clause. With *Knowledge-based inferences*, it is background knowledge, such as experiences, other texts or even the textual information from earlier sections which has been already encoded in long-term memory, that has triggered to interpret the meaning of a text. In simple terms, Text-based inferences means linking the new incoming textual information to what was previously understood from the text; Knowledge-based inferences link readers' related knowledge and experience.

Lastly, the CI Model attempts to explain the reading comprehension process by highlighting on the role of working memory. For example, at the level of Textual model comprehension, the meaning of words, clauses and sentences are built up and temporarily stored in reader's mind. The temporary storage is also referred as short-term memory. At the level of Situation model comprehension, readers construct a network from the textual information and retrieve relevant knowledge that is maintained in the long-term memory (Kintsch and Rawson, 2005). However, this issue about working memory is not a focus in this section, instead a detailed discussion will be provided in Chapter 4.

2.6.4 Section summary

This section reviewed the basic assumptions of Kinstch's (1998) CI Model. It is assumed that reading comprehension of a text is achieved through a Construction phase and an Integration phase. During the Construction phase, reading requires the reconstruction of information from sentences and text from

Microstructure (local) and Macrostructure (global). When the decoding of words becomes automatic, readers construct words in a sentence by the rules of syntax and grammar (Christmann & Groeben, 1999; Richter & Christmann, 2002). Subsequently, they integrate the gist coded across different sentences to build a Textual model. Moving towards the Integration phase, readers need to draw on their background knowledge and integrate with previous *Textual model* comprehension to build a Situation model comprehension (Kintsch, 1998; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). Furthermore, working memory plays an important role in CI model to explain reading comprehension process, which will be discussed in Chapter 4.

2.7 Towards a hybrid framework of reading comprehension assessment

One main purpose of reviewing the theories of reading comprehension was to identify an appropriate framework to assess academic reading comprehension in higher education. First, the SVR model proposes two primary components – decoding and language comprehension but it does not clearly indicate how to measure these two components. Also, this model focuses on basic word-level reading skills on the basis of research into early literacy, rather than students in higher education. The target university students in this study may rarely have decoding issues therefore a framework focusing on comprehension is needed. Further, the levels of reading comprehension theory provide a multi-dimensional perspective to assess comprehension at different levels. Barrett's Taxonomy stands out because it gives a detailed description about different levels of comprehension and specific subskills that should be included within each level.

Lastly, the CI Model conceptualizes the multiple levels of reading as two levels: the Textual model and the Situation model comprehension. To some extent, the essential ideas of the Textual model and Situation model comprehension are consistent with the assumptions in Barrett's Taxonomy. That is, the former (Textual model comprehension) is line with *Literal Comprehension and Reorganization* level because both deal with information which is explicitly stated in the text; the later (Situational model comprehension) is in line with *Inference, Evaluation and Evaluation* level because both deal with information which is implicitly stated in the text. Therefore, for the purpose of designing and developing a reading comprehension assessment, this thesis combines these two frameworks and proposes a hybrid framework of reading comprehension assessment. The hybrid framework of reading comprehension assessment is divided into literal-meaning questions and inferential-meaning questions; literal-meaning questions involve assessing reading abilities in literal comprehension and reorganization; inferential-meaning questions involve assessing ability in inferential comprehension and evaluation. Figure 2.4 gives details of the hybrid framework.

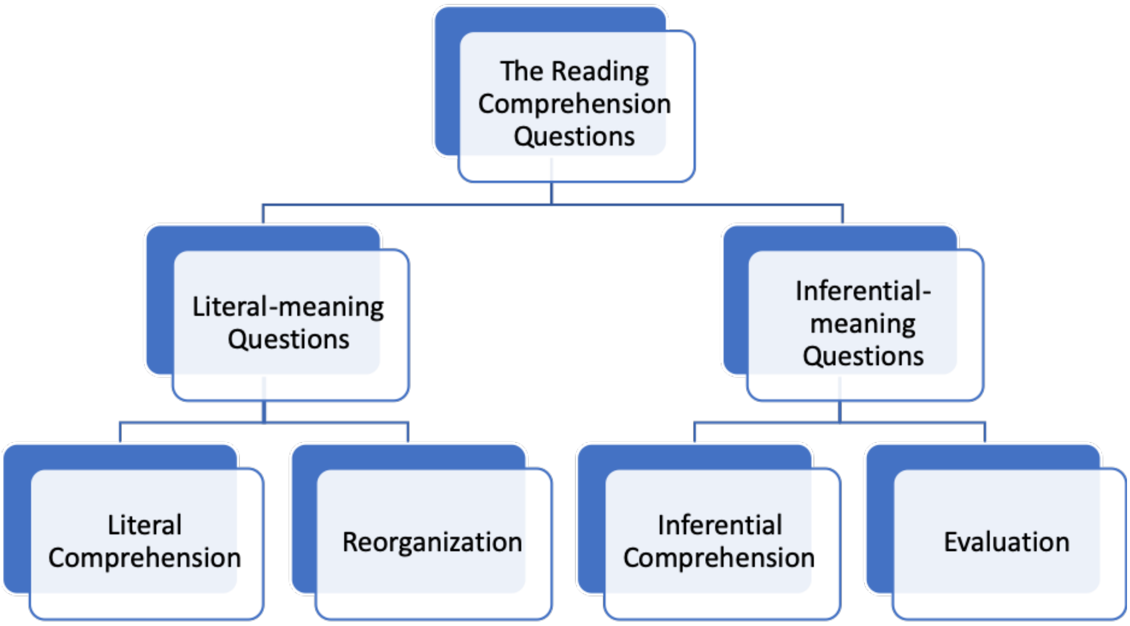


Figure 2.4 The hybrid framework of reading comprehension assessment

Nevertheless, the applicability and validity of Barrett's Taxonomy and CI Model in higher education context remains to be explored in this research. To date, most studies investigating reading comprehension of L1-English students focused on elementary school students or adolescent L2-English students at the beginning stages of learning English. Little is known about the applicability of Barrett's Taxonomy to young adult L1- and L2-English students who have completed school-level studies and entered into higher education. Further, as mentioned earlier, Barrett's Taxonomy has substantial impacts in classroom teaching of reading but there is little empirical evidence to support the existence of the levels and the subskills. Therefore, in order to assess the applicability and validity of Barrett's Taxonomy and CI Model in higher education, the proposed hybrid framework raises three questions:

1. What evidence is there for the division between literal and inferential reading comprehension?

2. Do the multiple levels of comprehension and the subskills involved in each level really exist?

3. Do the multiple levels of comprehension and the subskills involved in each level follow a linear sequential order?

Barrett's Taxonomy proposes a continuum of reading comprehension skills in which a student must first proficiently engage in tasks of literal comprehension before engaging in deeper interactions with the text. The continuum hypothesizes a linear progression of the difficulty between the levels of comprehension. It is also hypothesized that one skill within a level of comprehension is unquestionably a prerequisite for the skill within next level of comprehension (Vacca, Vacca & Gove, 1987). As this framework is built on these assumptions, three hypothesis are:

1. The levels of literal comprehension and inferential comprehension exist.

2. The multiple levels of comprehension and the subskills involved in each level exist.

3. The multiple levels of comprehension and the subskills involved in each level follow a hierarchical order.

Chapter 10 will continue to discuss the hybrid framework and respond to these hypothesis in light of the findings of the present study.

Chapter 3 Review of empirical research on screen and print reading

3.1 Introduction

To explore the differences in reading comprehension between print and screen, Chapter 2 has reviewed theories of reading in order to identify an appropriate framework to assess reading comprehension in higher education. Before employing this framework, it is necessary to look at empirical research and address the methodological issues in research into reading digital devices. The purpose of this chapter is to review empirical studies that investigate the differences between reading from paper and screen and to establish adequate criterion to conduct a comparative reading research and measure academic reading comprehension for the target participants – university-level students. This chapter is divided into four sections. In Section 3.2, it provides a brief overview of the studies encompassing the issue of reading from paper and screen. Section 3.2 identifies three current trends most relevant to this research. Section 3.3 moves on to describe the key factors affecting the reading tools and measures in greater detail: text (text length and types), task (the levels of reading tasks) and reader (academic level and the L1 of readers). Section 3.4 provides with a summary of some limitations of previous studies and considerations for the current study.

3.2 Overview

In recent decades, researchers in many fields including cognitive psychology, computer engineering and information technology have attempted to answer a fundamental question to what extent does the medium of screen, positively or negatively, change the nature of reading process and the results of reading comprehension. By the early 1990s, most studies concluded that people read slower, less accurately, and less comprehensively on screens than from papers (e.g., Gould & Grischkowsky,1984; Muter, Latremouille, Treurniet, & Beam, 1982; Smedshammar, Frenckner, Nordquist, & Romberger, 1989; Wright & Lickorish, 1983). Dillon (1992) conducted the earliest critical review of empirical research on differences between reading speed and accuracy from screens and paper. It was then concluded that although reading from a screen was slower than reading from paper, reading comprehension performance was not influenced by text medium. This review is regarded as a starting point in the discussion about print and screen reading. However, as the technology and issues with digital materials have largely improved in the last two decades, some issues discussed in Dillon (1992)'s review, such as cathode ray tubes and flickering screens, are no longer relevant. Recently, some schools are becoming paperless inside and outside classrooms, and this raises new questions for students and teachers. For instance, these paperless classrooms not only allow readers to access digital materials using a variety of devices such as tablets, mobile phones and e-readers; but also alter the size of text, and highlight texts and search related terms beyond single text with one simple click of a mouse.

Research published more recently has demonstrated mixed results. Some studies have agreed with previous conclusions, of slower reading speed and less accurate comprehension performance from screen (e.g., Kim & Kim, 2013). Others have found few significant differences in reading speed, accuracy of recall or comprehension performance between paper and screen (Margolin, Driscoll, Toland, & Kegler, 2013; Kretzschmar et al., 2013).

In a recent systematic review by Singer and Alexander (2017), they pointed out that analysing the research outcomes should not overlook the potential interactions among variables such as the variety of the devices, characteristics of the participants and text manipulation and how these interactions may impact on comprehension. According to their multi-dimensional view, Singer and Alexander's (2017) have raised important issues with respect to what areas have been well researched and what areas need further exploring (e.g. the validity and reliability of reading measures). However, Singer and Alexander (2017) did not include the research measuring meta-analysis and specify the research measuring literal and inferential comprehension when making broad comparison of reading performance. Clinton (2019) in her latest systematic review fills this gap. These reviews give important indications to help frame this study in the following aspects. First, the systematic reviews identifies three main trends most relevant to this study when comparing reading from screen versus print: 1) comprehension performances; 2) cognitive process; 3) metacognition. Second, the reviews highlight the most important factors in conducting well-designed comparative reading research.

3.3 Current trends

3.3.1 Comprehension performances

The current study investigates the differences between reading on paper versus on a screen and is concerned with three aspects: comparing reading comprehension performances, exploring cognitive processes and metacognition between the two medium. Research regarding the comparison of comprehension performances has provided a mixed picture with conflicting findings: some suggest that reading comprehension is poorer on screen than on paper (e.g., Mangen, Walgermo, & Bronnick, 2013; Noyes et al., 2014), while others report better comprehension results when participants process digital text than in print (e.g., Kerr & Symons, 2006; Verdi, Crooks, & White, 2014); still, some other studies find no significant differences in reading comprehension for print or electronic medium (e.g., Akbar, Al-Hashemi, Taqi & Sadeq, 2013; Margolin, Driscoll, Toland, & Kegler, 2013; H.K. Lee, 2004; Rockinson- Szapkiw, Courduff, Carter, & Bennett, 2013; Young, 2014).

One of the typical studies showing inferior comprehension performances from screen is by Mangen et al. (2013). They asked 15-year-old participants to answer a set of comprehension questions after reading a four-page text. Half of the participants read a PDF version on a 15-inch LCD monitor and they were allowed to scroll the pages; the other half read the same text from printed paper. The participants who read from paper text performed significantly better on the comprehension test than those who read the text on the screen. A number of studies have employed a similar experiment design (e.g., Mangen & Kuiken, 2014; Kerr & Symons, 2006; Mayes et al., 2001; Wastlund et al., 2005) and reached a similar conclusion that medium of text presentation influences comprehension performances. However, a crucial problem is that these studies did not indicated the underlying constructs of the reading measures (Singer and Alexander, 2017) and the conclusions were based on the overall results of the reading measures. In Mangen et al.'s (2013) study, the questions were designed to capture three aspects of reading comprehension (access and retrieve; integrate and interpret; reflect and evaluate) aligned with a standard assessment - OECD Program for International Student Assessment (PISA) for middle-school students. Nevertheless, the conclusions did not specify which of the three aspects of reading comprehension participants' performances tended to be superior from screen than paper. The construction of the reading measures is a key variable that influences the comprehension performances. In another study by Fisher, Lapp, Wood's (2011) for a similar population of 100 eighth grades students, different findings were found for different types of comprehension: no significant differences were found for comprehension of main themes between screen or paper but significant poorer performances were found on comprehension of local processing such as supporting details. Together, these studies suggest that reading on screen might be disadvantageous for some certain types of comprehension rather than global comprehension. Therefore, it is imperative for this research to conduct further research into specified and detailed reading constructs and compare results from different types of comprehension measures.

Linking to the hybrid framework of reading comprehension assessment that Section 2.7 proposed, in Mangen et al.'s study of *access and retrieve* is possibly more literal-meaning oriented as it requires recalling and retrieving factual information from the text, whereas *reflect and evaluate* is more inferential-meaning oriented. Meanwhile, the measures of *supporting details* in Fisher , Lapp, Wood's (2011) study could be considered literal-meaning or inferential-meaning oriented. This present study considers that the problem of Mangen et al.'s research is identified as reaching an overall conclusion from a global assessment; the problem of Fisher, Lapp, Wood's (2011) study is interpreting research results on basis of ambiguous nature of the reading constructs. As a result, it is impossible to rigorously test that at what level of comprehension or what types of reading tasks an observed discrepancy between the paper and screen conditions is likely to occur. Without such information it is not possible to conclude whether the influence is an effect from the text medium, or the comprehension demands required different levels of the reading measures, or a result of the interaction between these factors.

The main goal of the present study is to disentangle the effects of text medium on different levels of reading comprehension and to compare the reading performances under two medium through a uniquely designed experimental stimulus with specified constructs. Considerations related to the design of the reading stimulus and constructs will be elaborated in more detail in Section 6.4. The developments and adaptions of the reading measures for this study will be shown in Sections 6.5 and 6.6.

3.3.2 Cognitive processes

Attempts to explain the reasons why the observed differences in comprehension performances are found on a screen versus paper centre on cognitive processes. Assumptions of limited cognitive resources in readers' mind, referred as working memory, account for a large amount of research results (Li et al., 2013; Mangen et al., 2013; Wastlund, Reinikka, Norlander, & Archer, 2005). These studies explain that reading on a screen involves both comprehending the text and dealing with the medium such as moving a cursor or clicking a mouse. As one's cognitive resources are rather limited, the screen is likely to consume additional cognitive demands that is used for controlling the medium, leaving less cognitive capacity to process the text.

Another explanation is that the fixed layouts from printed text conveys structure information for a reader to construct a cognitive map (Hou et al., 2016). With a coherent cognitive map, a reader tends to have better spatial knowledge and more aware of one's place in the text (Crestani & Ntioudis, 2001). Such advantages from paper helps a reader to locate and recall textual contents (Li et. al., 2013). In contrast to paper, screen may weaken the spatial cues due to the instability of the screen text, which impedes a reader in constructing situational representation of the text (Kinstch, 2000) and burdens cognitive loads so as to impair comprehension. These arguments are based on an assumption that reading from screen constraints working memory capacity so that limited resources can be used. However, this lacks empirical evidence. Although a large number of studies indicate a strong correlation between reading comprehension from print and working memory capacity, this does not necessarily mean that the correlation also exists for reading comprehension from screen. Further research is needed to examine to what extent the correlation exists.

On the other hand, the assumption that that the fixed spatial cues of printed text supports reading, whereas the instability and intangibility of the digital texts impedes text comprehension (Mangen et al., 2013) might be true in the context of research with a population of primary or middle school students when the reading texts are not complex in content and not long at length. Printed texts could offer fixed markers of the physical text, such as pages, lines towards right corner or at the bottom (Piolat et al., 1997; Rothkopf, 1971; Zechmeister & McKillip, 1972); unlike printed texts, the onscreen texts may not provide tangible and physical cues for the readers (Mangen, 2006; Sellen & Harper, 2002). The problem is, in an authentic academic reading environment for university-level students, where the texts are information-rich and excessive, to what extent do the readers tend to rely on the physical spatial layouts or fixed markers to locate or recall information? Hence, these speculations require further correlational research to support or reject. This thesis aims to contribute this line of research by examining the potential relationships between working memory and reading comprehension on screen versus paper. To reach this aim, the literature will review the theories of working memory which are deeply rooted in cognitive psychology in next chapter.

3.3.3 Metacognition

Metacognition research has focused on students' metacognitive skills to explore the comprehension differences between reading print and onscreen. Metacognitive skills refer to as readers' ability to monitor their own reading, such as knowing when to re-read or read slowly to ensure adequate comprehension, being able to allocate and prioritize reading time, and choosing appropriate reading strategies for a variety of texts and purposes (Schunk & Zimmerman, 1994; Son, 2007). Slightly different from the research that examined the objective factors on reading comprehension (e.g. technical characteristics of text display, spatial layouts, navigational or annotative issues), metacognition is related to readers' subjective experience and ideas and can play a crucial role in guiding and regulating the reading process.

One important study assessing the influence of text medium on metacognitive ability was Ackerman and Goldsmith (2011). They did a comparative study of reading expository text (1000-1200 words long) between computer screen and paper under fixed-time and self-regulated time conditions on 70 undergraduate students. Poorer reading performance was observed on screen only under the condition of self-regulated time; in contrast, equal test performance was observed on paper. The researchers then suggested the differences between two mediums did not exist in cognitive but rather metacognitive aspects. In addition, less accurate prediction and monitoring was on screen possibly because participants from screen were significantly more overconfident with respect to their subsequent performance compared to print students. However, the participants in this study were reading in their native language (Hebrew) on screen so they might feel overconfident due to familiarity. When moving into the target situation of the present study where participants read electronic texts in their second language, it is interesting to ask whether students would feel overconfident or underconfident and how this might affect their metacognitive prediction and performances.

It seems fair to expect the transition from paper-based to screen-based reading entails a large degree of uncertainty with respect to the influence of text medium. These uncertainties pertain to effects of screen affordances on metacognitive ability for children (Støle, Mangen, Frønes, & Thomson, 2018). Nevertheless, little is known about how the digital screens may differently affect readers with different reading backgrounds, such as young adult learners and learners who read in a second language. Thus, this study aims to move beyond the school level of children and focus on university level of adults and those who come from a different language background.

As the studies on the metacognition have been based on participants' self-reported preference or self-perceptive data, the metacognitive accuracy was often uncertain (Ackerman & Goldsmith, 2011; Singer & Alexander, 2017). This is one of the reasons that the present study aims to emphasize the research that was included a direct measurement of participants' comprehension performance to construct a foundation to understand the possible effects of text medium (print or screen). However, this is not to say the metacognitive abilities are not important, or readers' self-perceptive information would not be collected when conducting the present research; this only means the issues on metacognition would be not a focus of this study.

3.4 Factors affecting relations between text medium and comprehension performance

In light of the view that reading comprehension is an active, constructive and meaning-making process (Goldman, 2015; Graesser, 2007; Kintsch & Kinstch, 2005; McNamara, 2012), this section reviews the empirical research that involved both print and screen reading which examined the effect by the medium on reading comprehension performance from phycological aspect. The purpose of this section is to identify the factors that should be considered to conduct a well-designed comparative reading research because they play an important role in relations between text medium and comprehension performance. Table 3.1 lists empirical studies conducted in recent decades, in which comprehension performance has been measured using an experimental test, or a combination of an experimental test with other methods (e.g., survey, questionnaires, observation and interview). Those studies which only used survey, questionnaire, interview or other methods were excluded in this table.

Table 3.1 Description of studies

Authors(year)	Text medium	No. of texts	Text length	Text genre	Level of tasks	Format of tasks	Academic level of participants	First language of participants	Language of texts	Validity & Reliability
Ackerman and Goldsmith (2011)	Computer and paper	6	1000-1200 each	Expository	Literal and inferential	MCQ	70 undergraduates	Hebrew (Israel)	Hebrew	NR
Ackerman and Lauterman (2012)	Computer and paper	5	1000-1200 each	Expository	Literal and inferential	MCQ	80 university students	Hebrew (Israel)	Hebrew	NR
Ben-Yehudah and Eshet-Alkalai (2018)	Computer and paper	1	458 words	Expository	Literal and inferential	MCQ	102 undergraduates	Hebrew (Israel)	Hebrew	Reliability reported
Chen and Catrambone (2015)	Computer and paper	3	1000 words each	Expository	Literal and inferential	MCQ; short-answer	92 university students	English	English	NR
Chen, Cheng, Chang, Zheng, and Huang (2014)	Tablet, computer and paper	3	1050-1099 words each	Expository	Literal and inferential	MCQ; summary	90 university students	Chinese	Chinese	NR
Connell , Bayless, and Farmer (2012)	Kindle, ipad and paper	1	NR	Expository	Literal	MCQ	201 university students	English	English	NR

Daniel and Woody (2013)	Computer and paper	1	one textbook chapter	Expository	General	MCQ	141 university students	English	English	NR
Dündar , H. & Akçayır, M. (2012)	Tablets and paper	NR	NR	Expository	General	NR	20 elementary students	Turkish	Turkish	NR
Davis & Neitzel (2012)	Computer and paper	2	1100 and 1200 words	Expository	General	MCQ	92 Grade 6 and 7 students	English	English	NR
Eden and Eshet-Alkalai (2013)	Computer and paper	2	600 words	Expository	General	Error correction	93 undergraduates	Hebrew (Israel)	Hebrew (Israel)	NR
Green , Perera, Dance, and Myers (2010)	Computer and paper	1	two pages	Expository	Literal	MCQ; recall	55 university students	English	English	NR

NR – not reported; MCQ – multiple choice question.

Authors(year)	Text medium	No. of texts	Text length	Text genre	Level of tasks	Format of tasks	Academic level of participants	First language of participants	Language of texts	Validity & Reliability
Hermena et al. (2017)	Tablet (ipad) and paper	2	604 each	Narrative	NR	MCQ	24 university students	Arabic	Arabic	NR
Hou, Rashid , and Lee (2017)	Tablet (ipad) and paper	1	30 pages	Narrative (comic)	General	MCQ	45 university students	English	English	NR
Hou, Wu and Harrell (2017)	Tablet (ipad) and paper	2	3469 and 3150 words (15 pages)	Expository and narratives	NR	MCQ; Sequence of events	81 adults over 50 yrs.	English	English	NR
Kim and Kim (2013)	Computer and paper	2	two pages	Expository	General	MCQ	108 high school students	English	English	NR
Kretzschmar et al. (2013)	Tablet, e-reader and paper	6	222 words each	Expository and narratives	Literal	Yes/No questions	36 young adults	English	English	NR
Magen, Walgermo, and Brønnick (2013)	Computer and paper	2	1400-1600 words each	Expository and narratives	General	MCQ; short-answer questions	72 10th grade students	Norwegian	Norwegian	Reliability reported
Margolin , Driscoll, Toland, and Kegler (2013)	Computer, tablet (kindle) and paper	10	542 words each	Expository and narratives	Literal and inferential	MCQ	90 university students	English	English	NR

Neijens and Voorveld (2018)	Tablets and paper	1	24 pages	Expository	Literal	Recall	90 university students	Dutch	Dutch	NR
Norman and Furnes (2016)	Tablet, e-reader, computer and paper	4	1000 words each	Expository	Literal	MCQ	100 university students	Norwegian	Norwegian	NR

NR – not reported; MCQ – multiple choice question.

Authors(year)	Text medium	No. of texts	Text length	Text genre	Level of tasks	Format of tasks	Academic level of participants	First language of participants	Language of texts	Validity & Reliability
Porion , Aparicio, Megalakaki,Robert, and Baccino (2016)	Computer and paper	1	one page	Expository	Literal and inferential	MCQ; Yes/No questions; True/false questions		French	French	NR
Singer and Alexander (2017)	Computer and paper	4	450 words each	Expository	Literal and inferential	Main idea; free recall; list key points	90 university students	English	English	Reliability reported
Singer Trakhman , Alexander and Silverman (2018)	Computer and paper	2	1800 words each	Expository	Literal and inferential	Main idea; free recall; list key points	57 university students	English	English	Reliability reported
Taylor (2011)	Computer and paper	1	one textbook chapter	Expository	General	MCQ	74 university students	Caucasian	English	NR
Usó & Ruiz-Madrid (2009)	Computer and paper	1	one journal article	Expository	Literal	True/False questions; open-ended questions	50 mixed level university students	Spanish	English	NR
Wästlund , Reinikka, Norlander, & Archer (2005)	Computer and paper	10	1000 words each	Expository	Shallow and deep	MCQ	72 university students	Swedish	Swedish	Reliability reported

Young (2014)	ipad and paper	3	NR	Expository	Literal	Open-ended questions	11 university students	English	English	NR

NR – not reported; MCQ – multiple choice question

This table presents a mixed picture of evidence, with conflicting findings regarding the medium of text presentation and participants' comprehension performance. Some research suggested that reading comprehension was poorer on screen than on paper (e.g., Mangen, Walgermo, & Bronnick, 2013; Noyes et al., 2014), while others reported better comprehension results when participants processed digital text than in print (e.g., Kerr & Symons, 2006; Verdi, Crooks, & White, 2014). Still, some other studies found no significant differences in reading comprehension for print or electronic medium (e.g., Akbar, Al-Hashemi, Taqi & Sadeq, 2013; Margolin, Driscoll, Toland, & Kegler, 2013; H.K. Lee, 2004; Rockinson - Szapkiw, Courduff, Carter, & Bennett, 2013; Young, 2014).

However, taking into account different attributes of texts, tasks and readers involved in the research may help to clarify this mixed picture. First, the texts in the research were various in length and text types (e.g., expository and narrative). Second, the reading tasks given to the participants were different in levels (e.g. literal or inferential) and formats (e.g., multiple-choice, short-answer, summary). Third, the participants were in a wide range of study level and age e.g. primary children or university adults) and first language (English or non-English). These factors may help to explain the contradictory research results and provide insights for the research design of this study.

The next sections will discuss in detail how the factors of texts, tasks and readers affect reading comprehension. Sections 3.4.1 and 3.4.2 will discuss text length and type; Section 3.4.3 will discuss the levels of reading tasks; Sections 3.4.4 and 3.4.5 will discuss readers' academic level and first language.

3.4.1 Text length

Text length is the first important factor to be considered when conducting a comparative study between screen and print. When the texts were shorter in length (about 500 words or no more than one page), no significant effects tended to be found for medium on comprehension (e.g., Ali et al, 2013; Dundar & Akcayiri, 2012; Eden & Eshet-Alkalai, 2013; Margolin et al 2013) or even better performance was shown for screen over print (e.g., Kerr & Symons, 2006). Conversely, when longer texts were involved in the reading process and participants were required to read more pages, the text medium tended to be more influential on comprehension outcomes, with print being the more effective and advantageous medium (Lenhard et al., 2017; Mangen et al., 2013; Singer & Alexander, 2017).

One example of non-significant comprehension differences between screen and print with shorter texts was Eden and Eshet-Alkalai's research (2013) involving 93 Israel university students. Participants were asked to edit and correct errors (six category text errors: mistyping, homophonic, morphologic, semantic, syntactic and clarity errors) while reading two papers with 600 words each in print versus personal computers. Results showed no significant difference between the average score of participants in the two formats: print = 30.40% (SD = 0.14) and digital = 30.10% (SD = 0.15), t = .30, df = 88, NS). Similar figures were also found for the scores on each type of error correction.

In the same vein, Margolin et al. (2013) used a similar length of texts with three types of reading formats. Three groups of 30 mixed level of English university students (total 90 participants) read 10 passages ranging from 505 to 571 words long. The passages included five expository and five narratives and both were presented via paper, computer screen and Kindle for the e-reader condition. After reading each passage participants completed 5 to 6 multiple choice comprehension, with a total number of 56 multiple choice questions for the ten passages. The results suggested no significant effect of medium presentation existed for the narrative and expository texts together ($F < 1$) and separately, $F(1, 87) = 1.03$, MSE = .01, $p > .36$. The texts used in Eden and Eshet-Alkalai's (2013) research were 600 words in length; likewise, the texts in Margolin et al's (2013) ranged from 505 to 571 words long. Texts at this length can be considered as relatively short; the researchers employed the texts at this length because shorter texts were thought to decrease the interferences and distractions from the digital medium (Singer & Alexander, 2017).

On the other hand, when longer texts were involved in the research, the text medium tended to be more influential on comprehension outcomes (Chen, Cheng, Chang, Zheng, & Huang, 2014; Mangen et al., 2013; Wästlund, Reinikka, Norlander, & Archer, 2005). Wästlund et al. (2005) investigated the influence of VDT (video display terminals) and print presentation of the reading comprehension by using a standardized Swedish language reading comprehension test — READ for Higher Education Entrance Examination. The test consisted of 10 pages consisting of five different tests, with an average of 1000 words each. They confirmed that the performance of 36 students in VDT condition was inferior to the performance of the other 36 students in the paper condition for consumption (multiple choice questions) and production (summarization) of information. Participants on VDT presentation also reported higher level of experienced stress and tiredness. The researchers attributed the inferiority of VDT presentation to both physiological (stress and tiredness) and psychological (increased cognitive demands) factors when processing longer online texts. Participants reading online needed to cope with a dual-task, operating reading tasks (multiple-choice questions) as well as handling computer equipment.

Mangen et al.'s (2013) research reported findings consistent findings with Wästlund (2005). Seventy-two primary students at the age of 15 or 16 in Norwegian schools read two texts (one narrative and one expository) about 1400 to 2000 words long. They were divided into two groups, with one group reading the print and the other read the same texts as PDFs without hyperlinks on a computer screen. Multiple-choice questions and short-answer questions were employed to assess students' comprehension performance. The main findings showed that students who read both texts in print scored significantly better on the RCA than students who read the text on computer screen. The authors suggested the issue of navigation such as scrolling on a computer screen would result in spatial instability, which may impede readers' comprehension on screen.

Similar findings of better performance in print with longer texts (more than 500 words) were observed with Chinese-speaking students (Chen, Cheng, Chang, Zheng, & Huang, 2014). A total number of 90 second-year college students were assigned into three groups and read three texts on the platforms of paper, tablet and computer. The three texts were about 1050 – 1099 Chinese characters at length. After reading each test participants were asked to complete five multiple choice questions and a summarization task in 80-120 words. It was found that students in the paper group performed better than those in computer-based group on multiple choice questions. The researchers explained the results on a basis of the increased cognitive load of scrolling and the unfamiliarity with the digital medium. Therefore, findings from empirical research (e.g., Sanchez &

Wiley, 2009; Wästlund, 2007) suggest that lengthy texts on screen normally poses more scrolling and page turning issues than print; if scrolling and page turning are required in sustained reading it may have an additional cognitive burden for readers.

In summary, when the texts being processed were shorter in length, there tends to be no significant effect for medium on comprehension (e.g., Ali et al., 2013; Dundar & Akcayiri, 2012; Eden & Eshet-Alkalai, 2013; Margolin et al., 2013) or comprehension was significantly better in the digital versus print medium (e.g., Kerr & Symons, 2006). When the texts length increase to more than one page or over 500 words, comprehension performances tend to be significantly better for print than screen reading (e.g., Davis & Neitzel, 2012; Mangen et al., 2013; Mayes et al., 2001). With respect to the research design target at an authentic academic reading context at an university level, the length of reading materials is believed to be over 1000 words. It is evitable to read on screen without involving scrolling or page turning. Whether such issues would impede reading comprehension for a populations of university students will be explored in this study.

Nevertheless, there are also a few exceptional studies to this length by medium pattern, showing significant differences between reading on screen paper (Hou, Rashid, and Lee, 2017; Hou, Wu and Harrell, 2017). In Hou, Rashid, and Lee's (2017) studies, the reading materials were a thirty-page narrative comic story; in Hou, Wu and Harrel's (2017) research, the reading materials were a combination of one narrative and one expository with both over 3000 words. One possible explanation for the non-significant results may differences in the types of the text. This factor will be discussed in the following section.

3.4.2 Text type

The most common types of experimental texts in empirical research were narratives and expository texts, with considerately more expository texts than narratives being used in the studies (Clinton, 2019). Narratives, such as myths and novels, are expressions of actual or fictions, event-based experiences (Graesser, 2007). Expository texts, such as newspaper, encyclopaedias, or even textbooks, inform readers by presenting information that explains principles and general behavioural patterns (Axelrod & Cooper, 1996). Generally speaking, narrative texts are considered to be easier to read than expository texts because understanding expository texts need more specialised knowledge rather than the common experiences which are involved in narratives (Graesser, McNamara, & Kulikowich, 2011). Given this, there are reasons to expect that reading on screen had a negative effect on reading performance for expository texts but made no difference with narrative texts, as was argued by Clinton in her systematic review (2019) on a basis of the empirical studies from 2008 to 2018.

In two studies where the researchers solely used narratives as the experimental test, the reading comprehension performances tended to be non-significantly different (Hermena et al., 2017; Hou, Rashid, and Lee, 2017) between screen and print. Both studies employed narrative texts, however the texts were different at length, 600 words for Hermena et al. (2017) compared with 30 pages long for Hou, Rashid and Lee (2017). Specifically, Hermena (2017) examined twenty-four native-Arabic-speaking university students reading two Arabic passages, with Passage A on iPad tablet and Passage B on paper. After reading each passage, the participants provided verbal responses to six multiple-choice comprehension questions related to the details of the text by the experimenter. Similarly, Hou, Rashid and Lee (2017) compared reading comprehension performances between three conditions – paper book, iPad tablet with full-page version, and iPad tablet with disruptive version. The participants read the comic without time limit and comprehension was assessed by a questionnaire including one open-ended question and thirteen multiple-choice questions. One possible reason for the non-significant differences between the tablet and paper conditions in both Hermena (2017) and Hou, Rashid and Lee’s (2017) studies could be because of the nature of narrative texts and detail-oriented questions. Another common reason explained by Hermena (2017) and Hou, Rashid and Lee (2017) was that the resemblance of the text displayed, such as same text layouts, control of scrolling, on two medium reduced the distractions and interruptions from e-tablet.

In conclusion, the text type, narrative or expository, is a salient attribute of experimental texts which could be contributing to the effect of the text medium on reading comprehension performance. However, the choice of using narratives or expository texts in the research is dependent on the participants’ age and academic level. Specifically, when studies involved early elementary students, researchers relied solely on narrative texts (e.g. Dündar, H. & Akçayır, M., 2012; Hermena et al., 2017; Ortlieb et al., 2014). Regarding the current research, the most common type of text that university students encounter is expository texts. Materials such as a chapter from a text-book or a journal article are considered in the current research design. In turns, the choice of text type will directly influence the task demand or the level of reading tasks.

3.4.3 The levels of reading tasks

The levels of reading tasks are considered to be a key issue in the current research design. The level refers to whether researchers measured reading comprehension as a global construct or as separate constructs.

On one hand, a number of researchers employed a general reading measurement, usually a standardised test (e.g., Ben-Yehudah and Eshet-Alkalai, 2018; Chen et al., 2014; Davis & Neitzel, 2012). These tended to measure reading comprehension as a global construct, and the levels of reading comprehension were not separated. Only looking globally at comprehension outcomes risks overlooking differences that might only be found when examining different levels of comprehension. Without such information, it is difficult to compare various research results. For example, Daniel and Woody (2013) used a general test and the results showed no significant results with the reading performances between screen and print; contradictory results with the performances were found in Taylor (2011). A possible reason for this disparity could be that the test used by Daniel and Woody (2013) asked only literal-level questions about explicit information in the text while the questions posed by Taylor (2011) were literal only, inferential only or a combination of literal and inferential questions. In this sense, the reading tasks were not equivalent in nature and thus the conclusions are incomparable.

Another danger of using a standardised test concerns the lack of detailed description of test specifications when reporting the research outcomes. Usually there may be information provided in standardised test manuals regarding test construction, validity and reliability; but such information is often underreported in the research, as indicated in Table 3.1 (only a few authors reported the reliability or validity). When using standardised tests, test specifications are pre-determined by the test designers and cater for a specific assessment purpose and a group of target test takers. The reading constructs are implicit for other researchers and thus the research analysis could only be on general outcomes. Without knowing that, the researchers are unlikely to be able to tell, at what types of constructs, participants might be advantages or disadvantages when reading on screen or paper.

On the other hand, some studies used researcher-developed reading tests and attempted to cover multiple levels of comprehension (e.g., Ackerman and Goldsmith, 2011). Still, the test specifications/constructs underlying the multiple test items (e.g., multiple-choice questions or short-answers) were rarely explicitly stated. This gives rise to ambiguity about the definitions of the level of reading comprehension, resulting in inconclusive results. For example, Singer and Alexander (2017) used main ideas and supporting ideas to measure comprehension performance. They claimed that their research was one of the few studies between 2008 and 2017 that measured reading comprehension from “different levels”, as opposed to those using a global construct. Indeed, Singer’s study was a good attempt to measuring reading comprehension from two test formats (main ideas by summary and recalling details by MCQ) while the researchers themselves considered them as different levels. It can be argued that main idea and recalling details were similar in nature – both retrieving factual information from the text involving literal comprehension and text memorization, rather than making inferences or understand intended meaning.

Another example is a study by Hou, Wu and Harrel’s (2017) who used multiple choice questions as literal comprehension measurement, and summarization as inferential comprehension measurement, because summarization involves reorganization of the test. They regarded the two test formats as separate measurements of shallow and deep comprehension. Similarly, Chen et al. (2014) tested ‘deep level comprehension’ by means of a summarization task and reported no significant differences for deep level comprehension across different medium platforms. In contrast to, Hou, Wu and Harrel’s (2017) and Chen et al. (2014), USO (2009) used two types of True/False question and open-ended question to test literal understanding, as explained by the authors, the answers to the questions can be explicitly found in the text. Definitions of literal/inferential measurements therefore vary across different studies.

However, one should note that the nature of the literal or inferential comprehension questions, may not be determined by the test formats (multiple-choice or summarization) but the underlying constructs. A multiple-choice question could test deep level of comprehension or inferential understanding of the text if it asks implicit textual information; a summarization task could also test literal understanding if it asks factual information of the text. To this end, lacking explicit definitions of the level of task would give rise to the difficulty in interpreting the results, the present research aims to fill this gap by designing a reading comprehension assessment to measure both literal and inferential level comprehension, on a basis of explicit and detailed reading constructs.

3.4.4 The academic level of readers

Associated closely with text types and the level of the task is readers' academic level. Specifically, when the participants were elementary students aged from 5-10, the research tended to solely include narrative texts (e.g., De Jong & Bus, 2004; Jones, 2011; Kim & Anderson, 2008; Ortlieb, Sargent, & Moreland, 2014) and the comprehension questions were primarily literal in nature. These studies revealed that when young children read the texts which are short and simple and answer the multiple-choice questions which are asking facts and details, the medium seems to have little effects on their comprehension performances (e.g., De Jong & Bus, 2004; Dundar & Akcayir, 2012). However, for older populations, such as high school students or students beyond school-levels, the research tends to include expository text at various lengths or combine multiple narrative and expository texts (e.g., Eshet-Alkalai & Geri, 2009; Lenhard et al., 2017). When the reading tasks often are complex and require a deeper level of understanding for older readers, compared with those only require shallower processing of text for students at earlier grades, the medium seems to matter in comprehension performances (e.g., Lenhard et al., 2017; Mangen et al., 2013; Singer & Alexander, 2017). Under such circumstances, text medium plays a more significant role in comprehension performance. Therefore, empirical research for this thesis should take into account the academic level, in conjunction with the factors of text type and reading task demands, as crucial parts of an integrated research design.

3.4.5 The L1 of readers

Previous studies including participants who were reading in their first language and used the same texts seem to suggest similar results for both the screen and paper reading conditions. For example, as shown in Table 3.1, the studies that used the participants' L1 texts included English (Connell et al., 2012; Davis & Neitzel, 2012; Hou, Rashid & Lee, 2017), Turkish (Dundar and Akcayir, 2012), Chinese (Chen et al., 2014) and Arabic (Hermena, 2017) and these have found non-significantly different performances between screen and print. However, two studies including Norwegian (Magen, Walgermo, and Brønnick, 2013) and Herbrew (Ackerman and Lauterman 2012; Ben-Yehudah and Eshet-Alkalai, 2018) have shown inferior screen performances.

There is limited research focusing on participants who were reading the texts in their second language. Two studies, shown in the Table 3.1, suggested mixed results. One research focused on L1-Spanish-speaking participants reading English academic texts and found the medium did not affect reading comprehension performances. Usó & Ruiz-Madrid (2009) studied fifty university students were from the discipline of Tourism but also engaged in an English for Academic Purpose course. The reading text was taken from an online journal. Twenty-five read it in a printed format, and a separate twenty-five read it in its online version with hyperlinked headings and references. Results showed that hypertextual medium did not affect these ESL learners' overall reading comprehension which was measured by five True/False questions and five open-ended questions.

Conversely, another study by Stakhnevich (2002) with a mixture of L1 participants reading an English text found superior performances from screen. The ninety participants were the students who had arrived in US for no more than two weeks and whose first language was non-English, including Chinese, Russian, German or Japanese. The participants performed better on 20-item multiple choice comprehension questions after reading a 2500-word text digitally. While this seems an exception to the length by medium pattern – poorer performance on screen with longer texts (discussed in Section 4.4.1), there are several specific features embedded in the online version that could result in such deviation. The online text included an online glossary and dictionary access, which can be helpful for the ESL readers. Moreover, the small sample size (n =

31) and the use of a single comprehension measure (only MCQ) can be considered methodological limitations. The additional online tools should also be considered as a factor affecting the performances in the present research design.

In summary, the medium of screen itself may not have significant impacts, but benefits for ESL students may be present when texts include embedded tools such as hyperlinks or online dictionaries. However, previous research is rather limited, neither Singer and Alexander 's (2017) or Clinton's (2019) systematic review report on research with this population of learners. Although the researcher (Clinton, 2019) agreed that L1 of the participants is a crucial factor affecting the influences of the text medium on comprehension, the relations between participants' L1 and text medium is underestimated. Thus, the current study focuses on the similar population and context – English as Second language learners who are new to an English-speaking university but includes a larger sample size and multiple comprehension measures.

3.5 The current study and research questions

To provide a multifaced view of text, task and reader to examine the effects of medium on reading comprehension, the present study questioned some of opinions in previous research (Margolin et al., 2013; Magen et al., 2013). To begin with, in Margolin et al.'s research (2013), they contended that little comprehension performance differences existed among three types of text medium (paper, computer screen and Kindle) and therefore the medium would not significantly limit reader's comprehension during reading. However, this view may overlook the factors such as the text length and reading task type which could have interacted with the medium and contributed to the non-significant results. The texts used in the research included 10 passages with about 500 words each, and this implied that the comprehension for each text may be not as demanding as one single passage with about 5000 words. Thus the effect of medium on reading comprehension for a short text may be not significant but it could be significant for a longer text.

Margolin's (2013) analyses were based on a general test, which was made up of fifty-six multiple-choice questions, for both narrative and expository text.

Some of answers could not be simply gained from a single word from text or from memory but required thought and understanding. This indicated that the questions/reading tasks could be various in nature, such as recalling facts or details (literal), or requiring understanding beyond textual meaning (inferential). Furthermore, the number of different types of questions was not specified in Margolin et al. (2013) research. In this regard, it was unknown whether the non-significant result may be due to performance on the literal questions or inferential questions. If a certain type of comprehension question accounted for the majority of 56 multiple choice questions, for example, the number of literal-level questions were much more than inferential level questions, the analysis on a basis of both type of questions would mask the small difference of the inferential type of questions. Therefore, it may be true that the medium had no significant effect on reading comprehension performance for one type of comprehension questions; still, it was unclear whether the text medium would have the same effect on reader's comprehension for another type of comprehension questions.

Second, the present study doubts the assumption that fixed spatial cues of printed text supports reading whereas the instability and intangibility of the digital texts impedes text comprehension (Mangen et al., 2013). This might be true for the primary-level students in Magen et al. (2013) research when reading texts were not rich and complex. Participants could rely on the fixed markers of the physical text, such as pages, lines towards right corner or at the bottom (Piolat et al., 1997). In contrast to printed texts, the onscreen texts could not provide tangible and physical cues for the readers (Mangen, 2006; Sellen & Harper, 2002). However, in an authentic academic reading environment for university-level students, where the text was complex in contents and excessive in length, it remains to be explored whether the readers still rely on the physical spatial layouts and fixed markers to memorize and understand information from the text.

Lastly, it is necessary to provide valid evidence from the perceptions of the readers, assuming the positive effect of fixity from printed texts and negative effect of intangibility from digital texts. In a number of empirical studies revealing better reading performance on print, the authors tend to interpret the results by the text materiality. However, this assumption is not well evidenced from the perspective of readers. That is, to what extent, it is true that readers would utilized the fixed cues or spatial layouts to help comprehend a printed text and provides theoretical and empirical justification for each factor – the length of the text, the levels of reading tasks and the characteristics of the reader. For example, the length of the text is set above 500 words, which is the normal length in an authentic academic reading task, Chapter 6 will describe this in detail.

Therefore, the comparative research design of this study takes previous research limitations into consideration and considers the interactions between the factors of text, task and reader and text medium. In order to examine whether there are differences in reading comprehension performances between screen and print at different levels of comprehension for L1-English and L1-Chinese-speaking university students, the hybrid framework of reading comprehension assessment (discussed in Section 2.7) helps to shape the research questions of this study:

1. Do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

2. For L1-English-speakers only, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

3. For L1-Chinese-speaking participants, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

4. How do subskill scores differ across different text presentation for L1-English- and L1-Chinese-speakers together?

5. How do subskill scores differ across different text presentation for L1-English-speakers only?

6. How do subskill scores differ across different text presentation for L1-Chinese-speakers only?

7. When reading on screen how do the subskill differ between L1-English- and L1-Chinese-speakers?

8. When reading on print how do the subskill scores differ between L1-English- and L1-Chinese-speakers?

Chapter 4 Theoretical perspectives of working memory

4.1 What is working memory?

> [Working memory is] … a temporary storage system under attentional control that underpins our capacity for complex thought. (Baddeley, 2007, p. 1)

The term working memory was initially invented in 1960s and evolved from the concept of a unitary short-term memory (STM) by Atkinson and Shiffrin (1968). It has been expanded into the idea that "working memory is a temporary memory system that is used to help to undertake complicated cognitive tasks such as reasoning, learning and language comprehending" (Baddeley, 2010, 2:05). On this basis, working memory has been conceptualized in various ways. For example, working memory has been defined as the use of temporarily stored information in the performance of more complex cognitive tasks (Hulme & Mackenzie, 1992); or it is referred to a mental workspace for manipulating activated long-term memory representations (Stoltzfus , Hasher, & Zacks, 1996). A more recent definition is an active memory system which is responsible for temporary maintenance and simultaneous processing of information (Bayliss , Jarrold, Baddeley, Gunn, & Leigh, 2005). Thus, working memory can be seen as the quantity of cognitive resources which are utilized to temporarily store and simultaneously manipulate information during thinking and reasoning tasks.

This definition that working memory (WM) requires storage and processing of information is rooted in the concept of short-term memory (STM) (Baddeley , 1992). Sometimes, WM and STM are used interchangeably by many psychologists and researchers (Dehn , 2008). However, for the purpose of this thesis, it is important to point out that WM and STM are separable and distinguished in the following aspects. First, STM passively holds information whereas WM actively processes information. Second, STM has no management functions while WM has executive and coordinative functions. Third, STM can operate independently of long-term memory but WM relies heavily on long-term memory. The current study is interested in WM; pointing out the distinctions is mainly because a vast majority of theories and empirical research on WM has incorporate STM as a subsidiary system and tend to include STM without discriminating their differences. Such differences about how STM and WM have developed could have a clear picture when dating back to earlier cognitive theories and models on human memory.

The next section will review some fundamental frameworks that contribute to understanding human memory because they can help to understand text processing differences between screen and print.

4.2 Models of memory

Despite numerous theoretical models explaining human memory in the literature, this section focuses on four fundamental models in cognitive psychology which are most relevant to this study. The models can help to understand text processing differences between screen and print and to explain the comprehension performances from cognitive psychology perspective. Section 4.2.1 starts with Information Processing Model which underpins the basic approach to examine human memory. Section 4.2.2 moves to one of the most accepted and influential models proposed by Atkinson and Shiffrin (1968), the three-component Multi-store Model. Though influential, the linearity of this Multi-store Model underplays the complexity of the human brain and has received increasing criticism, and thus various theories were developed in order to further assess the inherent processes. Following this line of thought, Section 4.2.3 describes the Levels of Processing Model created by Craig & Lockhart (1972), arguing that the degree to which the information is elaborated will affect how well the information was learned. Section 4.2.4 details the original and revised Multi-component Model of working memory by Baddeley and Hitch (1986 & 2000). Of these, Baddeley and Hitch's (2000) model will be highlighted as it provides a theoretical framework to examine the relationship between working memory capacity and reading comprehension across different medium.

4.2.1 Information Processing Model

The concept of WM is deeply rooted in the cognitive model of human mental processing, which has been widely known as the Information Processing Model since the 1960s. This model uses computer processing as a metaphor to describe how human minds tackle the information from sensory input to storage and behavioural responses. For example, the eye receives visual information and decodes information into the brain. This information can be stored, retrieved and transformed using "mental programs", with the results being behavioural responses. According to this model (see Figure 4.1), the cognitive processing is constituted by a set of separate but interconnected subsystems, with memory components being the heart of the system (Gagne, Yekovich , & Yekovich, 1993). There are also a variety types of information processing involved in this model, for example, selective perception, encoding, storage retrieval and response organization. At its core, the Information Processing Model identifies WM as one component of information processing. Thus, WM can be considered as a cognitive processing approach to understand mental functions in human's mind.

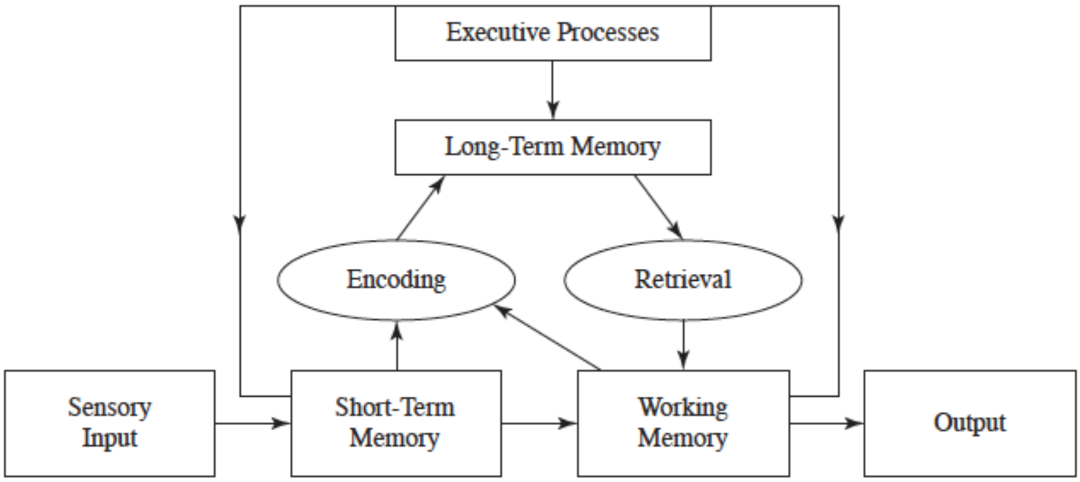


Figure 4.1 Example of an information processing model (Dehn, 2008, p.12)

4.2.2 The Multi-store Model by Atkinson and Shiffrin

The Multi-store Model by Atkinson and Shiffrin (1968) is one of the most accepted and influential development in accord with the information processing model in 1960s and 1970s. The Multi-store Model, also referred to as the modal model (Atkinson & Shiffrin,1968), divides memory into three types of components (see Figure 4.2): sensory memory or store, short-term memory (STM) and long-term memory (LTM). The first component, sensory memory or store, also known as sensory register, is associated with three types of perceptual processing: a. iconic memory (visual information), b. echoic memory (auditory information), c. haptic memory (tactile information). These types of sensory memory can be only stored for a very short duration, for example, iconic memory lasts for milliseconds; haptic memory lasts for one or two seconds

(Goddard, 2012). The iconic memory – holding written information such as letters, words or pictures, plays a key part in reading either on paper or on screen. When given attention by human brains, the contents in sensory memory are selected and filtered into the *STM* – the second component in this model. The STM can hold information between 15 and 30 seconds, without being rehearsed, according to Atkinson and Shiffrin (1971). When the information is being rehearsed through subvocal repetition (maintenance rehearsal), it will maintain in the STM and transfer into LTM – the third component. Meanwhile, if new information enters into STM it replaces the previous information already in there; consequently, the information that cannot be rehearsed or transferred into the LTM will disappear. Therefore, the Multi-store Model assumes a sequential three stages for information processing: the information is first passed from sensory memory and then filtered into the STM via attention; when being rehearsed the information in STM continues transferring into the long-term memory.

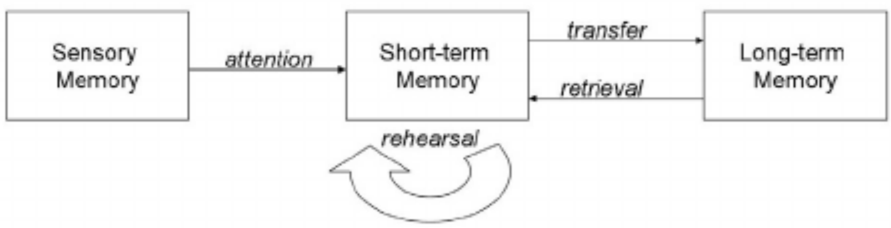


Figure 4.2 The Multi-store model by Atkinson-Shiffrin (1968) (Kelleher & Dobnik, 2019)

One of the most important elements to understand multi-store model is the *short-term memory* capacity. This refers to the amount of information that can be stored in short-term memory. As described by Atkinson-Shiffrin (1968), STM has very limited capacity. Information in STM will be quickly forgotten if not rehearsed. Still, the information that can be rehearsed at one time has long been assumed to be approximately seven units for a healthy adult, according to a classic article – *Magical number seven, plus or minus two* by Miller (1956). The author speculated the idea that STM could only hold five to nine chunks of information (seven plus or minus two). Later, the concept of chunking and the limited capacity of short term memory is considered as a consensus of subsequent theories of memory.

A chunk is referred to as a meaningful unit, or a memorable group of items that combines discrete pieces of information (Miller, 1956), for example, creating a pair or a set from two or several adjacent words in a written text. On this basis, the number of the chunks is limited to five to nine whereas the size of each chunk can be varied depending on individual's expertise, language proficiency, age and strategies. In other words, chunking can be used as a method for keeping groups of information accessible for later recall. For example, when given a list of individual words to remember one can divide them into several chunks according to different strategies; the more words in each chunk the larger capacity in STM. Therefore, the STM has very limited capacity but it may be increased by applying strategies. In terms of reading English written text, it is reasonable to speculate the more familiar with the reading strategy and more proficient with English language, readers can hold greater textual information in STM. This provides some insights to examine the differences between L1-English and L1-Chinese participants in the present reading test.

However, Atkinson and Shiffrin's model has been criticized for being an oversimplification by presenting a three-stage sequential theory. The linearity of this theory, though influential, tends to reduce the complexity of the human brain, and thus various theories have been developed in order to further assess the inherent processes. This thesis focuses on two of the most important theories: the Levels of Processing by Craik, & Lockhart (1972) and the Multi-component Model of working memory by Baddeley and Hitch (2000). The first criticism of Atkinson and Shiffrin's model (1974) is concerned with transferring information from the STM to the LTM. There is little evidence supporting the hypothesis that rehearsal is the key process to facilitate transferring information into LTM (Craik, & Lockhart, 1972). For example, the maintenance rehearsal tends to ignore the fact that other factors such as motivation or strategy also underpin the learning process. These limitations are dealt with by the Levels of Processing Model by (Craik, & Lockhart, 1972). The next section 4.2.3 will discuss the Levels of Processing Model in detail.

Second, the Multi-store Model has been criticised for being oversimplified because it suggests STM operates in a single and uniform faculty. In contrast, Baddeley and Hitch (1974) argued that STM is more than one simple unitary store and consists of different components, on a basis of evidences from neuropsychology. Baddeley and Hitch (1986) later developed the idea of short-term memory and multi-store model and proposed the concept of *WM* and *Multi-component Model*. The term WM gradually replaced STM; the Multi-component Model gradually replaced Multi-store Model (Baddeley, 2000).

In summary, this section has reviewed the three components in the Multi-store Model by Atkinson and Shiffrin's – sensory store, STM and LTM , and their linear relationship. Most importantly, the concept that STM has limited capacity not only has influence on the idea of working memory which was later proposed by Baddeley and Hitch (1974), but also becomes the central idea in cognitive psychology, to be discussed in Section 4.2.4. However, Atkinson and Shiffrin's model has received criticism from two aspects: oversimplifying the transfer process from STM into LTM (Craik & Lockhart, 1972); and assuming STM is unitary (Baddeley and Hitch,1974). These limitations has been dealt with in the Levels of Processing Model and Multi-component Model, discussed in the next two sections.

4.2.3 The Levels of Processing Model

Craik and Lockhart (1972) propose the Level of Processing Model in order to improve on Atkinson and Shiffrin's account of transfer from STM to LTM. This model focuses on the depth of processing involved in memory rather than its structure. The depth of processing is defined by Craik (1973, p.48) as "the meaningfulness extracted from the stimulus rather than … in terms of the number of analyses performed upon it." This depth of processing has two levels: shallow processing and deep processing. *Shallow processing* involves two forms: 1) visual processing, also referred to as structural processing, means encoding the physical qualities by visual properties such as the shape, colour, letter; 2) phonemic processing refers to encoding auditory information such as the syllable or the sound of a word. *Deep processing* involves in sematic processing which means encoding the meaning of a word and linking it with existing knowledge. To give an example, if one is trying to decode a word *elephant,* the shallowest level is to look at the word *elephant* in general such as colour and size (visual processing); subsequently, the less shallow level is to think the number of syllables or the pronunciation of the word (phonemic processing); consequently, the deepest level is to link the meaning of the word to the existing knowledge about an elephant (semantic processing). This view is consistent with Simple

View of Reading, turning back the reading comprehension theories in Chapter 2, decoding written texts draws on phonological knowledge. Therefore, the Levels of Processing model focuses on the levels of information processing and assumes two principal levels – shallow and deep.

The key assumption of this model is that the degree to which the information is elaborated upon will affect how well the information is learned. For example, deep levels of analysis produce more elaborate, longer lasting, and stronger memory traces than shallow levels of analysis. This implies that the deeper the level of processing, the easier the information is to recall; the shallower the level of processing, the easier the information is to forget. It is, therefore, believed that semantic processing which involves in making deeper meaningful analysis is more effective to remember and recall information compared with shallow processing (visual and phonemic processing). Regarding the present study, it is interesting to ask whether the medium, screen and print, affect how well readers learn and elaborate the textual information; or on what type of medium the level of written textual processing is deeper/shallower. Chapter 11 will response to this issue in light of the research findings.

The Levels of Processing Model developed Atkinson and Shiffrin's model (1968) develops two aspects of our understanding of the links from short-term memory to long-term memory. First, Craik and Lockhart (1972) distinguished between maintenance rehearsal and elaborative rehearsal in their level of processing model. Maintenance rehearsal involves repeating analyses previously carried out, whereas the elaborative rehearsal involves deeper or more semantic analysis of the information. Yet, Atkins and Shiffrin (1968) considered *rehearsal* primarily as *maintenance* rehearsal in their multi-store model. More recently, Raaijmakers and Shiffrin (2003) also agree on the importance of elaborative rehearsal in memory processes. Second, the Levels of Processing Model assumes that it is the elaborate rehearsal by deeper encoding that leads to better STM. Still, this view contrasts with Atkinson and Shiffrin's (1968) idea that merely (maintenance) rehearsal by holding information in the STM would guarantee transfer to LTM. Therefore, the development of Craik and Lockhart's (1972) is the idea that elaborative rehearsal (deeper process with information) contributes to LTM.

Regarding the course of reading on screen and on paper, the ideas of shallow and deep processing will help to explain the interactions between the reader and the text across difference medium.

4.2.4 The Baddeley and Hitch model of working memory

The Multi-component Model of working memory has its roots in the traditional proposal of STM (Broadbent, 1958) and is supported by neuropsychological data from patients. Traditionally, it is assumed that STM is principally verbally based and plays a useful role in general cognition (Atkinson & Shiffrin, 1968). However, neuropsychological evidence found that impairment of patients' verbal short-term memory had little impact on their broader cognition (Shallice & Warrington, 1970). This presents problem to the Multi-store/modal Model in that impaired STM may not affect input into LTM . In dealing with such issue, Baddeley and Hitch (1986, p. 34) replaced the idea of STM with the term *WM*, defined as "a system for the temporary holding and manipulation of information during the performance of a range of cognitive tasks such as comprehension, learning and reasoning." They stepped forward to propose a Multi-component Model of working memory (1974 & 2000) in which some components are responsible for information storage while others are responsible for information processing.

Baddeley and Hitch's original model (1974) comprised three components of WM — a central executive, a phonological loop and a visuospatial sketchpad, shown in Figure 4.3. However, Baddeley and Hitch (2000) proposed a revised model (Figure 4.4) since the original model received criticism for ignoring the links between various components in working memory and long-term memory. The major change is adding an entirely new component "episodic buffer". These four components will be discussed fully as follows.

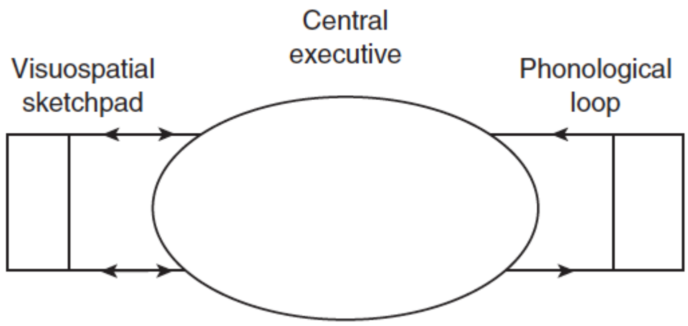


Figure 4.3 The original Baddeley and Hitch (1974) model of working memory (Baddeley , 2015, p.19)

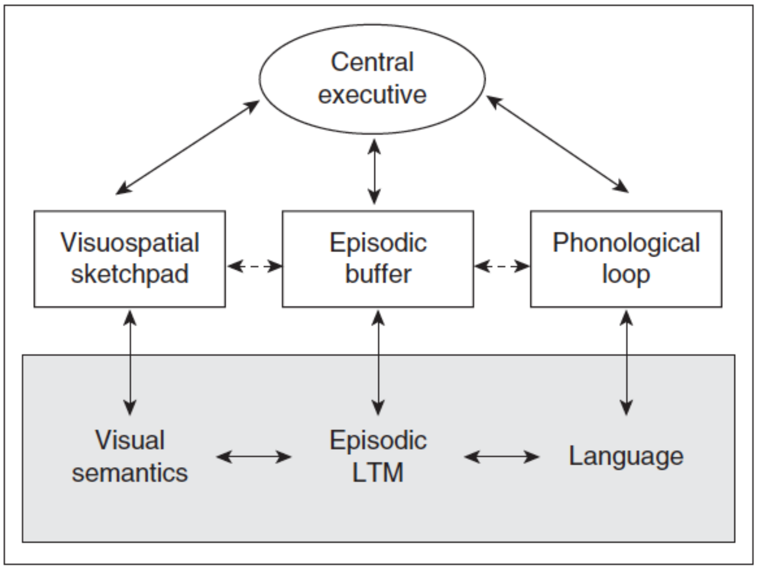


Figure 4.4 The revised Baddeley and Hitch (2000) model of working memory (Baddeley, 2015, p.19)

1. The central executive

The central executive is considered as the essence of WM. It is used for controlling attention, ensuring that WM resources are directed appropriately to achieve the tasks that have been undertaken. Meanwhile, the most important role of the central executive is to coordinate information from the other two subsystems or slave systems: the phonological loop and the visuospatial sketchpad. Baddeley (1986) uses a metaphor of a company to describe the way in which the central executive operates. The company boss (central executive) makes decisions about which issues need attention and which can be overlooked. The boss can only deal with a limited number of tasks (limited capacity with working memory) concurrently but can select strategies and coordinate information from different aspects (resources in phonological loop and visuospatial sketchpad). Thus, the central executive in WM provides overall regulation and control of the working memory system and coordinates activities from different components.

The phonological loop and *the visuospatial sketchpad* are two temporary storage slave systems under the central executive. Unlike the central executive, both of the two subsystems are regarded as passive storage mechanism without any capacity for controlling attention or decision-making (Henry , 2011). The information held by the two slave systems can only last for short periods of time and fades rapidly. Specifically, the phonological loop is for holding speech-based information while the visuospatial sketchpad is for holding visual and spatial information. For example, in the process of reading print, readers may make decisions what information to focus according to reading purposes (central executive) and then direct their cognitive resources to the written words (phonological loop) or to the sequences and location of the written words (*visuospatial sketchpad*). When moving to reading on screen, researchers (Mangen et al., 2013), as discussed in Sections 3.3, tended to speculate the visuospatial WM was impaired due to the instability of the text presented on screen. Yet it is uncertain that whether readers on screen make use of both phonological and visuospatial WM or less depend on visuospatial WM to process the textual information. This present study aims to further explore this issue.

2. The phonological loop

The phonological loop, according to Baddeley and Hitch (1974), is divided into two further subcomponents: the phonological store and the articulatory rehearsal mechanism. The phonological store acts as an inner ear for holding speech-based form information which involves in spoken words and written words. Spoken words enter the store directly whereas the written words will be first converted into spoken code before entering the phonological store. The articulatory rehearsal mechanism acts as an inner voice for repeating and refreshing the information. This structure has been supported by a range of experiments exploring the properties (Baddeley, Gathercole and Papagno , 1998) of the phonological loop model.

3. The visuospatial sketchpad

The visuospatial sketchpad is the other storage slave systems in the working memory model. This component is responsible for holding visual and spatial information for short period of time so that it can be used for thinking, remembering and processing tasks (Logie , 1995). The visual features of information refer to “what” such as form and colour; the spatial features refer to “where” such as location. In a more recent study, Baddeley (2007) has extended the role for visuospatial sketchpad as:

The sketchpad is a subsystem that has evolved to provide a way of integrating visuospatial information from multiple sources, visual, tactile and kinaesthetic, as well as from both episodic and semantic long-term memory (p. 101).

This means the visuospatial sketchpad is hypothesised to deal with visual, spatial and kinaesthetic information. This involvement of kinaesthetic information may require further research (Smyth & Pendleton, 1989), however, a number of studies (e.g. Vicari , Bellucci & Carlesimo, 2006) have supported the existence of at least two separate mechanism within the sketchpad to deal with information about visual appearance such as colour, shape and pattern as well as information for spatial location such as single locations or movements between locations. Nevertheless, there is a lack of research on visuospatial sketchpad by memory researchers compared with phonological loop (Baddeley, 2007) probably because verbal materials are more tractable in terms of experimental manipulation.

However, the conceptualization of the visuospatial sketchpad is closely associated with the current research. Regarding the onscreen and print reading comprehension research, previous assumptions indicate that readers performed better on paper because they could see the tactile fixed cues and feel the spatial extension of a physical text (Kerr & Symons, 2006; Magen, 2006 & 2010; Mange et al, 2011; Sellen & Harper, 2002). Yet there is limited empirical research to provide evidences for this assumption. Therefore, one of aims of the current study is looking for some empirical evidences to support or reject the hypotheses found in literature. Also, Baddeley and Hitch's (2007) theories on visual sketchpad of working memory can help to explain the research findings that the reading comprehension performance on paper is superior or inferior to onscreen performances.

4. The episodic buffer

The episodic buffer is the most recent subcomponent added to the WM model (Baddeley, 2000) on a basis of the original model (Baddeley & Hitch, 1974). Unlike the phonological loop or the visuospatial sketchpad which hold particular types of information (e.g. auditory, visual, spatial or kinaesthetic), the first feature of the episodic buffer is that it can deal with information from many different modalities and bind them together (Baddeley, 2007). As in Baddeley 's (2007) description,

[the episodic buffer is] a temporary system that is able to combine information from the loop, the sketchpad, long-term memory, or indeed from preconceptual input, into a coherent episode. (p. 148)

For example, the information about a scene may involve in visual information, speech sounds and movements. It is the episodic buffer that is hypothesised to link such information together into a meaningful unit. In terms of reading comprehension, the concept of episodic buffer is very similar to the Integration phase of CI Model, explaining the importance of linking textual information and activating existing long-term knowledge to reach a meaningful understanding of the text.

A second feature of the episodic buffer is that it links the central executive and LTM so that the stored knowledge can be assessed and utilised during ongoing memory and task processing. This episodic buffer, as Baddeley notes (2007), can act as a backup store to supplement the phonological loop or the visuospatial sketchpad as well as acting as a link to long-term memory. Maughan and Brown (1991) provided evidence to support the notion of an episodic buffer which provides access to LTM knowledge. They found that remembering lists of non-words was more difficult than remembering familiar words; if students learnt the meaning of the nonwords, their recall performances improved. The episodic buffer is assumed to be the mechanism that can help to improve the recall. Nevertheless, the roles of episodic buffer and its interrelationships with other components of WM continue to be refined and specialised (Baddeley, 2000) as to be determined by further research.

Therefore, two main features of the episodic buffer are: a. blending information from different sources into a coherent memory experience; b. allowing long-term knowledge to be accessed and utilized in the working memory system. However, the component of episodic buffer is a relatively recent addition to the multi-component working memory model (Baddeley, 2000 & 2007) and requires further research.

4.3 Key assumptions on working memory capacity

As communicated in Section 3.3.2, a number of studies comparing reading from screen versus print use working memory capacity to explain the differences in comprehension performances (Mangen et al., 2013). A key debate about WM is whether the working memory resources are distributed in separate subsystems or shared by a single unitary system. Some researchers in favour of General Capacity Hypothesis (e.g. Engle , Cantor & Carullo, 1992) argue there is a single pool of resources flexibly allocated between processing and storage components. That means, when the processing demands of the task are high, less capacity will be available to meet storage requirements (Daneman & Carpenter, 1980). By contrast, others supporting Separate Resources Hypothesis (e.g., Halford, Wilson & Phillips, 2001) hypothesise that there are separate capacities limits for short-term memory components (e.g. phonological short-term memory or visuospatial short-term memory) and processing (the central executive). Nevertheless, the debate is far from reaching an agreement as there is evidence to support general capacity hypothesis (Engle et al., 1992; Turner & Engle, 1989) as well as separate capacity hypothesis (Duff & Logie, 2001; Towse, Hitch & Hutton, 1998). More recently, working memory theorists suggest “there are most likely separate resources, with separate limits for storage and processing, while at the same time some shared general resources” (Dehn, 2008, p.40). These different assumptions have an impact on the measurements of WMC (to be discussed in Section 4.4).

Despite of the ongoing debate on general capacity hypothesis or shared capacity, the general consensus by cognitive psychologists is that the capacity of working memory is rather limited, meaning that human cannot store and manipulate endless amounts of information, which is usually 7 items as discussed in Section 4.2.2. In other words, the types of processing and remembering tasks that can be undertaken concurrently will be constrained by working memory resources. This study considers this important conception as a theoretical foundation to explain the differences on reading performances (e.g. Magen et al., 2011).

4.4 Measurements of working memory capacity

Although there are different assumptions regarding WMC, a more practical issue is what specific tasks can be taken to measure the WMC for the purposes of empirical research. From a theoretical perspective, a distinction was made between simple WM tasks and complex WM tasks. From a methodological perspective, a distinction was made between verbal and non-verbal. This section discuss three different span tasks that are widely applied.

Theoretically, a distinction is often made between simple WM (or referred to as short-term memory) tasks and complex WM tasks. The simple working memory/short-term memory tasks emphasize storage/maintaining of information; the complex WM tasks require both storage and processing/manipulation of information (e.g. Baddeley, 1986; Baddeley & Hitch, 1974; Cowan , 2010; Logie, 2011). In essence, the distinction of simple and complex working memory tasks is on a basis of the separate capacity hypothesis and in accord with Baddeley and Hitch (2000) Multi-component Model of working memory (see Section 4.2.4). Simple working memory/short-term memory tasks link with the two distinct slave systems for separate storage of verbal (phonological) and visuospatial information (Allen , Havelka, Falcon, Evans, & Darling, 2015; Davis , Rane, & Hiscock, 2013; Logie, Zucco, & Baddeley, 1990). Complex WM tasks link with central executive for processing and control information (Baddeley, 2012; Cowan, 2008; Gathercole, Durling, Evans, Jeffcock, & Stone, 2008; Gathercole et al., 2006).

Methodologically, the memory span tasks are divided into verbal (e.g. reading span tasks, forward and backward digit span tasks) and non-verbal (e.g. operation span task and counting span task) (Conway et al., 2005). One of the most influential verbal-based measurement is the *reading span task*. This was initially carried out by Daneman and Carpenter (1980), with a focus on the process involved in reading comprehension. They devised a series of working memory tasks: requiring participants to read aloud a series of short sentences at the same time retaining the last word from each sentence for subsequent recall. Participants usually start with two sentences and this increases to a point that they are no longer able to recall all the last words. This typical point is the subjects' *working memory span*. In the version of Daneman and Carpenter's reading span task (Conway et al., 2005), the sentences are presented in groups that range in size from two to six sentences. For example, a participant starts with two sentences and might hear:

The sailor sold the parrot.

The teacher opened the book.

Then if the subject is to successfully recall two words "parrot" and "book", the test will continue for the subject to attempt three sentences and recall the three last words. If the participant fails to recall all three words, the experiment will terminate, and the subject's reading span will be two. Based on this simple version of reading span test, Daneman and Carpenter (1980) added a dual-task version by adding a true-false component to the task. The participants not only had to recall the last word but also identify the truth or falsity of each sentence. The sentences were chosen from general knowledge quiz books and covered various domains. Conway et al. (2005) suggested that adding the second task could prevent the participants from merely memorizing the last word without devoting attention to reading the sentences. Turner and Engle (1989) later developed a version that changed the additional task identifying true-false into identifying syntactically the correctness of a sentence. Therefore, the central idea of the reading span test of Daneman and Carpenter (1980) is to jointly tap the storage and processing functions of working memory, which is also known as *complex working memory*. This approach has been widely used to measure WMC and predict its relationship with reading comprehension in empirical research (discussed in next section).

Apart from reading span task, non-verbal memory span tasks have also frequently been used to measure WMC. The *operation span task* replaced reading sentences in Daneman and Carpenter's version, and required subjects to solve mathematical operations while attempting to remember words. The *counting span task* (Case et al, 1982) involves counting shapes and remembering the count totals for subsequent recall. In the version of Case et al.'s counting span task, participants orally counted and pointed their fingers at green dots in the face of disruption by yellow dots. On the surface, these three span task seems apparently different, but in essence, they are complex span tasks sharing the general capacity hypothesis. Therefore, these span tasks are primarily concerned with central executive component and designed with a dual-task paradigm – to force storage/maintain of information in the face of processing/manipulating.

The advantage of the complex span tasks (e.g. RST) is that they directly tackle the issues related to the central executive, which is perhaps the most crucial component of working memory system. On this ground, one can subsequently work on the practical problems such as reading comprehension or reasoning tasks (Baddeley, 1992). However, the disadvantage lies in over-dependence on the complex working memory tasks which are based on an arbitrary construction, meaning it is less likely to undertake a detailed analysis of the storage and processing component. To this end, simple working memory span tasks can serve as additional method to measure the capacity of different subsystems of (e.g. digit span tasks, Corsi-block span tasks or other adaptions of operational span tasks and counting span task).

In reading research, a variety of span tasks have been applied to measure the WMC and predict its relationship with reading comprehension. Section 4.5 will continue to explore the relationship between the WMC measured by the different span tasks and reading comprehension to identify the most appropriate measure for this study.

4.5 The current study and research questions

This section explains the reasons why this thesis choose the Multi-component Model of working memory (Baddeley and Hitch, 2000) as the theoretical foundation and design three different span tasks as WMC measures. There are three theoretical reasons for choosing the working memory model and two methodological reasons for choosing different span tasks.

4.5.1 Theoretical reasons

The interest of this thesis is to explore reading comprehension performances in different medium and to understand what extent working memory capacity can account for the performances. Baddeley and Hitch's working memory model (2000) is used as the theoretical underpinning for the research design and discussion for the current study. There are three main reasons for choosing the working memory model. *First*, the working memory model has become a major explanation for memory and language processing in recent research and has received wide support from neuropsychological experiments. *Second*, a number of studies presume a strong relationship between working memory capacity and reading comprehension and takes this as an account for differences between onscreen reading and print reading comprehension performances (Singer and Alexander, 2017). However, little research has provided empirical evidences for a relationship existing between working memory capacity and onscreen reading comprehension performance. *Third*, the working memory model describes a comprehensive and clear four-part structure, which can explain many different types of remembering and manipulation (e.g. verbal, visual or spatial) of information. Therefore, the multi-component model of working memory model provides theoretical sophistication but also a methodological template for the span task used in this research.

4.5.2 Methodological reasons

Research concerning working memory and reading comprehension has primarily used verbal-based span task (e.g reading span task) to investigate how working memory influences reading comprehension. However, the current study attempts to use three tasks, namely, Forward Digit Span task (FDS), Corsi-block Span task (Corsi) and Backward Digit Span task (BDS) for two reasons. *First*, three span tasks are designed with an aim to tap at both the structure and function of working memory. The FDS aims to capture the phonological loop slave system, also known as the phonological short term memory; the Corsi task aims to measure the visuospatial sketchpad slave system, also known as the visuospatial short term memory. The BDS task aims to tap into the central executive. The first two span tasks (FDS and Corsi) are to measure simple short term memory which are primarily responsible for storage; the third BDS task is to measure the complex working memory. This combination follows the dual-task paradigm on WMC measurement – forcing storage in the face of processing, but also has an additional goal to look at two sperate slave systems.

Second, this study will use digit span task over reading span task to measure WMC to avoid measuring overlapping variance with the reading comprehension test. Measuring WMC by reading span task in reading research has a potential risk, that is, it is hard to conclude whether any correlation between the WMC and reading comprehension is due to the capacity of working memory or the ability in reading. For this reason, in the present study it is desirable to use non reading-based measurements and provide domain-independent evidences of working memory to reading comprehension across different medium.

Therefore, Baddeley and Hitch (2000) Multiple-Component of working memory is used as a theoretical underpinning to investigate how the WMC relates with onscreen and print reading comprehension performances; a combination of FDS, Corsi and BDF task is used to measure WMC to explore how the various components of the working memory model might be involved in onscreen and print reading comprehension.

4.5.3 Research questions

In order to examine the relationship between working memory capacity and performance outcomes at different levels of reading comprehension from screen and print, the Multi-component Model of working memory helps to shape the research questions on working memory capacity test of this study. These are

1. a. Is there a relationship between *working memory capacity* and *literal* reading comprehension across different text presentation?

   b. Is there a relationship between *working memory capacity* and *inferential* comprehension across different text presentation?

2. a. To what extent does *working memory capacity* predict *literal* reading comprehension across different text presentation?

b. To what extent does *working memory capacity* predict *inferential* reading comprehension across different text presentation?

3. a. To what extent does working memory measured by different tasks (FDS, BDS, and Corsi-block) predict *literal* reading comprehension across different text presentation?

   b. To what extent does working memory measured by different tasks (FDS, BDS, and Corsi-block) predict *inferential* reading comprehension across different text presentation?

NB: It should be noted that in the WMCT part, the research questions will be tested on L1-English and L1-Chinese-speaking participants together rather than separately. There are two reasons: 1) the purpose of the WMCT is to explore the factors accounting for the performances in the first RCA part but not focus on the differences of WMC between L1-English and L1-Chinese-speakers; 2) the number of sample size of separate L1-Enlgish and L1-Chinese (n = 30) was not sufficient to conduct the statistical test.

Pilot study

Chapter 5 Pilot study: Purpose, methodology, results and discussion

5.1 Purposes

To recap, the main purpose of this thesis is to examine the differences of reading comprehension from screen and print. Chapter 1 has explained the current research target at an Anglophone University context and the population of first-year L1-English and L1-Chinese-speaking undergraduate students. The next three chapters have reviewed the literature in the areas of the theories of reading, empirical studies addressing digital reading issues and models of working memory. Review of the literature has revealed two critical methodological issues of conducting this research: 1) the lack of a theoretical framework to design reading comprehension assessment for higher education context; 2) the lack of a valid and reliable reading comprehension test. To solve the first problem, Chapter 2 has proposed a hybrid framework of reading comprehension assessment (Section 2.7), however, a fundamental question within this framework needs to be answered first – What evidence is there for the division between literal and inferential reading comprehension assessment? The pilot study is in an attempt to find some evidences for such division. To solve the second problem, this pilot study attempts to make use of the Barrett's Taxonomy as test specifications to develop reading measures. Therefore, the purpose of the pilot study is to validate the proposed theoretical framework and researcher-designed reading test, ensuring a valid and reliable reading test for the following main study. As the focus of the pilot study is on the validation of the reading test itself, the polit study design is on a basis of reading on print condition only to reduce the possible effects from the text medium.

5.2 Methodology

5.2.1 Participants

Twenty four participants took part in this pilot study. All of them were Chinese students from the business school who were at different levels of study in the University of Leeds. The selection of L1-Chinese students only was due to time and resource constraints. In further study, both L1-Chinese and L1-English students should be considered.

They had all finished high school education in China before they came to the UK. The participants were randomly divided into two groups (Group A and Group B). Each group had twelve students. Group A (answering questions without the text) participants completed reading tasks without the text while group B (answering questions with the text) participants completed reading tasks with the text. Demographical information was collected from a self-report questionnaire. Details for each group are given in Tables 5.1, 5.2 and 5.3.

The rationale for asking participants in Group A to answer the questions without the text was to create a reading condition which was closer to a real academic setting. Reading every piece of authentic reading material, such as a chapter in textbook or a journal article is not like taking a reading comprehension test. More often, students engage in academic reading tasks without the texts after some time of reading. Thus, it is interesting to investigate whether the reading conditions would influence comprehension

Table 5.1 Group A (answer questions without the text) Participants' information

Participant	Age	Level of study at university	Duration of study in the UK	IELTS reading	English reading per week
1	23	1st year UG	4 years	6	8 hours
2	18	1st year UG	1 year	6.5	5 hours
3	18	1st year UG	2 years	5.5	2 hours
4	19	1st year UG	1 year	5.5	3 hours
5	23	1st year UG	1 year	6.5	5 hours

6	22	3rd year UG	3 years	6	3 hours
7	25	1st year UG	1 year	6.5	6 hours
8	20	2nd year UG	2 years	6	5 hours
9	20	2nd year PG	2 years	6	5 hours
10	22	1st year PG	1 year	7	8 hours
11	24	1st year UG	3 years	7	10 hours
12	19	2nd year UG	3 years	6	5 hours

Table 5.2 Group B (answer questions with the text) Participants' information

Participant	Age	Level of study at university	Duration of study in the UK	IELTS reading	English reading per week
1	20	2nd year UG	2 years	6	5 hours
2	22	3rd year UG	1 year	6	6 hours
3	21	2nd year UG	2 years	5.5	2 hours
4	18	1st year UG	1 year	5.5	2 hours
5	19	1st year UG	1 year	6	3 hours
6	18	1st year UG	2 years	6	3 hours
7	19	2nd year UG	3 years	6	2 hours
8	20	2nd year UG	1 year	6	2 hours
9	24	1st year PG	4 years	7	8 hours
10	23	1st year PG	3 years	6.5	5 hours
11	19	1st year UG	1 year	6	5 hours
12	19	1st year UG	1 year	5.5	2 hours

Table 5.3 Comparison of participants' backgrounds in Group A and Group B

Participant	Age	Level of study at university	IELTS reading	English reading per week
Group A	21 years and 1 month	2 years	6.22	5.42
Mean (SD)	2.39	-1.04	-0.5	-2.31
Group B	20 years and 2 months	1 year and 8 months	6	3.75
Mean (SD)	-1.95	-1.02	-0.43	-2.01

As shown in Table 5.3, the participants from both groups were very similar in age, years of study, English reading proficiency (as indicated by the IELTS reading scores) and hours of English reading per week. Participants were young adults between 18 and 24 years old and the average age for each group was between 20 and 22 years. More than half of the participants were at the beginning of their studies at undergraduate (UG) or postgraduate (PG) level. The time studying in an English-speaking environment was less than five years; on average close to two years for each group. All the participants had reached the English minimum requirements in reading (most English-speaking universities accept IELTS score in reading of 5.5 to 6). They spent on average between 3 and 6 hours a week on reading English for academic purposes.

5.2.2 Materials and test design

Based on the theoretical framework discussed in the previous section, the test materials consisted of two parts. Part one was a reading text around 3000 words at length; part two consisted of two related reading tasks. The text was part of a chapter *The Marketing Environment* from a textbook *Principles and Practice of Marketing* aimed at first-year undergraduate students in business. Details are given in the Table 5.4. Task one consisted of sixteen open-ended questions; task two was a summarization task for selected paragraphs of a text. Both reading tasks were designed by the researcher. Then a tutor in the business school confirmed that the questions were similar to those that students would be expected to be able to answer during an *Introduction to Marketing* course.

After the reading test a short questionnaire was administered to obtain demographic information such as participants' age, level of study, year of study in the UK, English reading proficiency, and hours of English reading per week. The information was presented in Tables 5.2 and 5.3.

Table 5.4 Source and design of the reading test

Title: *The Marketing Environment*
Source*: Principles and Practice of Marketing, Third Edition by Jim Blythe*
Words count: About 3000 words
Task one: Sixteen open-ended comprehension questions
Task two: Summary consisting of no more than 150 words

5.2.3 Scoring scale

Task one (open-ended questions) and task two (summary writing) were assigned equal total scores. The first task, the open-ended question, had a total of 16 test items, and each item was assigned 2 points. The scale employed an unweighted partial credit system (Yu, 2005): 2-point responses included all correct information necessary for a complete response; 1-point responses included part of the correct information necessary for a complete response; 0-point responses included no correct information.

The scoring scheme for summary writing was based on content-related criteria. It was assigned 32 points the same as task one. To determine the summary criteria, two PhD students from the Business School and one tutor who taught the course *Introduction to Marketing* were asked to produce a summary on the same paragraphs. They all agreed on the eight most frequently occurring points. Each of these key elements was assigned 4 points. A partial credit system also applied for each element: 4-points for a fully adequate statement; 2-points for a part adequate statement (as long as a keyword was mentioned in the answer); 0-points for no inclusion of the key words. The total maximum score for the comprehension test was 64.

5.2.4 Data collection procedures

The reading test was administered within a 90-minute session. During the session, participants were required to read the text within 50 minutes and then complete two reading tasks within 40 minutes (20 minutes for each task). The test took place in one classroom in the School of Education on different days within two weeks. For each day, a group of two or three students participated in the study. They followed the procedures as shown in Table 5.5.

Table 5.5 Reading test stages and time allocation

Stage	What to do	Time (mins)
1. Pre-reading	read texts (and/or questions)	50
2. While-reading	read texts (and/or questions)	
3. Post-reading	task 1: 16 open-ended questions	20
	task 2: summary writing	20

To help guarantee the procedures' consistency and reliability, the test was administered by the researcher. At the pre-reading stage both groups were presented with the reading text and questions (see Table 5.6). However, at the post-reading stage, the texts were collected from Group A participants after the first 50 minutes. In other words, Group A had to complete the open-ended questions without referring back to the texts. Conversely, Groups B completed the questions with the text.

Table 5.6 Reading test procedures

	1. Pre-reading	2. While-reading	3. Post-reading
Group A	text and questions	text and questions	questions only
Group B	text and questions	text and questions	text and questions

Two research questions were asked:

1. Do scores on literal and inferential questions differ according to the reading condition (with or without text)?
2. Do scores on open-ended questions and a summarization task differ according to reading condition (with or without text)?

5.2.5 Statistical analysis

The IBM SPSS (Statistical Package for Social Science) program 26 was used to analyse data.

5.2.6 What test to choose?

To determine what kind of test should be used, the research questions were revisited. With regard to the first question, ‘Do scores on literal and inferential questions differ according to the reading condition (with or without text)?’ This involves two types of test outcome – scores on the literal questions and scores on the inferential questions. In a situation in which there is more than one dependent variable/outcome, a MANOVA (multivariate analysis of variance) should be used to test for differences. It can be used to identify interactions between independent variables/factors and to conduct planned contrasts to see which groups differ from each other (Field, 2013, p. 624). A MANOVA test can also be used to answer the second research question, ‘Do scores on open-ended questions and a summarization task differ according to reading condition?’, since there are two further variables/outcomes (scores on the open-ended questions and a summarization task) to be examined.

5.2.7 Rationale for MANOVA

5.2.7.1 Why not a t-test or an ANOVA?

The situations where t-test, ANOVA and MANOVA can be used are different. First, both the t-test and the ANOVA are designed to test for statistical differences on a *dependent variable* among two or more *independent variables*. Field (2013) suggests that a MANOVA can be thought of as an ANOVA for situations where there are several dependent variables. Since the present research aimed to investigate the potential differences among several dependent variables, the MANOVA is the appropriate test.

5.2.7.2 Why not separate t-tests or an ANOVA?

There are two reasons why separate tests may not be conducted. By conducting multiple tests on the same data, it tends to ignore the relationship between variables but also increases the chance of obtaining Type 1 error. That is, the results might be due to the accumulation of error when multiple tests are carried out. In addition, the relationship between dependent outcomes is ignored when conducting separate t-tests or an ANOVA. To take the example of the first research question, the correlation between two dependent outcomes (scores on literal questions and inferential questions) would not be observed. The same issue is applicable when investigating the second research question. As summarized by Huberty and Morris (1989), the MANOVA has greater power than the ANOVA to detect effects because it takes account of correlations between dependent variables.

5.2.8 Checking assumptions before conducting MANOVA

Several assumptions needed to be fulfilled before conducting a MANOVA test.

a. Independence. There was no relationship between the observations in each group as different participants were assigned to the two groups (with text and without text).
b. Random sampling. The data were randomly sampled as the participants were randomly selected from the population.
c. Multivariate normality. Multivariate normality cannot be checked directly at one time. Thus, the normality of each dependent variable was checked numerically by using the *Shapiro-Wilk test* (for use when the sample size is less than 50, which was the case here) and graphically by using a *Normal Q-Q Plot*. The results of the *Shapiro-Wilk test* suggested the data for each dependent variable were normally distributed: open-ended question ($p = 0.09$); summarization ($p = 0.06$); literal questions ($p = 0.11$); and inferential questions ($p = 0.28$).

In addition, the normal Q-Q plot graphically illustrated the normal distribution of the data. Figure 5.1 to Figure 5.8 show that the data were normally distributed as they follow a near-linear pattern.

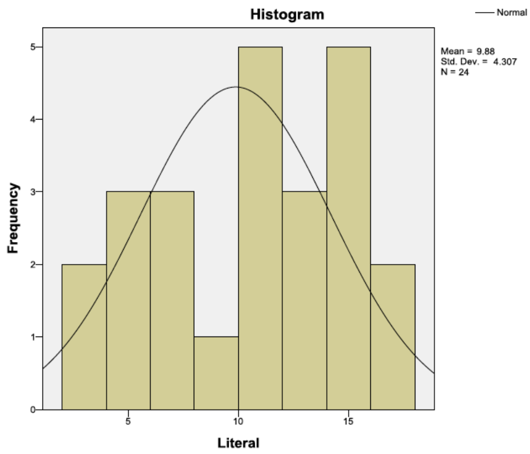


Figure 5.1 Histogram of Literal Questions

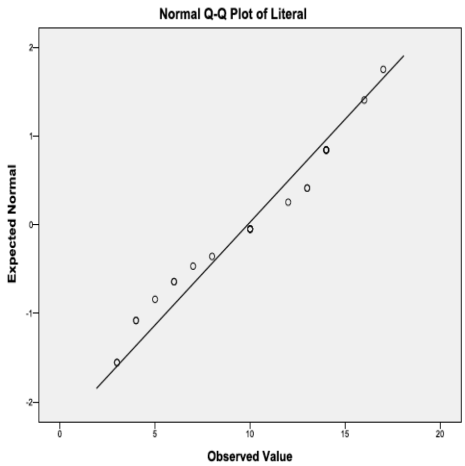


Figure 5.2 Q-Q Plot of Literal Question

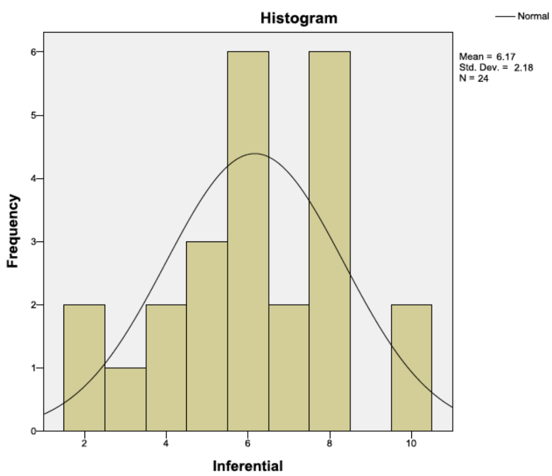


Figure 5.3 Histogram of Inferential Questions

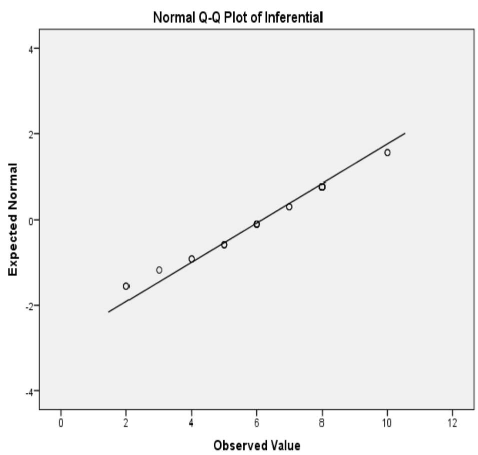


Figure 5.4 Q-Q Plot of Inferential Questions

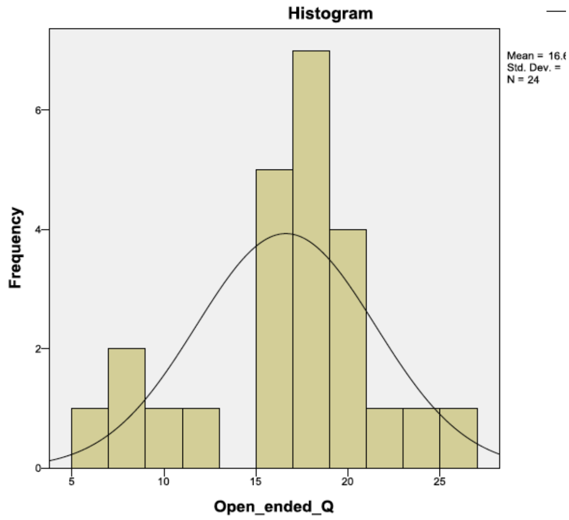


Figure 5.5 Histogram of Open-ended questions

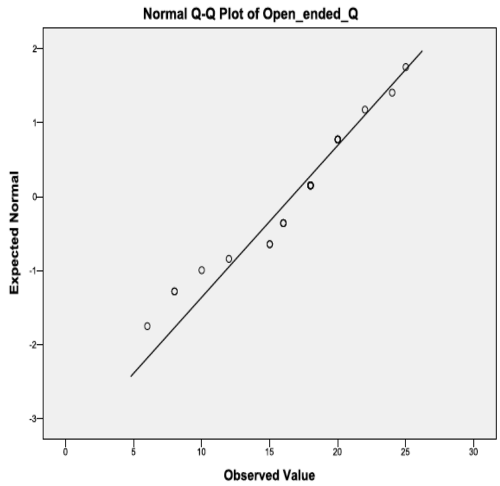


Figure 5.6 Q-Q Plot of Open-ended questions

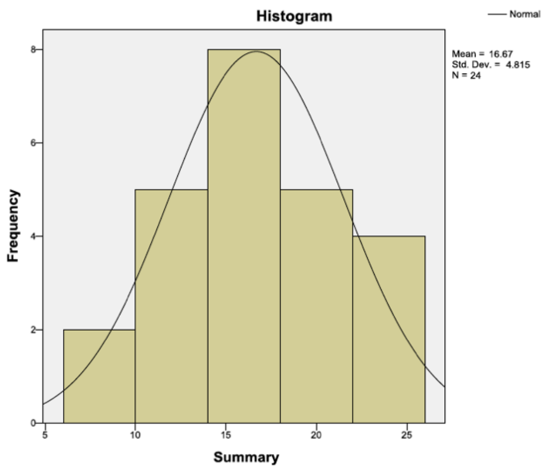


Figure 5.7 Histogram of Summarization task

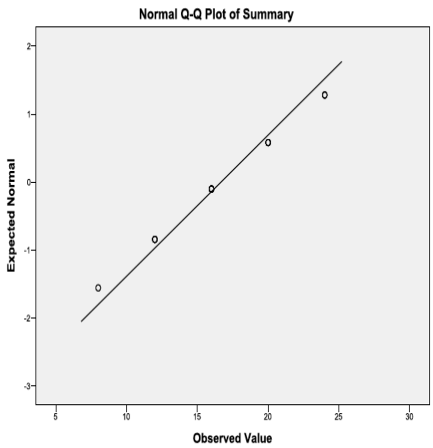


Figure 5.8 Q-Q Plot of Summarization task

d. **Homogeneity of covariance matrices.** A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between Groups A and B were equal for the summarization task scores (p=0.48); scores on the literal questions (p=0.57); and scores on the inferential questions (p=0.83) (see Table 2.12). However, the results for scores on the open-ended questions (p = 0.01) suggested that the variances between the two groups were not equal. Therefore, the data from the open-ended questions did not fulfil the assumption of variance required to conduct a MANOVA test. Thus, the MANOVA was not an appropriate test to use for answering the second research question. Instead, independent t- tests were used to answer the second research question. This means that scores on the open-ended questions and scores on the summarization task will be compared separately.

5.3 Results

5.3.1 Research question 1: Scores on the literal and inferential questions

The maximum score possible on the open-ended questions was 32 (20 from the literal questions and 12 from the inferential questions). With regard to the literal questions, the mean scores for students without the texts and with the text were 7.75 (SD 4.00) and 12 (SD 3.59) respectively. With regards to the inferential questions, the mean score for students without the texts and with the text were 2.16 (SD 5.35) and 2.07 (SD 4.00) respectively.

Table 5.7 Scores of different types of questions (literal and inferential) according to reading condition

Question type	Group	Mean	Std. Deviation
Literal	A. Without the texts	7.75	4.00
	B. With the texts	12.00	3.59

	Total	9.88	4.30
Inferential	A. Without the texts	6.83	2.16
	B. With the texts	5.50	2.07
	Total	6.17	2.18

The results of a MANOVA Pillai's trace test showed a significant effect of reading condition on the scores for the literal and inferential questions, $V = 0.54$, $F = (2, 21) = 12.36$, $p = 0.00$. Furthermore, separate univariate ANOVAs on the outcome variables revealed that there was a significant difference between group A and B on the scores for the literal questions, $F\ (1, 22) = 7.5$, $p = 0.01$, but no significant effect of reading condition on the scores for the inferential questions, $F\ (1, 22) = 2.4$, $p = 0.13$. Thus, to answer the first research question, scores on the literal questions differed according to reading condition. The group of students reading with the texts obtained higher mean scores (12.00) than those without the texts (7.75).

5.3.2 Research question 2: Scores on the open-ended questions and summarization task

The maximum score on each test was 32. With regard to the open-ended questions, the mean score for students without the texts and with the text was 14.58 (SD 5.73) and 18.67 (SD 2.77) respectively. With regard to the summarization task, the mean score for students without the texts and with the text was 15.33 (SD 5.35) and 18.00 (SD 4.00) respectively (see Table 5.8).

Table 5.8 Mean scores (and standard deviations) on the open-ended questions and summarization task

Test format	Group	Mean	Std. Deviation
Open-ended Q	A. Without the texts	14.58	5.73
	B. With the texts	18.67	2.77
	Total	16.63	4.87

Summary	A. Without the texts	15.33	5.35
	B. With the texts	18	4
	Total	16.67	4.82

As discussed earlier, a MANOVA test was not applicable to answer the second research question because the assumption of homogeneity of variance was violated for the scores on open-ended questions. Thus, a Mann-Whitney U Test (which is an alternative t-test to test for differences in nonparametric data), and an Independent t-test were used to examine the effect of reading condition on scores on the open-ended questions and summarization task respectively.

To answer the second research question, a Mann-Whitney U Test was conducted and showed that scores for Group A participants (who answered the questions without the text) (Mdn=15.5) were not significantly different from those for Group B (who answered questions with the text), (Mdn=18.00), U=101.50, z=1.73, p=0.09, r=0.35. In other words, reading condition had no effect on the scores of open-ended questions. Second, an Independent t-test was conducted to compare scores on the summarization task for the two reading conditions (Group A and Group B).There was no significant difference in scores for the without-text group (Group A) (M=15.33, SE=1.54) and the with-text group (Group B) (M=18.00, SE=1.15), $t(22)$= -1.38, p= 0.18. These results suggest that reading condition (without or with text) had no significant effect on the scores for the summarization task either.

5.3.3 Reliability and Validity

5.3.3.1 Production of the reading test

Drawing on the literature in test design, the production of this Reading Comprehension Assessment followed the framework of Bachman and Palmer's (2010) design in Assessment Development Use (Figure 5.9). Overall, the text production involves five important stages: 1.Initial planning; 2.Design; 3.Operationallization; 4. Trailing; 5. Assessment use.

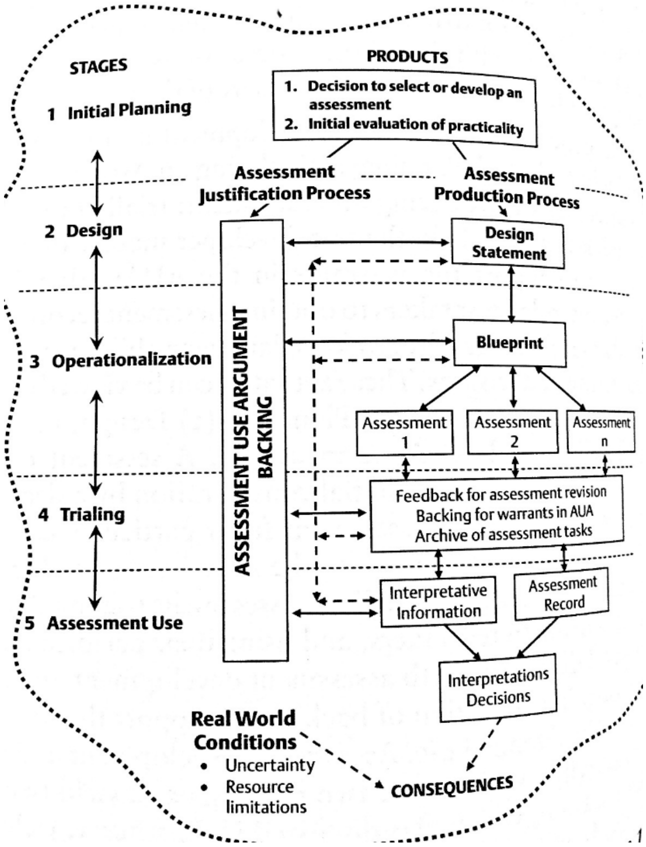


Figure 5.9 Assessment development and use, Stages 1-5 **(**Bachman and Palmer, 2010)

According to the table, the procedures of the tests design were as follows.

a. Establish purpose of the test. The main purpose of the reading comprehension test was to produce a practical and reliable research instrument for the main study. This explains why the test only used two test formats (open-ended questions and summarization).
b. Draw up test specifications and scoring scale.
c. Write the test. A reading test was constructed which reflected current reading theories and academic reading characteristics.
d. Moderate/trial the test. Before participants took the pilot test, it was trialled by two PhD students and a tutor studying/working in the field of Business Studies. There was agreement among all three regarding the questions and answers.
e. Administer the test. For some practical reasons (the pilot test was held during the busy exam time at the end of semester), the participants could not take part in the test at one time simultaneously. But the researcher

administered the test personally and guaranteed the same test procedures were followed each time.

f. Mark and report results. The tests were marked individually according to the agreed answers by the tutors in this module and PhD students in the Business School.
g. Evaluate the test. Participants' test scores and the answers to each question were analysed and this analysis informed the choice of questions included in the final test version.

5.3.3.2 Reliability

Reliability means a measure should consistently reflect the constructs that it is measuring (Field, 2013). To assess whether the pilot reading test demonstrated internal consistency, a Cronbach's alpha test was used to check its reliability. The results showed that the sixteen open-ended questions and the summarization task had acceptable levels of reliability, as Cronbach's alpha was 0.78. When deleting the summarization task, the reliability can be increased to 0.82. However, Cronbach's alpha provides an overall reliability coefficient for a set of variables (e.g. questions). The pilot reading test covered different underlying dimensions (literal questions and inferential questions). Cronbach's alpha will not be able to distinguish between these. In order to check their reliability, further statistical tests, for example, a factor analysis or a principal components analysis (PCA) should be conducted. Removing the summarization task to increase the overall reliability is also taken into consideration. Thus, one of the research topics for the main study will be an investigation into the divisibility of reading constructs (also corresponding to the literature in reading subskills).

5.3.3.3 Validity

In order to improve the validity of the reading test, a tutor in the Business School confirmed that the test questions were similar to those students would be expected to be able to answer when reading the textbook chapter for the *Introduction to Marketing* course. The answers to the open-ended questions and the summarization task were produced by two PhD students and the tutor. Specifically, the validators compared their answers, discussed any disputes, and finally reached an agreement. But before the main study a further five tutors and five PhD students from the Business School will double-check these comprehension questions and answers. Since the research results suggested that the with-text-group of participants (Group B) achieved higher test scores than the without-text-group of participants (Group A) when answering the literal questions, this might indicate validity of the literal comprehension questions (they measured what they were supposed to measure). However, the validity of the inferential questions may need further statistical testing in the main study. To evaluate the construct validity, a factor analysis will be conducted to assess whether the questions test what they are supposed to test. By comparing the constructs emerging from the answers and the constructs in the theories, the validity of the reading tests can be assessed.

According to Bachman and Palmer's (2010) test design framework, revisions need to be made and more questions added based on the pilot test results and feedback. Analysis of the pilot results indicated the necessity of adding more questions to the original test. The initial 16 open-ended questions will be extended to 42 questions (with an equal number of literal-meaning questions and inferential-meaning questions) in order to facilitate further factor analysis.

To summarize, it has to be acknowledged that a reading test cannot represent all of the constructs in real academic reading comprehension. However, the present measure has been designed to best capture academic reading constructs and incorporate these constructs into a reading comprehension framework. This is in order to produce a valid and reliable research instrument to further explore the issue regarding screen reading comprehension. Previous research lacks agreement on the constructs of academic reading comprehension, meaning that various reading tests have been used in previous research. This might be one reason for the conflicting findings regarding screen reading comprehension. The following sections will discuss the objectives, methods of the main study and outlines the research plan.

5.4 Discussion

The results of the pilot study have shown that reading condition had no effect on the scores for the open-ended questions, the summarization task, or for the inferential questions. However, it did have an effect on scores for the literal questions. There are some possible explanations for these findings.

Firstly, the lack of significant differences between the scores on open-ended questions and the summarization task, may rely on the readers' comprehension process. During the while-reading stage, participants were presented with both the text and the questions. Before answering the questions, readers might have constructed their understanding of from the text and their own schema. In spite of the absence of the text, both groups of participants relied upon their own understanding to produce the answers. Another possible explanation relates to the test questions. Even though some answers were explicitly stated in the test, participants were expected to employ appropriate reading skill to reach the answers. But they tended to obtain the answers by drawing on their subject-related knowledge.

Third, better scores on literal questions may be because the groups with texts in front of them could obtain answers for literal questions by directly referring back to the texts; conversely the groups without the text may have failed to recall the textual information due to limited short-term memory capacity when the texts were unavailable. Furthermore, the results provide evidence for the validity of literal-question items as they measure what they were supposed to measure (Alderson, Clapham, & Wall, 1995). To this end, it is sensible that with-texts-students answered reading questions had better results than those without texts.

Finally, the fact that there was no significant effect of reading condition for inferential questions suggests that the act of referring to the text did not significantly influence participants' scores when they answered inferential questions. There are at least two ways in which this finding can be explained. The first is that the answers for the inferential questions are not easily obtained by directly assessing texts. That is, some inferential questions cannot be answered via direct retrieval of the information in texts. Another possible reason relates to the schema theory. That is, when answering the inferential questions, it is possible that both groups drew upon their personal experience and prior knowledge on this subject. It can be assumed that participants' content schema were similar in that they had the same first language, came from the same discipline but also reached the threshold level of English proficiency to study in a English-medium university (as indicated by IELTS reading score). When answering inferential questions in which answers were not explicitly stated in the text, retrieval of the text was not a crucial factor in order to produce the answer. Thus, there was no significant effect of reading condition for the inferential questions.

5.5 Conclusion and adjustments

The testing instrument within the pilot study faced some difficult issues. First, the taxonomy and list of subskills in earlier work is designed for elementary or high school students and teachers at that time. They are not meant to aid university level students which are the target language users in this research. Second, Barrett's (1968) frameworks were originally designed for young children who learn English as their mother tongue rather than adults who learn English as second or foreign language. It is unknown whether the reading process of native-English-speaking young children is similar to adults using English as a second/foreign language. Considering students' individual age, educational level, culture difference and English language proficiency, some skills still need to be further explored, in particular those needed by English as a second language students in higher education. Despite the uncertainty, issues about whether these subskills are appropriate in academic reading needs further research in the main study.

Returning the aims of the pilot study, the first aim regarding the division between literal and inferential assessment has been achieved but the second concerning the reliability and validity of a self-designed test still needs to be fulfilled. First, the significant difference between literal and inferential comprehension questions on both conditions suggests there exists some evidence for the division between literal and inferential comprehension assessment for students in higher education context. This finding, to some extent, supports the hybrid framework of reading comprehension assessment proposed in section 2.7, indicating the rationality and feasibility of comparing reading comprehension performances from two dimensions – literal and inferential.

However, regarding the second aim – validating the self-designed reading test in terms of validity and reliability, has not been fully achieved. The pilot study checked the overall reliability of all the test items as acceptable (Cronbach's alpha 0.78). Yet Cronbach's alpha did not distinguish the reliability between two dimensions of literal and inferential. In addition, the pilot study did not statistically check the validity of the test because the sample size (24) and the number of test items (16) were not sufficient to conduct a statistical test. Continuing to fulfil this aim, the main study needs to further validate the reading test by means of statistical tests, using a factor analysis or a principal components analysis (PCA) technique. Thus, the main study has two adaptations on the basis of the pilot reading test:

a. To increase the number of test items by linking each open-ended question to each subskill, rather than literal and inferential dimensions;

b. To remove the summarization task to avoid measuring confounding subskills.

Main study: Methodology and Results

Chapter 6 Part One: Reading Comprehension Assessment

6.1 Introduction

So far, Chapter 2 has proposed a theoretical framework for designing reading assessment by reviewing theories of reading; Chapter 3 has examined previous empirical studies and raised methodological issues relating to design an experimental reading test. Chapter 4 turns to the theories and models of working memory in attempt to find relevant literature from cognitive psychology to understand text processing. The pilot study in Chapter 5 tested a self-designed reading assessment which was developed from the proposed framework in Chapter 2 and summarized two possible adaptions. Then the following chapters 6 to 9 move to the main study of this thesis, describing the way how the main study is designed, arranged and implemented. The main study includes three empirical parts, each of which being regarded as independent but inherently related. Chapter 6 and 7 are both on the first part of the main study – *Reading Comprehension Assessment (RCA).* Chapter 8 and 9 will focus on the second – *Working Memory Capacity Test (WMCT)* and third part – *a follow-up interview* respectively.

Chapter 6 describes the research methodology of the RCA and Chapter 7 details the findings. This chapter starts with the research aims and overall research design (6.2). This is followed by a description of the philosophical underpinning (6.3) with an explanation of how the current study design aligns with the philosophical stances. Next section (6.4) discusses the research approach employed in the RCA, including context, recruitment process and participants, design, procedures and materials. It follows a discussion of the hybrid theoretical framework adapted from the pilot study (6.5). Guided by the adapted framework, reading subscales and test items are developed correspondingly into an experimental reading test (6.6). The penultimate section reports on the data analysis (6.7). Section 6.7.1 details the process of validating the experimental test in terms of validity and reliability; Section 6.7.2 details the processes of selecting and correcting statistical tests to answer the research questions.

6.2 Research aims and overall design

As discussed in Chapter 1, the overall aim of this thesis is to examine the differences in reading comprehension from screen and print but also to explore the underlying factors that are contributing to the differences. The main study attempts to achieve the aim by adopting a mixed-method approach and conducting three empirical studies. First, the RCA aims to examine whether possible differences exist in reading comprehension performances when reading from screen and print through an experimental reading test which takes into account two dimensions of reading, namely, literal and inferential. The second part – WMCT continues to investigate the factors from cognitive psychology and examine the relationship between working memory capacity (WMC) and comprehension performances from screen and print. Finally, the follow-up interview is to triangulate the data collected from the RCA and WMCT. This not only provides qualitative data to compare text processing differences between screen and print but also gives access to the participants' different perceptions and their contexts that help to shape the perceptions (e.g. cultural and educational backgrounds, leaning experiences and habits etc.).

Using this mixed-method approach rather than mono method, as indicated by Tashakkori and Teddlie (2003), has three areas of advantages. First, mixed-method can answer simultaneously confirmatory and exploratory research questions. Conducting experimental tests (e.g. RCA) can directly answer whether reading comprehension performances from screen and print exist any differences (confirmatory); having an in-depth interview helps to explore what factors contribute to the possible differences (exploratory). Second, mixed-method provides stronger inferences through depth and breadth to unfold the complexity of research issue. Interpreting the experimental reading results needs to delve into the participants' cognitive process (depth) and understand participants' individual differences (breath) such as learning habits, language proficiency etc. Third, mixed-method gives opportunity through divergent findings for an expression of different viewpoints. This allows a logic of triangulation of combining quantitative test results and qualitative interview data to understand different aspects of the issue of digital reading.

The next section moves to the philosophical assumptions of this research, which further explains the study design.

6.3 Philosophical underpinning

In social science research, *post-postpositivist* and *constructivist* research are anchored at opposite ends of a paradigm continuum (Betzner, 2008). This is because *post positivism* is associated with quantitative methods and highly formal rhetoric but *constructivism* is associated with qualitative methods and informal rhetoric. Given such a situation, the *pragmatist* stands between two ends and strives to bridge the gap between the scientific method and structuralist orientation and naturalistic method and freewheeling orientation (Creswell and Clark, 2011). Then, for the current thesis, the understandings of the world nature and the functions of being a researcher draws on both *post-positivist* and *pragmatist* paradigms (Creswell , 2013; Creswell and Clark, 2011). This is embedded in four elements: axiology, ontology, epistemology, methodology (Creswell and Poth, 2018).

The axiological assumptions accepts the “value-free” stance of conducting the research (Crowther and Lancaster, 2008) but also recognises the “value-laden” nature of interpreting results (Creswell and Poth, 2018). This thesis acknowledges a detached, neutral and independent relation between the researcher and what is researched. Independent means the research can be objective and the researcher can maintain minimal interaction with the research participants when carrying out the research (Crowther and Lancaster, 2008). In this thesis, two experimental tests (RCA and WMCT) were adopted to measure the participant’s’ reading comprehension performances and working memory capacity to examine their relationship by a series of statistical tests. This follows a hypotheses-and-deductions research process. On the other hand, this thesis understands that “complete objectivity is impossible’ (Gay and Airasian, 2000, p.205) in social science research. For example, the measurements adopted in the current study were designed by this thesis herself; thus is impossible to conduct a purely objective research. When interpreting the results, it is impossible to escape this thesis’s personal perceptions and engagement with the issue that is being researched. The suggestion is to reduce the bias by ensuring the validity of the measurements and adopting a non-judgemental stance.

The *Ontological* assumptions, or "assumptions of the nature of reality", of this thesis accept the notion that an objective reality exists apart from human experience and thus social research should work in an evidence-based culture (Alston and Bowles 2013; Engel and Schutt 2014). However, this thesis also acknowledges that the 'reality' is grounded in the environment (Goles and Hirschheim, 2000) and can only be encountered through human experience (Tashakkori and Teddlie 2008). Regardless the endless philosophical arguments, this thesis adopts a *pragmatist* approach*,* linking the choice of approach directly to the purpose of and the nature of the research questions posed (Creswell, 2003). The experimental results from the participants provides evidence for the 'truth'; still, the meaning of the 'truth' is inseparable from the participants' past experience and habits (Dillon et al, 2000) and is dependent upon the their "historical and cultural norms' (Creswell and Poth, 2018, p.24). The *ontological* assumptions centres on evidences and effectiveness; but the evidence is described as socially constructed (Plath, 2006) and the effectiveness possibly depends variedly upon the context (Morgan, 2014).

Epistemology, or "assumptions about how we know the world, how we gain knowledge", (Kaushik & Walsh, 2019, p. 2) rejects the traditional philosophical dualism of objectivity and subjectivity, (Biesta, 2010) by advocating solving practical problems in real world (Creswell and Clark, 2011). Rather than describing this thesis as postpositivist or constructivist, this thesis abandons dichotomies and focuses on two different approaches to achieve the research purposes (Morgan, 2014).

Methodologically, this thesis embraces *pragmatism*, focusing on "solving practical problems in the real world and helping researchers to achieve their purposes" (Kaushik & Walsh, 2019, p.3). It accepts multiple methods or mix-method approach to gain knowledge about the world (Johnson and Onwuegbuzie, 2004). Unlike *postpositivism* – which typically supports quantitative approaches and deductive reasoning, or *constructivism* – which typically endorses qualitative approaches and inductive reasoning, the researcher, in favour of *pragmatist,* stands between two opposite ends and adopts a flexible and reflexive approach to research design (Feilzer 2010;

Morgan 2007; Pansiri 2005). Thus, this thesis adopts an experimental test (quantitative) and an semi-structured interview (qualitative) to collect objective and subjective data from participants. In addition, this data analysis follows an abductive reasoning cycle, continuing moving back and forth between deduction from experimental results and induction from interview data. In this way, the researcher get active involvement in creating data and establishing theories (Goldkuhl 2012; Morgan 2007).

6.4 Methodology of Reading Comprehension Assessment

6.4.1 Context

As previously explained, the purpose of the first part RCA was twofold. First, the RCA continued the exploration of issues about academic reading constructs at higher education that the pilot study failed to solve. Second, the RCA aimed to produce a valid and reliable research instrument to assess reading comprehension for future research in higher education. Therefore, the RCA aims to collect data in a setting where reading from print or screen, is familiar and authentic to the participants in their academic study. This is similar to Ortlieb et al.'s intervention study (2014) aiming to investigating the role of medium in a more naturalistically way. There was no time limit in reading the text or answering the comprehension questions. Participants could also choose their preferred medium to read and were allowed to use any forms of dictionaries.

6.4.2 Recruitment process and participants

The decision to focus on first-year undergraduate students enrolled in a business course was based on two considerations. First, a recent review of the literature about screen and print reading suggested that more than 75% published studies involved undergraduate readers (Singer & Alexander, 2017b). Hence, the outcomes from this research can have a direct comparison with other up-to-date studies reported in the literature. Second, the first year of undergraduate is a crucial transition period for students who come out from school-level study or from non-English-speaking countries. This is because the reading experience in higher education is largely different from school-level reading tasks, as well as from a non-English-speaking educational system. Therefore, employing first-year undergraduate readers would be most suitable for addressing the research goals.

During the recruitment process, a range of strategies were used, starting with the ‘snowball’ recruitment strategy/sampling (Dörnyei, 2007; Seidman, 2013). First, ethically approved advertisements (Appendix A) were posted to relevant groups for Chinese-speaking freshmen through an online social communities (i.e. Wechat) and English-speaking freshman through Facebook between March,2017 to June,2018. During the period, emails were sent to the teaching staff in the language centre and tutors in business schools in Leeds and York and invited participants who studied in Business. Classrooms were visited in person with teachers’ permission and both Chinese and English-speaking students from a module called Introductory Business Course were invited to take part. A tutor who teaches Chinese language course in Sheffield was also contacted and they shared the invitation with more English-speaking students with business backgrounds.

By the end of June of 2018, a total number of 112 (42 females) first-year undergraduates in business schools from three UK universities (Leeds, York and Sheffield) took part in this research. Half (56) were L1-Chinese-speakers while the other half were L1-English-speakers. All of the Chinese participants and thirty of the English participants were recruited in an Introductory Business Course. The remaining 26 business English students were enrolled in a Chinese language course. Participants ranged in age from 18 to 21 and had not been diagnosed with any reading or learning disability.

By the end of March of 2019, 3 of 112 dropped out the following working memory capacity test; 11 more students joined the research. In order to keep the consistency with sample size in the RCA and WMCT, the data were collected from final 120 participates, including 60 L1-English-speaking and 60 L1-Chinese-speaking students.

6.4.3 Design

A 2 (type of question) × 2 (type of presentation) between-subject design was used. 120 participants were randomly assigned to two groups (60 in each group) and read the texts on print and on laptop screen respectively and completed forty-two open-ended comprehension questions. Each group consisted of equal number of L1-English-speaking (30) and L1-Chinese-speaking participants (30). The independent variable was the text presentation with two levels (screen and print). The dependent variable was comprehension accuracy, measured by the scores of correct comprehension questions (literal and inferential type of questions).

The reason for using a between-subject design rather than within-subject design was to reduce the impact of one test condition on the performance of another condition, which was called carryover effect (Creswell, 2018). Participants may become exhausted or uninterested after taking part in repeated tests. On the hand other, taking the measurement tests repeatedly might help the participants become more familiarised and skilled, which means they may be able to produce the better results on the later test due to the practice effect. This can skew the results and make it difficult to determine whether any effect is due to different test conditions or a result of practice.

6.4.4 Procedures and participants

Participants were led to a temperature-controlled classroom upon arrival. Each participant completed the task in the same room as individual sessions. Before undertaking the task, participants read and signed a statement of informed consent. Following consent, participants were given detailed instructions about the task. The participants were told that they would be reading one text and answering open-ended comprehension questions relevant to the text. Instructions indicated that they would be reading on paper or on a screen. After reading instructions, participants were given chance to ask any questions before starting.

During the test, participants would be able to refer back to the original text when answering questions. Importantly, participants were instructed that the reading task would not be timed, and they were encouraged to read at their own pace. This was in order to create a similar reading atmosphere to university-level academic reading rather than a school-level test environment.

Navigation of the screen reading software Adobe Acrobat Reader software was discussed with each participant before reading on screen. Participants were informed that they could annotate and highlight on the PDF texts, but they could not search for online resources to answer the questions.

Following the reading task, participants completed a demographic questionnaire, which included questions about age, gender, years/months studying/living in an English-speaking country, time spent on English reading per week, time spent reading on electronic devices per week, English reading proficiency (i.e. IELTS scores).

Table 6.1 Comparison of demographic information of the participants between screen and print

Text medium	Age	Female No.	Duration of study in the UK	IELTS reading	English reading per week
Screen	18 years and 10 month	38 (60)	9 years and 3 months	6.37	10.56 hours
Mean (SD)	2.89		-1.07	-0.56	-2.87
Print	19 years and 2 months	46 (60)	9 years and 5 months	6	11.78 hours
Mean (SD)	-2.54		-1.19	-0.48	-2.45

6.4.5 Materials

The reading material was a 2500-word text on the topic of global marketing, which was the same text used in the pilot study. Two presentation modes of the text were produced: 1. Screen presentation which was the original electronic textbook in PDF presented on a 13' inch laptop screen with high resolution display. The researcher chose the scrollable pdfs for the participants, as opposed to page-style digital text format because university textbooks and journal articles are mostly electrically stored in scrollable style; conversely, page-style pdfs are common with entertainment books. 2. Print presentation in which the text was photocopied on A4 size white paper from the original textbook.

Paper-based answer sheets for the comprehension questions were provided separately for both groups. It is considered that screen-based answer sheets might increase screen readers' cognitive load when they switch tasks on a smallsized laptop screen. participants might be In addition, there was an in-built dictionary software provided for participants on screen; physical dictionaries or translation dictionary Apps on mobile phones were allowed to use for participants from paper.

6.5 The hybrid framework of reading comprehension assessment

The framework (Figure 6.1) used to design the reading test for the main study was consistent with the pilot study. Two types of comprehension questions were employed: literal-meaning questions and inferential-meaning questions. The literal-meaning questions were designed to measure students' reading ability for explicitly stated ideas; the inferential-meaning questions were designed to measure students' ability to read for implied meanings.

However, one key difference between the framework employed in the main study and the pilot study was the reading construct. That is, the pilot study designed sixteen comprehension questions but the factor analysis failed to specify the underlying reading constructs. Yet the framework in the main study incorporated 14 subscales in total and each type of question (literal and inferential) had seven subscales respectively. These subscales underpinned the reading constructs for the present research but also were considered as test specifications for the reading test. The current subscales were revised from the original version of Barrett's Taxonomy under *Literal Comprehension, Reorganization, Inferential Comprehension and Evaluation* level (see Section 2.4.3, Figure 2.2). Some of the subscales were revised or deleted according to the topic and contents of the selected text. For example, the original subscales in Barrett's version "inferring character Traits" was deleted, "judging desirability and acceptability" was revised as "judging strengths and weakness of a position", as the selected text was an academic exposition from a undergraduate level textbook.

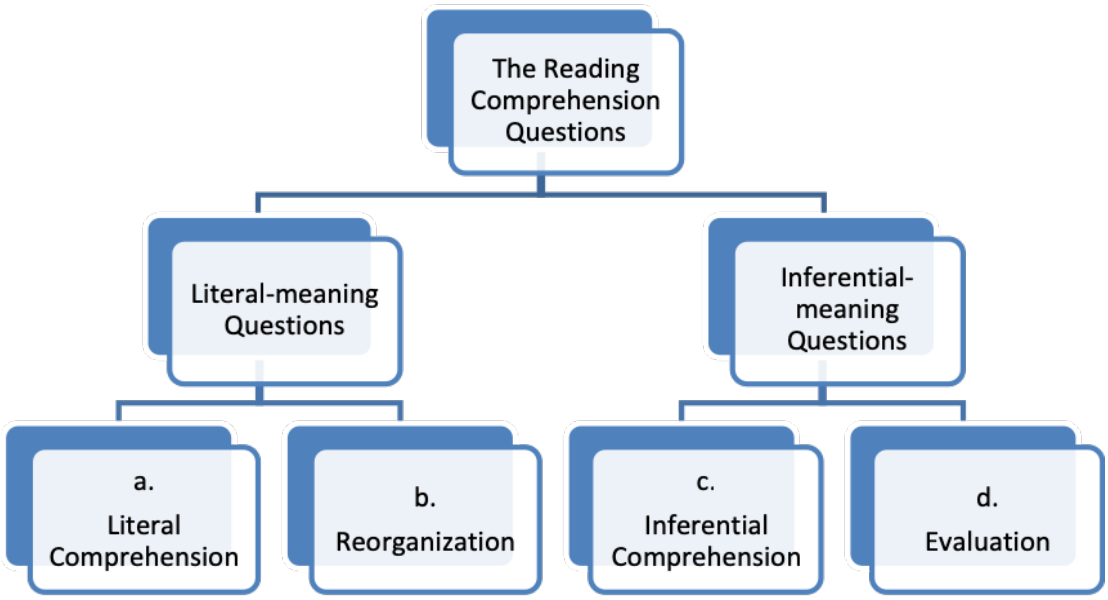


Figure 6.1 The hybrid framework of reading comprehension assessment

6.6 Development of reading subscales and test items

The reading test covered 42 reading comprehension questions. Each subscale represents one subskill, three reading questions were designed on a basis of one subskill. The next sections will explain the subscales in detail and give example questions for each subscale.

6.6.1 Literal comprehension

Literal comprehension requires readers to recognize or recall a single fact or a series of details that are explicitly stated in the text. It contains three main subscales:

1). Recognition or recalling of details;

2). Getting main ideas of a paragraph and larger body of text;

3). Recognition/Recalling comparison or cause and effect.

Example questions:

1) Recognition or recalling of details: The reader is required to locate or identify from memory such facts as the names, the time, the setting or an incident described, when such facts are explicitly stated in the selection

e.g. What factors will affect people making business decisions in an organization?

The answers can be found in the text "competition, customer characteristics, behaviour of supplies and distributors and legislative and social framework."

2) Getting main ideas: The reader is asked to locate or identify or to produce an explicit statement in or from a selection that is the main idea of a paragraph or a larger portion from the selection.

e.g. What is the main idea of the Introduction in this chapter?

The answer can be located and produced in the first sentence “No business operates in a vacuum” or produced as “everything within business relate with each other.”

3) Recognition/Recalling comparison or cause and effect: The reader is required to identify similarities, differences and reasons.

e.g. How does the network between people in an organization bring about better living standards?

The answers are stated in the text, "people are contributing to the society welfare and in return they get satisfactions of their own needs".

6.6.2 Reorganization

The reorganization requires reader to analyse or organize explicitly stated information in the selection. It contains four subscales: classifying, outlining, summarizing and synthesizing.

1) Classifying: placing information into groups.

e.g. In what ways can factors in the marketing environment be classified?

The answers can be found in the text “the environment can be classified as macro and micro but also internal and external factors. The macro factors ….”

2) Outlining: organizing a selection in outline form.

e.g. List some stakeholders that will have an effect on marketing decisions.

The answers can be found in the text "neighbours, suppliers, competitors, customers or governments".

3) Summarizing: paraphrasing or condensing a selection.

e.g. What is the relationship between the firm and its environment?

The answers can be summarized according to the text "the larger firms have greater control over the environment but also they have more difficulties than smaller firms to adapt to environment changes".

4) Synthesizing: consolidating stated information from more than a single source.

e.g. What do governments do to control the economy and stabilize the markets?

The answers need to be found in different places from the text “have a fine balancing act; set interest rates; control taxation and expenditure."

6.6.3 Inferential comprehension

The inferential comprehension requires students to use the literal contents of the text, personal experience and knowledge as basis for conjecture and hypothesis. The information is not explicitly stated in the text. It contains four subscales:

1) Inferring supporting details;

2) Inferring main ideas;

3) Inferring cause and effect relationship;

4) Predicting outcomes.

1) Inferring supporting details: suggesting additional facts that might have made the point more informative and persuasive.

e.g. Can you give example to explain the problem of classifying the macro and micro environment factors?

The answers need to be based on readers’ own experience and find relevant examples to illustrate the problem.

2) Inferring main ideas: producing main ideas when it is not stated explicitly in the text.

e.g. What conclusions can you draw according the description of EU report in 2002?

3) Inferring cause and effect relationship: readers hypothesize author's intention and motivation, the reasons and outcomes.

e.g. What brought about the decrease in the birth rate of Eastern European countries during 1990?

4) Predicting outcomes: on the basis of reading an initial portion of the selection readers predict the consequences of the selection.

e.g. What would be the consequences of depopulation and an aging population?

6.6.4 Evaluation

The evaluation requires students to make judgments about the content of a reading selection by comparing it with external information such as readers' own experience, knowledge or values. It contains three subscales in this selected academic exposition text:

1) Judging facts and opinions;

2) Judging strengths and weakness of a position;

3) Judging adequacy and validity.

1) Judging facts and opinions; distinguishing the reality and assumption, supported and unsupported ideas

e.g. What evidences can you find to illustrate the increase in single-person households in several European countries?

2) Judging strengths and weakness of a position: make decisions of good, bad, right and wrong.

e.g. In your view, how do you think the argument "line between necessities and discretionary purchase is somewhat blurred"?

3) Judging adequacy and validity: judging whether information in a text agrees with other source of information, such as an alternative explanation.

e.g. How do you expect the working population change in the original fifteen member states of EU? Please justify your answers.

Overall, the reading test had 14 subscales and 42 reading comprehension questions. Three questions were designed for each subscale (see Table 6.2).

Table 6.2 Description of subscales and reading comprehension questions

Subscales	Questions
1. Recognition or recalling of details	1. What factors will affect people making business decision in an organization? 2. Please give one or two examples of different types of marketing factors mentioned in the text. 3. How do non-profit organizations benefit from government control?
2. Getting main ideas of a paragraph and larger body of text	1. What is the main idea of the Introduction section? 2. In the "Real-life marketing" case (p. 35), what was the main information used by the companies like Tie Rack and Sock Shop to make their marketing decisions? 3. What is the main idea of the section “cultural environment” (p. 36)?

3. Recognition/Recalling comparison or cause and effect	1.How does the network in an organization bring people better living standards? 2. Why are the boundaries between the internal and external environment sometimes hard to decide? 3. Given that some macro factors are common for all firms, why will they affect firms differently?
4. Classifying	1. How does the author classify the marketing environment? 2. How does the quality of economic change assessment vary in different countries ? 3. What was the main shift to the population in West Europe over the past fifty years?
5. Outlining	1. List some stakeholders that will have an effect on marketing decisions. 2. List the consequences of economic recession mentioned in the text. 3. What changes would the increasing number of single-person households make to the market?
6. Summarizing	1. Summarize the relationship between the firm and its environment. 2. What is a boom-and-bust economic cycle? 3. How do you illustrate the distinction between "necessities and discretionary purchases"?

7. Synthesizing	1. What do governments do to control the economy and stabilize the markets? 2. What were the changes for the fifteen states members of the EU during the thirty years after the mid-1970? 3.What factors will influence the socio-cultural environment?
8. Inferring supporting details	1. How does a firm relate to the environmental factors mentioned in the text? 2. What is the problem of classifying the macro and microenvironment factors? Please give examples. 3. Can you infer what measures the government took to recover the economy in 2008 recession?
9. Inferring main ideas	1. What conclusions can you draw from the description of the EU report in 2002? 2. What information does Figure 2.2 suggest? 3. What would happened to Spain after its government encouraged Latin American Spaniards to return home?
10. Inferring cause and effect	1. In the Costain West Africa case, why did the Costain Company survive despite the financial crisis in 2008? 2. Why is the external environment impossible to control?

	3. What brought about the decrease in the birth rate of Eastern European countries during 1990?
11. Predicting outcomes	1. Can you predict other changes (not mentioned in the text) that increasing number of single-person households would bring to consumer-product companies? 2. Can you predict the results if the EU makes computers or office equipment supplies open to all members? 3. What would be the consequences of depopulation and an aging population?
12. Judging facts and opinions	1. Can you think of any evidence to illustrate the increase in single-person households in several European countries? 2. What changes had already happened and what did the author predict in the future according to EU report in 2002? 3. Do you agree with the authors' expectations on demographic shifts as a result of expansion (p. 34)?
13. Judging strengths and weakness of a position	1. In your view, how far do you think the conclusion that "the line between necessities and discretionary purchase is somewhat blurred"? 2. What do you think of the EU intervention in agriculture markets? 3. What is your opinion about the age demographic changes to the West European market?

14.Judging adequacy and validity	1. Is it true that the government in the 19th century was too poor to control the economy due to the defence of realm? Why? 2. How do you expect the working population to change in the original fifteen member states of EU? 3. Based on your own experience, can you give some other examples to illustrate "think small" rules (p. 35)?

6.7 Data analysis

The IBM SPSS (Statistical Package for Social Sciences 26) program was used to analyse data in the present study. This section first discuss the validity and reliability of the research instrument and then explained what statistical tests were taken for each research question.

6.7.1 Validity and reliability of the reading test

Reading comprehension is a construct that is difficult to directly measure as a single variable. It is abstract and multifaceted. Instead reading subscales and corresponding comprehension questions were developed as part of this thesis, each measuring different aspects of reading comprehension. In order to discover the number of factors influencing variables and to analyse which variables "go together" (DeCoster, 1998), factor analysis was employed. As Yong and Pearce (2013) suggest, factor analysis is useful for studies that involve a battery of tests which can be reduced to a smaller set of underlying subskills (Rummel, 1970). Hence, using factor analysis is a way of reducing the loads of comprehension questions onto meaningful categories of subscales.

The purpose of employing factor analysis was two-fold: 1) to produce a valid and reliable reading measure; 2) to identify a hypothetical framework of reading comprehension. For purpose one, factor analysis and reliability analysis were done for the whole dataset as well as at the subscale level. For the second purpose, possible underlying factors were extracted and analysed on the basis of revised measures.

6.7.1.1 Choice of Factor Analysis approach

Factor analysis is an umbrella term and it has two types (Field, 2000). One is confirmatory factor analysis and the other is exploratory factor analysis. The confirmatory factor analysis aims to test a specific hypothesis while the exploratory factor analysis attempts to explore the data. Principle Component Analysis (PCA) serves the aims of exploratory factor analysis. The purpose of this study is to identify the underlying traits within the reading test. As Bachman (1990, p.260) in the field of language testing explains “in the exploratory mode, we attempt to identify the abilities or traits that influence performance on tests by examining the correlations among a set of measures”.

6.7.1.2 Checking requirements for factor analysis

To perform a factor analysis, two requirements should be met: sampling adequacy and patterned relationship among variables (Field, 2013). To measure the sampling adequacy Kaiser-Meryer-Olkin (Kaiser, 1970) (KMO) test is suggested to check whether the dataset is suitable to produce reliable and distinct factors (Field, 2015). The KMO statistic value varies between zero to one. A value of zero indicates a factor analysis is likely to be inappropriate. Kaiser (1970) suggests the value should be greater than 0.5 as a bare minimum. Values below 0.5 indicate either a need to collect more data or to reconsider what variables to include. When the overall KMO statistic value is unsatisfactory, it is useful to check the KMO value for individual variables to identify the problematic variables. This is shown in the anti-image correlation matrix (Field, 2013). The values are supposed to all be above 0.5. If any individual variable is below this value, we can exclude it from the analysis and re-examine the overall KMO value.

To check the sampling adequacy of the dataset in this study, a KMO test was performed for all the variables and the results showed that the KMO test of adequacy of sampling was .54. Since it was greater than 0.5, it therefore met the threshold of the KMO value. As for an exploratory analysis, the variables with the lowest KMO values ($p < .50$) were removed. Removing some of the items intended to explore the potential problematic individual items. The first trial was to remove individual variables which were in 0.30s (Item 7, 10 and 13) and the final KMO value for the retained variables (39 items) was improved to .576 . The second trial was further removing Item 14 apart from Item 7, 10 and 13; the KMO value was improved to .582 and the individual values were above .40s. To summarize, the KMO test for the all the original 42 questions met the minimum requirement but could be improved by removing some items. At this stage all of the variables (questions) were retained for later exploratory analysis.

Table 6.3 KMO test for 39 items

KMO and Bartlett's Test (Remove question 7, 10 and 13)		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.576
Bartlett's Test of Sphericity	Approx. Chi-Square	1106.336
	df	741
	Sig.	.000

Table 6.4 KMO test for 38 items

KMO and Bartlett's Test (Remove question 7, 10, 13 and 14)		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.582
Bartlett's Test of Sphericity	Approx. Chi-Square	1062.587
	df	703
	Sig.	0

Next, it is important to check if the dataset has a patterned relationship amongst the variables. This means two extreme situations should be avoided; the correlations between variables cannot be either too weak or too strong. First, if the correlations between variables are too low, we could not argue that the test questions measure the same underlying dimension(s) due to the weak relationship between the variables (Tabachnick and Fidell, 2007). The objective way to check whether the overall correlations are too small is Bartlett's test of Sphericity. Bartlett's test tests the null hypothesis that the original correlation matrix is an identity matrix. If the value is significant (significance level of $p < 0.05$) then it means the correlations of the sample are significantly different from zero. In this case, the Bartlett's test value of the sampling was $p = 0$, which was significant. This indicated the correlation matrix was not an identity matrix therefore there were some important relationships between the variables.

Second, if the variables are extremely highly correlated, it becomes impossible to determine the unique contribution to a factor of the variables (Field, 2013). In that case, it is difficult to identify what each factor contributes to the independent variables. A remedy to avoid the risk of this extremely high correlation is to eliminate variables with correlations higher than .90. Another check is the determinant value which can be detected by looking at the bottom of the correlation matrix. This determinant is supposed to be greater than the necessary value of 0.00001; otherwise it may cause Multicollinearity. When Multicollinearity occurs, one variable can be used to predict another variable so as to create redundant information. The dataset in this study had no variables with correlations higher than .90 suggesting there were no variables extremely highly correlated. The determinant value was 5.697^{-8} (which was smaller than 0.00001).

6.7.1.3 Factor Analysis and reliability analysis of whole dataset

After checking that the data-set met the assumptions required for conducting factor analysis, the next steps were extraction and rotation.

A Principal Component Analysis (PCA) was conducted on the whole dataset of 42 reading comprehension questions. The number of positive eigenvalues determines the number of factors/components to be extracted. This is supported by Rietveld and Van Hout (1993, p. 259), "the number of positive eigenvalues determines the number of dimensions needed to represent a set of scores without any loss of information." They further suggest some rules of thumb for determining how many factors should be retained (p. 273).

1. Retain only those factors with an eigenvalue larger than 1 (Guttman-Kaiser rule);
2. Keep the factors which, in total, account for 70-80% of the variance;
3. Create a screen-plot, keep all factors before the breaking point or elbow.

However, the Kaiser Criterion does not apply to datasets with average extracted communalities below .70 (Field, 2013). That is, we have to ensure that the average extracted communalities are at least more than .70 and keep those factors with an eigenvalue larger than 1. The average extracted communalities in this dataset was .72 indicating that it was appropriate to extract the factors with eigenvalues over 1. The results (Table 6.5, in Appendix) showed 16 components which in combination explained 71.95% of the variance.

Table 6.5 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	5.388	12.828	12.828	5.388	12.828	12.828	2.392
2	2.961	7.050	19.878	2.961	7.050	19.878	2.764
3	2.581	6.146	26.025	2.581	6.146	26.025	2.145
4	2.157	5.135	31.159	2.157	5.135	31.159	1.911
5	2.025	4.821	35.980	2.025	4.821	35.980	2.086
6	1.917	4.565	40.546	1.917	4.565	40.546	2.809
7	1.844	4.390	44.936	1.844	4.390	44.936	2.501

8	1.609	3.830	48.766	1.609	3.830	48.766	2.437
9	1.494	3.558	52.324	1.494	3.558	52.324	2.058
10	1.395	3.320	55.644	1.395	3.320	55.644	1.739
11	1.322	3.147	58.791	1.322	3.147	58.791	1.786
12	1.191	2.835	61.626	1.191	2.835	61.626	2.040
13	1.156	2.752	64.378	1.156	2.752	64.378	1.859
14	1.096	2.610	66.988	1.096	2.610	66.988	1.881
15	1.054	2.510	69.498	1.054	2.510	69.498	1.982
16	1.028	2.447	71.945	1.028	2.447	71.945	2.170
17	.985	2.346	74.291				
18	.924	2.200	76.491				
19	.835	1.987	78.478				
20	.777	1.849	80.328				
21	.717	1.708	82.036				
22	.678	1.614	83.650				
23	.630	1.500	85.150				
24	.622	1.482	86.632				
25	.555	1.322	87.954				
26	.527	1.255	89.209				
27	.474	1.128	90.338				
28	.465	1.106	91.444				

29	.449	1.069	92.513				
30	.410	.977	93.490				
31	.387	.922	94.412				
32	.323	.768	95.180				
33	.284	.677	95.857				
34	.277	.658	96.515				
35	.269	.639	97.155				
36	.232	.552	97.707				
37	.214	.510	98.217				
38	.179	.427	98.644				
39	.178	.423	99.067				
40	.141	.337	99.403				
41	.129	.306	99.709				
42	.122	.291	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

However, extraction alone cannot indicate the factors to be retained without the next step: rotation. Factors need to be rotated for better interpretation since extraction cannot provide information on factor loadings. That is, in terms of this study, it is not clear that what specific questions are loaded on the subskills in the reading battery. The goal of factor rotation is to have each variable load on as few factors as possible but to maximize the number of high loadings on each variable (Rummel, 1970). Thus, in this study, the variables that relate to the reading subskill of summarization for example, should load highly on the summarization factor but have zero loading on the subskill of predicting (if we presume that these sub-skills are not correlated). Hence, rotated solutions provide us detailed information on factor loadings for each individual variable and with such information we can further interpret the meaning of different factors.

The next step is to choose what type of rotation is appropriate for the dataset. Generally speaking, the choice is between orthogonal rotation and oblique rotation. Orthogonal rotation assumes that the factors are uncorrelated or not highly correlated (DeCoster, 1998; Rummel, 1970). Conversely, Oblique rotation assumes the variables are correlated with one another. This research rests on the assumption that reading is a trait which consists of sub-abilities (Hughes, 1989) and that these sub-abilities are to some degree related. Also, it will be recalled that the reading test questions (which map onto individual variables in the dataset) were designed to measure these underlying sub-abilities. Thus, it can be assumed that the 42 questions employed in the present reading test are related variables. Therefore, an oblique rotation was deemed the most appropriate type of rotation (in SPSS it refers to direct Oblimin option) for the present data set. The significance value of a factor loading was set as .4 (the widely-accepted criteria was .5 but .4 was considered in case of this study).

A FA of the whole dataset (Table 6.5) revealed 16 factors when a factor loading of .04 and above were taken into account. For each subscale, some items do not appear to load onto their own factor. For example, Questions 6, 15 and 29 loaded onto one factor. This is not consistent with the designed hypothetical framework. That is, Question 6 was supposed to measure outlining (literal); whilst Question 15 was deemed to capture the subskill of getting the main idea (literal); and Question 29 was supposed to measure inferring cause and effect (inferential). Thus, it seemed difficult to identify what this particular factor is because these three questions were supposed to measure three different underlying subskills. For this reason, factor analysis was conducted on 14 individual subscales in the next section after running it on the whole dataset.

A reliability analysis suggested that the overall alpha for the total 42 items was .79, which was considered as acceptable. The alpha improved from .79 to .80 when the following items “what factors will affect people making business decision in an organization?" and "List two consequences of economic recession mentioned in the text." Thus, it is suggested that these two items might be deleted. Yet it can also be argued that when deleting the items stated above the overall alpha reliability would not be improved dramatically. At this stage, a decision was made to keep all 42 items and examine the reliability at the subscale level.

6.7.1.4 Factor analysis and reliability analysis of subscales

In order to produce a valid and reliable test, 14 subscales representing 14 hypothetical reading subskills were revised on the basis of the factor analysis and reliability results for each individual subscale. A strategy for revising the measure at the subscale level was established. Specifically, any items with a loading of less than .04 on the first factor were deleted; any items that would improve the alpha for the subscale if they were deleted were also taken out. The alpha criteria of reliability below 0.5 was considered unacceptable.

Subscale 1 Recognition or recalling of details

Factor Analysis results showed that all three items loaded onto Factor 1 at 0.4 or above thus no item was deleted. Reliability analysis showed that the alpha was .54, and this cannot be improved by deleting any items. The decision was to retain all the items (3 items).

Subscale 2 Getting main ideas of a paragraph and larger body of text

Factor Analysis results showed that all three items loaded onto Factor 1 at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .62, and this cannot be improved by deleting any items. Thus, no changes to subscale 2 were required (retain all 3 items).

Subscale 3 Recognition or recalling comparison or cause and effect

Factor Analysis results showed that all three items loaded onto Factor 1 at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .54, and this cannot be improved by deleting any items. The conclusion was suggesting retaining all 3 items.

Subscale 4 Classifying

Factor analysis showed that all items loaded onto Factor 1 at 0.4 or above except for "How does the author classify the marketing environment?" state the question number too. Thus, this item was deleted. Reliability analysis showed that alpha was .34 to .40 when deleting the same above item. However, this subscale still did not meet the criteria .50. The conclusion was suggesting deleting the entire subscale. (3 items)

Subscale 5 Outlining

Factor analysis showed that all the items loaded onto Factor one at 0.4 or above except for the item “What changes would the increasing number of single-person households make to the market?” The reliability analysis showed that the alpha increased from .37 to .53 when deleting the following item “What changes would the increasing number of single-person households make to the market?” The conclusion was suggesting deleting the stated item (retain 2 items).

Subscale 6 Summarizing

Factor analysis showed all the items loaded onto Factor one at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .57 and this cannot be improved by deleting any items. The conclusion was retaining all 3 items.

Subscale 7 Synthesizing

Factor analysis showed that only one item loaded onto Factor one at above 0.4 thus to delete the other two items “What do governments do to control the economy and stabilize the markets?” and “What were the changes for the fifteen states members of the EU during the thirty years after the mid-1970?” question numbers. Reliability analysis showed that the alpha was .36 and this cannot be improved by deleting any items. The alpha fell short of meeting the criteria reliability level of .50. The conclusion was suggesting deleting the entire subscale.

Subscale 8 Inferring supporting details

Factor analysis showed all the items loaded onto Factor one at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .66 and this cannot be improved by deleting any items. The conclusion was retaining all 3 items.

Subscale 9 Inferring main ideas

Factor analysis showed all the items loaded onto Factor one at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .56 and this cannot be improved by deleting any items. The conclusion was retaining all 3 items.

Subscale 10 Inferring cause and effect

Factor analysis showed all the items loaded onto Factor one at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .65 and this cannot be improved by deleting any items. The conclusion was retaining all 3 items.

Subscale 11 Predicting outcomes

Factor analysis showed all the items loaded onto Factor one at 0.4 or above except for the item "Can you predict the results if the EU makes computers or office equipment supplies open to all members?". Add Question no. Thus, this item was deleted. The reliability analysis showed that the alpha improved from .50 to .65 when the following item was deleted "Can you predict the results if the EU makes computers or office equipment supplies open to all members?", thus this item should be deleted. The conclusion was deleting the item stated above and retaining the other 2 items.

Subscale 12 Judging facts and opinions

Factor analysis showed all the items loaded onto Factor one at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .51 and this cannot be improved by deleting any items. The conclusion was retaining all 3 items.

Subscale 13 Judging strengths and weakness of a position

Factor analysis showed all the items loaded onto Factor one at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .62 and this cannot be improved by deleting any items. The conclusion was retaining all 3 items.

Subscale 14 Judging adequacy and validity

Factor analysis showed all the items loaded onto Factor one at 0.4 or above thus no item to be deleted. Reliability analysis showed that the alpha was .70 and this cannot be improved by deleting any items. The conclusion was retaining all 3 items.

6.7.1.5 Post-hoc power analysis

At the completion of the study, the post-hoc power analysis was conducted to calculate if there were sufficient subjects (N = 120) to detect the difference between each group. For the total scores of literal and inferential questions, power analysis indicated a 99% chance of detecting an effect size (defined by Cohen, 1992) between the screen and print group as significant at the 5% level. However, power analysis indicated a 59% chance of detecting an effect size for the literal questions and a 99% chance for inferential questions between two groups as significant at the 5% level.

6.7.1.6 Section summary

In the reading test the total number of items and subscales was originally 42 and 14, however, after the revision this was reduced to 34 items and 12 subscales. The original subscales “classifying”, and “synthesizing” were deleted. Two items from the subscales “outlining” and “predicting outcomes” were also deleted. After that the overall alpha reliability of the retained 32 items was 0.78 and this could not be improved by deleting any items. The following table shows the revised 12 subscales and the reading comprehension questions within each subscale.

Table 6.6 The validated reading subscale and comprehension questions

Subscales	Questions
1. Recognition or recalling of details	1. What factors will affect people making business decision in an organization? 2. Please give one or two examples of different types of marketing factors mentioned in the text. 3. How do non-profit organizations benefit from government control?
2. Getting main ideas of a paragraph and larger body of text	4. What is the main idea of the Introduction section?

	5. In the "Real-life marketing" case (p.35), what was the main information used by the companies like Tie Rack and Sock Shop to make their marketing decisions? 6. What is the main idea of the section “cultural environment” (p.36)?
3. Recognition/Recalling comparison or cause and effect	7. How does the network in an organization bring people better living standards? 8. Why are the boundaries between the internal and external environment sometimes hard to decide? 9. Given that some macro factors are common for all firms, why will they affect firms differently?
4. Outlining	13. List some stakeholders that will have an effect on marketing decisions. 14. List the consequences of economic recession mentioned in the text.
5. Summarizing	16. Summarize the relationship between the firm and its environment. 17. What is a boom-and-bust economic cycle? 18. How do you illustrate the distinction between "necessities and discretionary purchases"?
6. Inferring supporting details	22. How does a firm relate to the environmental factors mentioned in the text? 23. What is the problem of classifying the macro and microenvironment factors? Please give examples. 24. Can you infer what measures the government took to recover the economy in 2008 recession?

7. Inferring main ideas	25. What conclusions can you draw from the description of the EU report in 2002? 26. What information does Figure 2.2 suggest? 27. What would happened to Spain after its government encouraged Latin American Spaniards to return home?
8. Inferring cause and effect	28. In the Costain West Africa case, why did the Costain Company survive despite the financial crisis in 2008? 29. Why is the external environment impossible to control? 30. What brought about the decrease in the birth rate of Eastern European countries during 1990?
9. Predicting outcomes	31. Can you predict other changes (not mentioned in the text) that increasing number of single-person households would bring to consumer-product companies? 33. What would be the consequences of depopulation and an aging population?
10. Judging facts and opinions	34. Can you think of any evidence to illustrate the increase in single-person households in several European countries? 35. What changes had already happened and what did the author predict in the future according to EU report in 2002? 36. Do you agree with the authors' expectations on demographic shifts as a result of expansion (p.34)?

11. Judging strengths and weakness of a position	37. How far do you think the conclusion that "the line between necessities and discretionary purchase is somewhat blurred"? 38. What do you think of the EU intervention in agriculture markets? 39. What is your opinion about the age demographic changes to the West European market?
12. Judging adequacy and validity	40. Is it true that the government in the 19th century was too poor to control the economy due to the defence of realm? Why? 41. How do you expect the working population to change in the original fifteen member states of EU? 42. Based on your own experience, can you give some other examples to illustrate "think small" rules (p.35)?

6.7.2 Statistical analysis in response to the research questions

The aim of the RCA is to investigate whether any differences exist in reading comprehension performances when reading from screen and print. The hybrid theoretical framework for reading comprehension assessment (Section 6.5) has provided a guideline to measure and compare reading comprehension performances from two dimensions (literal and inferential) and individual subscale-level (12 validated subscales) under two modes of text presentation. To recap, there were eight research questions regarding part one (RCA), RQ 1 to 3 aim to examine the possible differences of reading from screen and print when taking into account two dimension – literal and inferential; RQ 4-6 aim to examine the differences from the perspective of subscale-level. For a further examination of the nature of reading, RQ 7-8 aim to compare the differences between scores of L1-English and L1-Chinese-speaking participants. This followings discuss which statistical tests are most appropriate to use in relation to each of these questions.

6.7.2.1 Research question 1 – analysis plan

Do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

This involved one independent variable (mode of text presentation) with two levels (print group and screen group) and one dependent variable/outcome (total scores). When comparing the total scores between two groups, an Independent T-test will be used. However, when comparing the results from literal-meaning questions and inferential-meaning questions separately between screen and print group, a one-way MANOVA (multivariate analysis of variance) will be used because there is more than one dependent variable involved.

A plausible statistical test to answer research question 1 – 3 is to conduct a two-way MANOVA, for a situation where there is two or more dependent variables and two independent variable. This seems suitable for the current experimental design, including two dependent variables (scores on literal and inferential-meaning questions) and two independent variables (the mode of text presentation-screen versus print; the first language of the participants – English versus Chinese). However, the primary aim of doing a two-way MANOVA is to understand whether the effect of one independent variable on the dependent variables (collectively) is dependent on the value of the other independent variable, which is called an "interaction effect". Research question 1 – 3 is mainly interested on the "main effects" of each independent variable on the dependent variables, rather than their interactions. Thus, it was that the results from L1-English-speaking participants (RQ2) and L1-Chinese-speaking participants (RQ3) would be looked at separately.

6.7.2.2 Research question 2 – analysis plan

For L1-English-speaking participants, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

Similar to research question 1, an Independent T-test will be used to compare the total scores. A one-way MANOVA (multivariate analysis of variance) will be used to compare scores from two types of questions respectively. This is because there are two dependent variables involved (scores on literal-meaning questions and scores on inferential-meaning questions).

6.7.2.3 Research question 3 – analysis plan

For L1-Chinese-speaking participants, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

For the same reason as research question 2, an Independent T-test and a one-way MANOVA will be used.

6.7.2.4 Research question 4 – analysis plan

How do subskill scores differ across different text presentation for L1-English- and L1-Chinese-speakers together?

This involved one independent variable (mode of text presentation) with two levels (print group and screen group) and more than one dependent variable/outcome (scores on the validated 12 subscales). The dependent variables in a MANOVA test cannot be too corelated to each other, according to Tabachnick & Fidell (2012), suggesting that no correlation should be above r = .90. In effect, previous factor analysis (6.7.1.3) has checked the dataset in this study had no variables with correlations higher than .90. Theoretically, a one-way MANOVA can be used to compare the scores from 12 subscales/subskills respectively. Alternatively, multiple t-tests can also be used to compare the scores from screen and print on each subskill with taking a risk of obtaining Type 1 error.

6.7.2.5 Research question 5 – analysis plan

How do subskill scores differ across different text presentation for L1-English-speakers only?

Similar to research question 4, a one-way MANOVA or multiple t-tests will be used to compare scores from 12 subskills respectively.

6.7.2.6 Research question 6 – analysis plan

How do subskill scores differ across different text presentation for L1-Chinese-speakers only?

For the same reason as research question 5, a one-way MANOVA or multiple t-tests will be used will be used.

6.7.2.7 Research question 7 – analysis plan

When reading on screen how do subskill scores differ between L1-English and L1-Chinese participants?

Different from RQ1-5 with an interest to compare scores from screen and print, RQ7-8 aim to compare the scores from L1-English-speaking and L1-Chinese-speaking students. This involves one independent variable (the first language of participants) with two levels (L1-English-speaking and L1-Chinese-speaking participants) and 12 dependent variables (scores from screen on 12 subskills). Theoretically, a one-way MANOVA can be used. However, taking the small number of participants in each group (N=30) into account, multiple t-test is considered to be a safe alternative.

6.7.2.8 Research question 8 – analysis plan

When reading on print how do subskill scores differ between L1-English and L1-Chinese participants?

For the same reason as research question 7, a one-way MANOVA or multiple t-test can be used.

6.7.3 Checking assumptions for the statistical tests

Several assumptions needed to be fulfilled before conducting each statistical test to answer the research questions.

6.7.3.1 Research question 1 – assumption checking

Do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

An independent t-test is a parametric test based on the assumption of approximately normal distribution, therefore, the normality of the dependent variable was checked first. For *RQ 1 "Do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?* The total scores of both the screen and the print group were checked numerically by using a Kolmogorov-Smirnov test (for use when the sample size is greater than 50, which was the case here), and graphically by generating Normal Q-Q Plot.

The results of the Kolmogorov-Smirnov test suggested that the data for the dependent variable (total scores) was normally distributed, $p = .20$. In addition, the normal Q-Q plot graphically illustrated that the data were normally distributed as they follow a near-linear pattern (Figures 6.2 and 6.3).

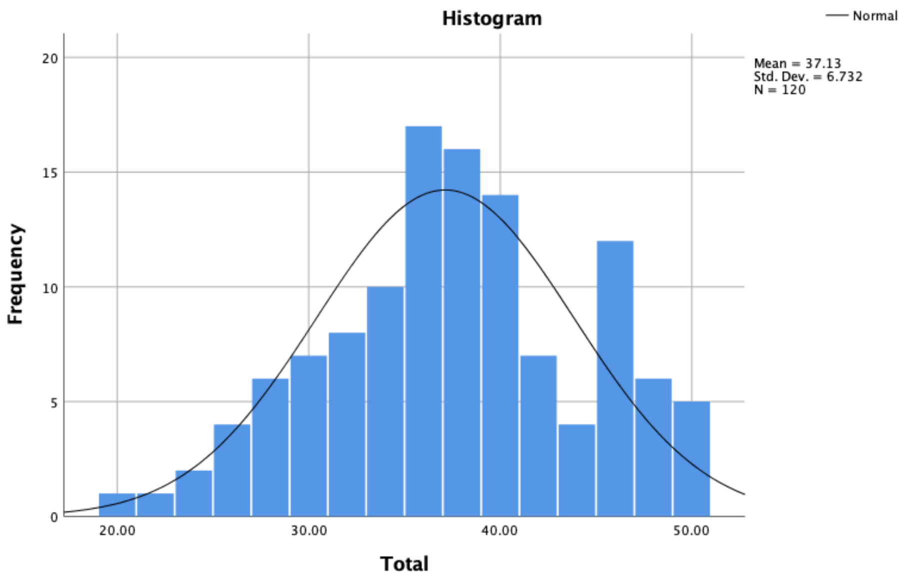


Figure 6.2 Histogram of total scores

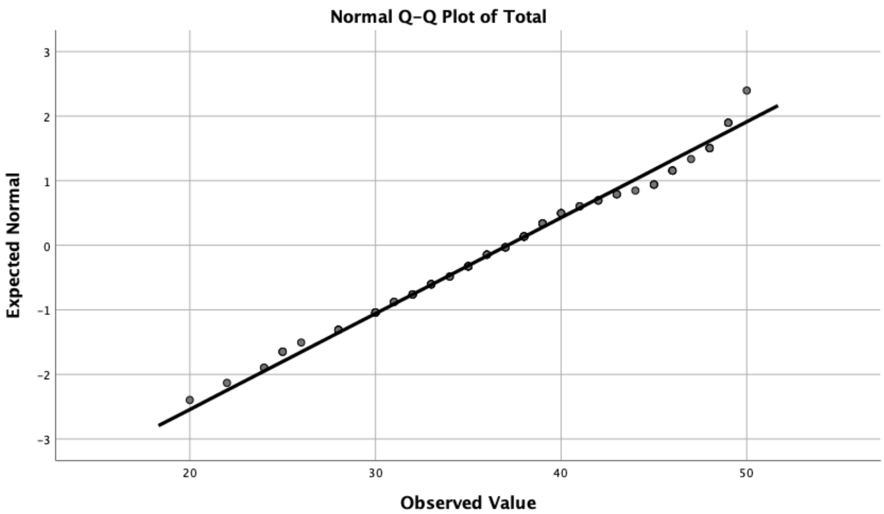


Figure 6.3 Q-Q plot of total scores

Four essential assumptions have to be met before conducting a MANOVA test.

a. Independence. There was no relationship between the observations in each group as different participants were assigned to the two groups (print and screen groups).
b. Random sampling. The data were randomly sampled as the participants were randomly selected from the population.
c. Multivariate normality. Multivariate normality cannot be checked directly at one time. Thus, the normality of each dependent variable (the literal-meaning questions and the inferential-meaning questions in this case) were checked numerically by using the Kolmogorov-Smirnov test, and graphically by using Normal Q-Q Plot.

The results of the Kolmogorov-Smirnov test suggested that one dependent variable – the scores of literal-meaning questions was not normally distributed, $p = .00$ ($p < .05$); another dependent variable – the scores of inferential-meaning questions was normally distributed $p = .08$ ($p > .05$). However, Field (2003, p. 184) explains that "even in a small sample size a significant test won't have the power to detect non-normality". Having examined the graphs in Figure 6.4 to 6.7 it is found that the data is not fairly skewed and follows an approximately near-linear pattern. Thus, the normal distribution can be assumed and it will not increase type 1 error rate in a MANOVA test.

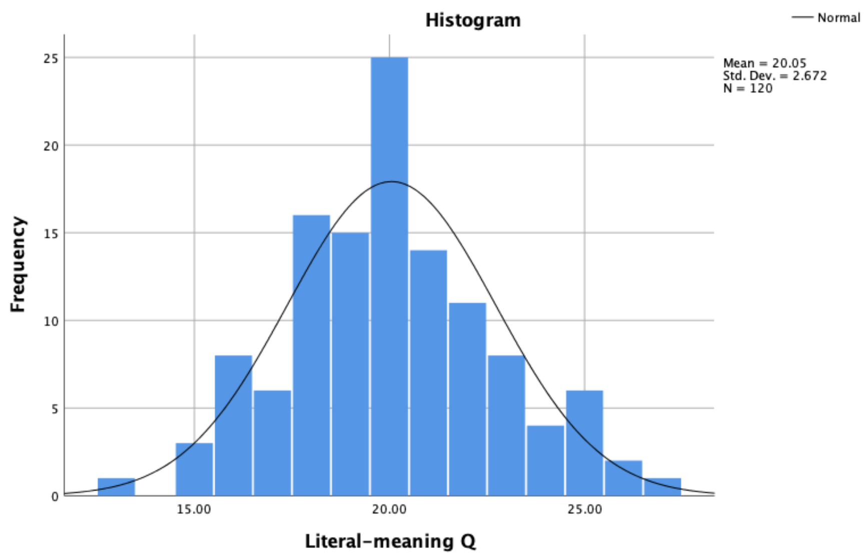


Figure 6.4 Histogram of literal-meaning questions

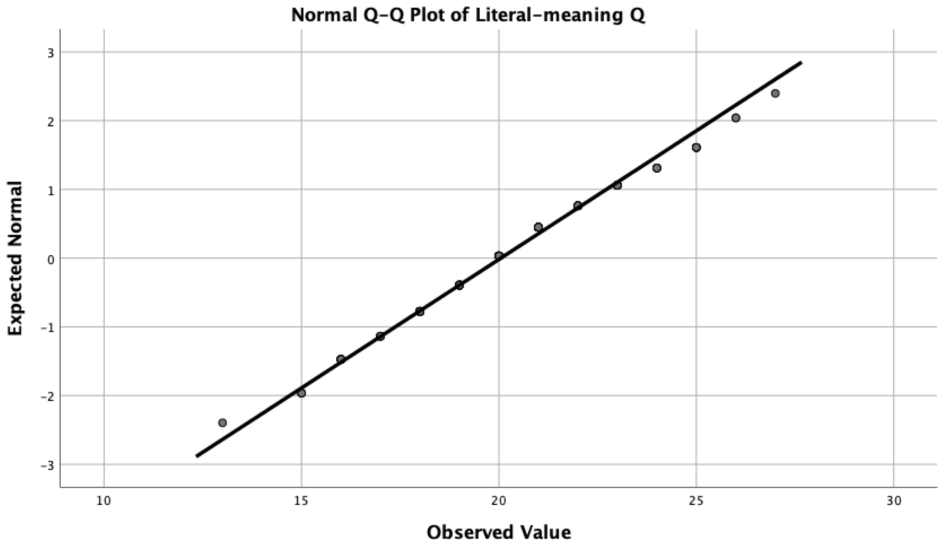


Figure 6.5 Q-Q plot of literal-meaning questions

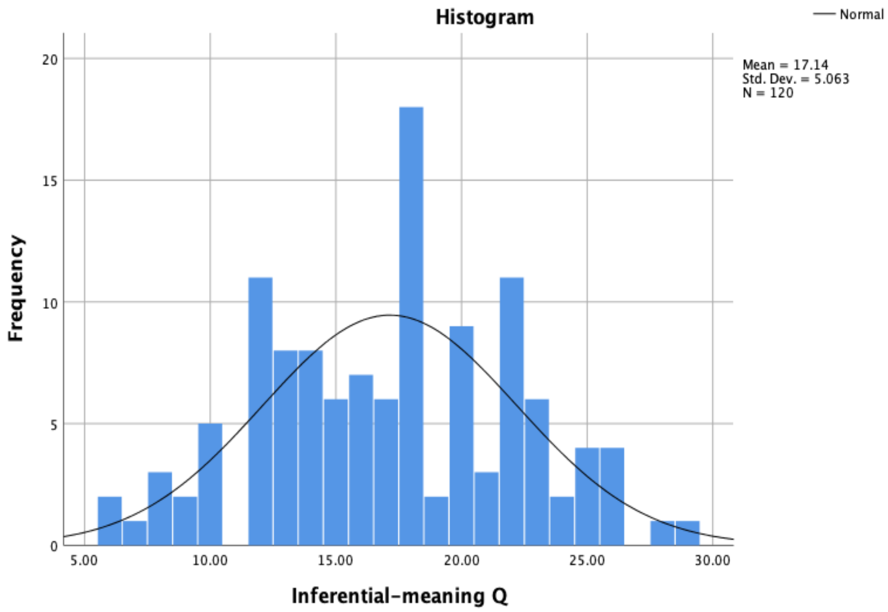


Figure 6.6 Histogram of inferential-meaning questions

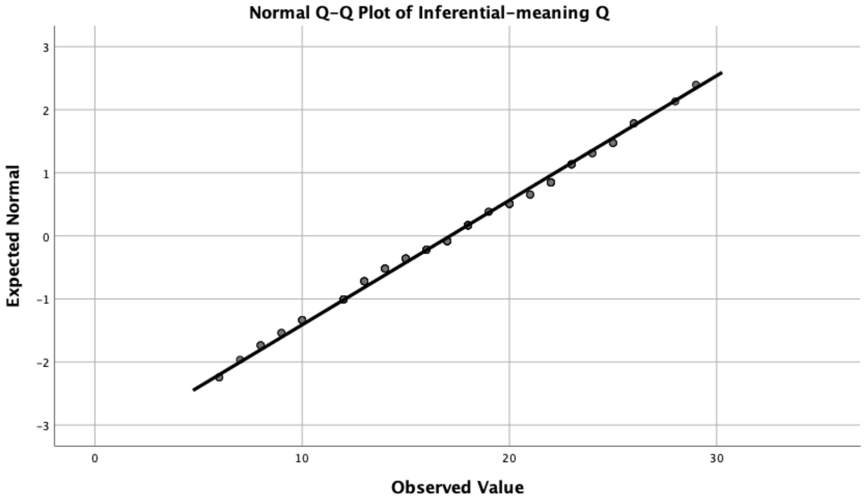


Figure 6.7 Q-Q plot of Inferential-meaning questions

d. Homogeneity of covariance matrices. A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the screen group and the print group were equal for both literal-meaning questions (p = .81) and inferential-meaning questions (p = .45). Thus, there was homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance ($p > .05$).

To summarize, as the assumptions have not been violated, an Independent t-test is appropriate to use to compare the total scores of reading comprehension performances between print and screen; a one-way MANOVA test is suitable to use to compare the scores from literal-meaning and inferential-meaning questions for L1-English-speaking and L1-Chinese-speaking participants together.

6.7.3.2 Research question 2 – assumption checking

For L1-English-speaking participants, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

In order to conduct an Independent t-test and a one-way MANOVA test to answer research question 2 for L1-English-speaking participants only, the data for the dependent variable (total scores from L1-English-speakers) should be normally distributed. Additionally, the dependent variables for scores for literal-meaning questions and scores for inferential-meaning questions should also meet the assumption of normal distribution and homogeneity of variance.

The total scores of L1-English-speakers were checked numerically by using a Kolmogorov-Smirnov test, and graphically by using a Normal Q-Q Plot. The Kolmogorov-Smirnov test suggested that the distribution of the data for total scores (p = .06) and for the inferential-meaning scores (p = .18) was normal; the scores for the literal-meaning questions was not normally distributed (p = .00). However, having examined the graphs from Figure 6.8 to 6.13，it is found that the data is not fairly skewed and follows an approximately near-linear pattern. Thus, the normal distribution can be assumed and it will not increase type 1 error rate in a MANOVA test.

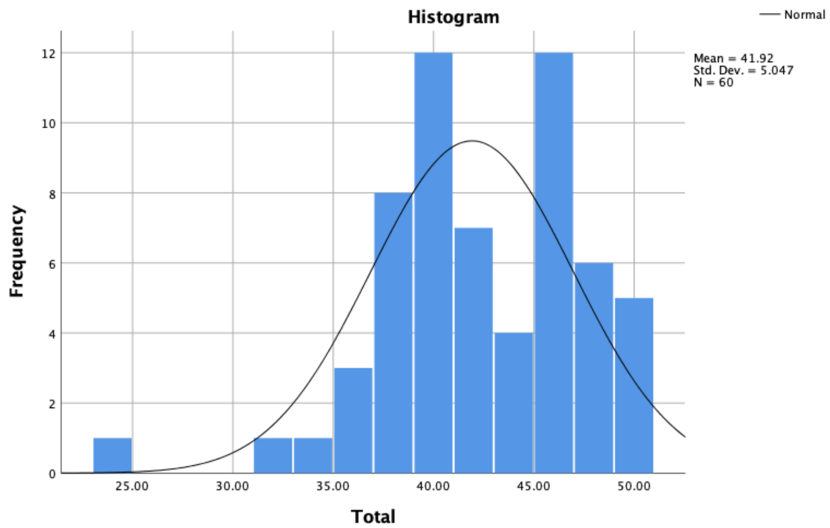


Figure 6.8 Histogram of total scores (L1-English-speaking)

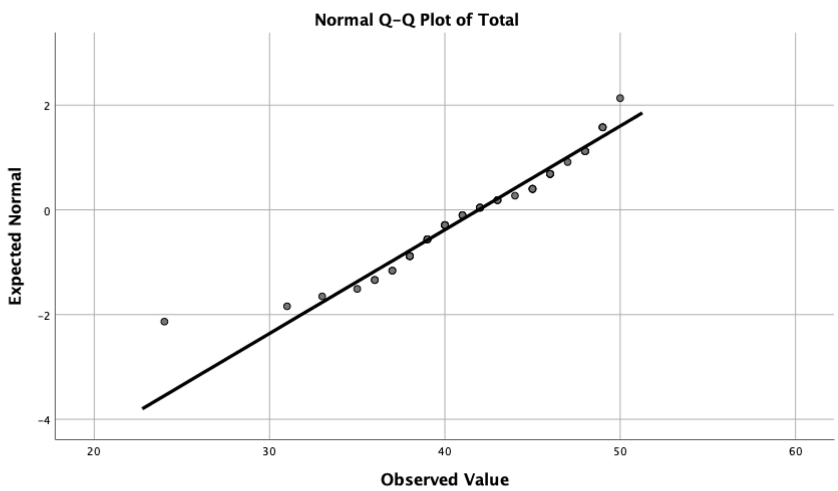


Figure 6.9 Q-Q plot of total scores (L1-English-speaking)

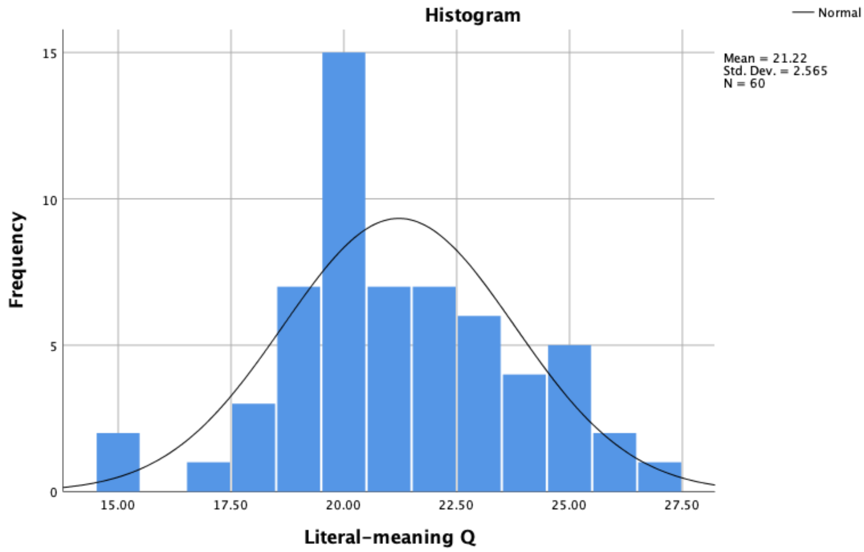


Figure 6.10 Histogram of literal-meaning questions (L1-English-speaking)

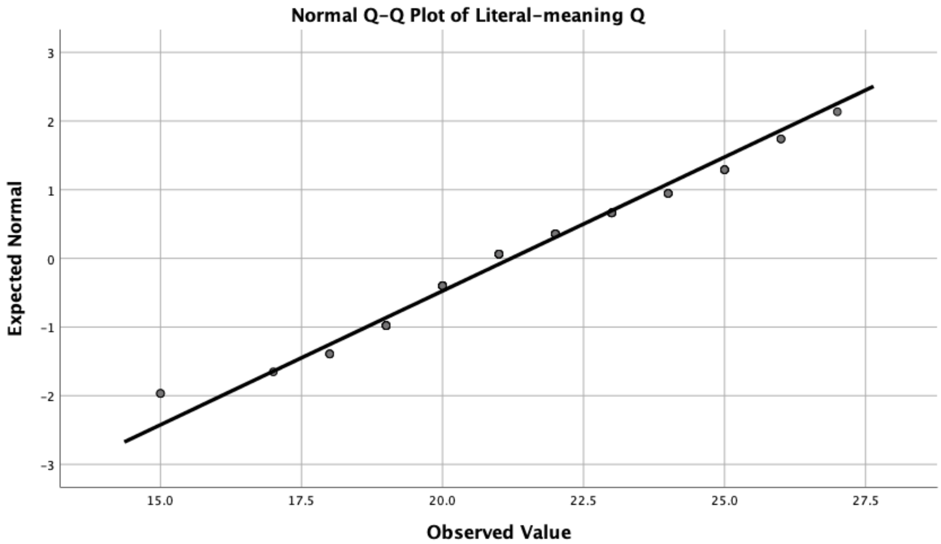


Figure 6.11 Q-Q plot of literal-meaning questions (L1-English-speaking participants)

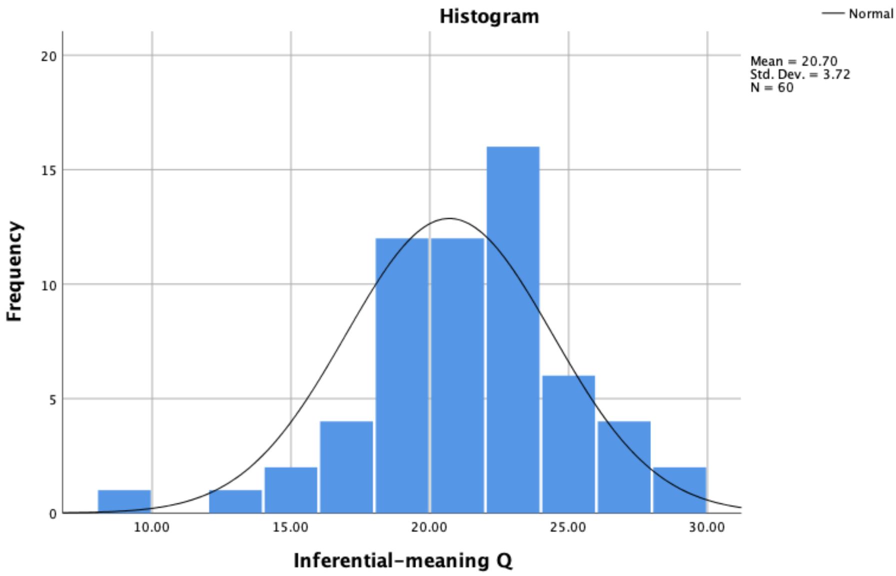


Figure 6.12 Histogram of inferential-meaning questions (L1-English-speaking)

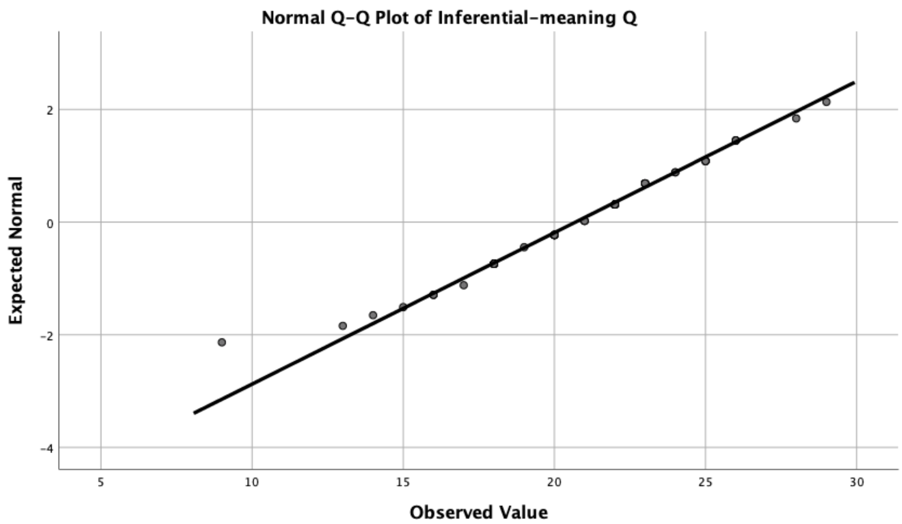


Figure 6.13 Q-Q plot of inferential-meaning questions (L1-English-speaking)

A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the screen group and the print group were equal for both literal-meaning questions (p = .95) and inferential-meaning questions (p = .91). Thus, there was homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance (p > .05).

Thus, it was appropriate to use an independent t-test to examine total scores and a one-way MANOVA test to compare separate scores of literal-meaning and inferential-meaning questions in research question 2.

6.7.3.3 Research question 3 – assumption checking

For L1-Chinese-speaking participants, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

In order to conduct an Independent t-test and a one-way MANOVA test to answer research question 3 for L1-Chinese-speaking participants only, the data for the dependent variable (total scores from L1-English-speakers) should be normally distributed. Additionally, the dependent variables for scores for literal-meaning questions and scores for inferential-meaning questions should also meet the assumption of normal distribution and homogeneity of variance.

The total scores of L1-Chinese-speakers were checked numerically by using a Kolmogorov-Smirnov test, and graphically by using a Normal Q-Q Plot. The Kolmogorov-Smirnov test suggested that the distribution of the data for total scores ($p = .16$) was not normal; but the scores for the literal-meaning scores ($p = .46$) and inferential-meaning questions was normally distributed ($p = .22$). In this case, an alternative way is to use a Mann-Whitney U test (equal to a t-test for the non-parametric data). However, Field (2013) mentions that, albeit the t-test is robust against normality, it is problematic only when the data has serious outliers or far from normal in a small sample size. Having examined the graphs from Figure 6.14 to 6.19, it is found that the data is not fairly skewed and without any serious outliers. At this stage, a decision was made to carry on an independent t-test to compare the total scores.

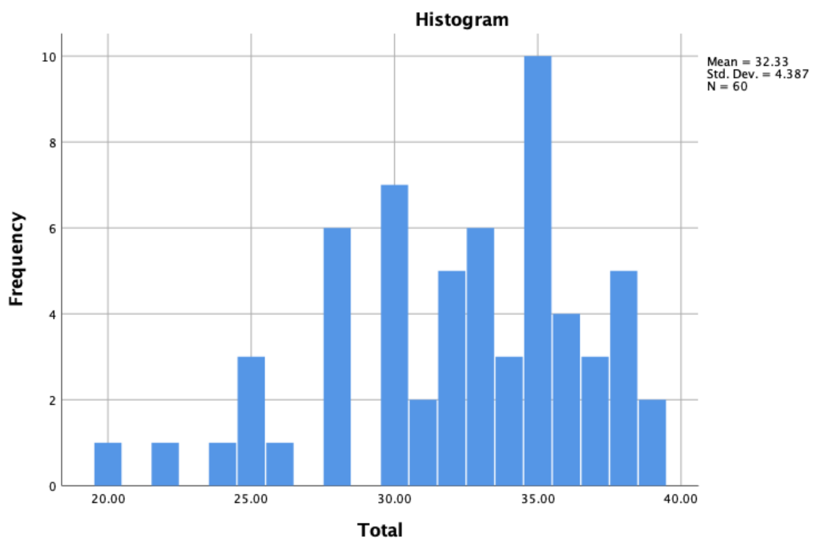


Figure 6.14 Histogram of total scores (L1-Chinese-speaking)

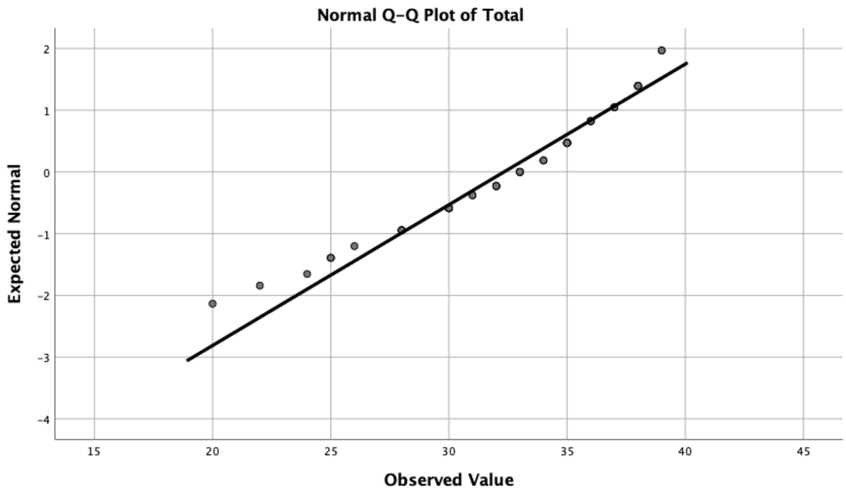


Figure 6.15 Q-Q plot of total scores (L1-Chinese-speaking)

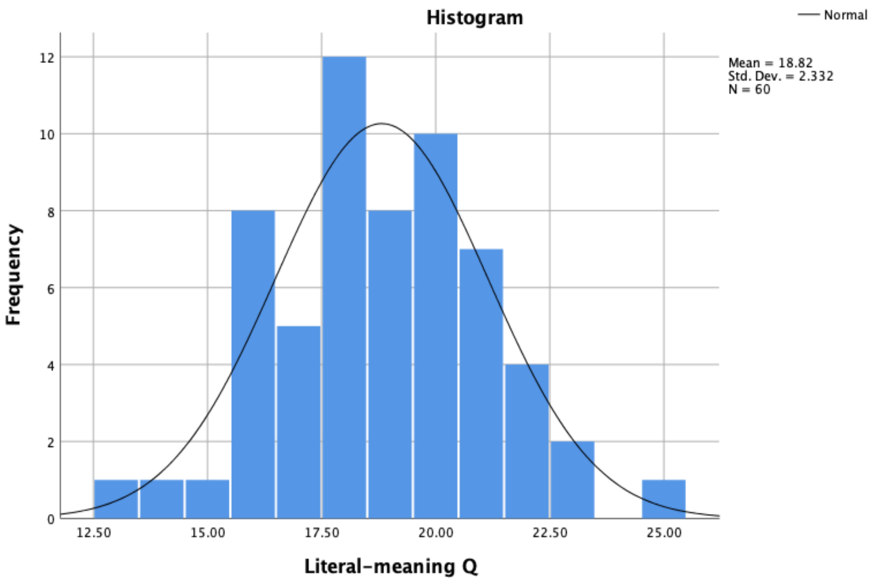


Figure 6.16 Histogram of literal-meaning scores (L1-Chinese-speaking)

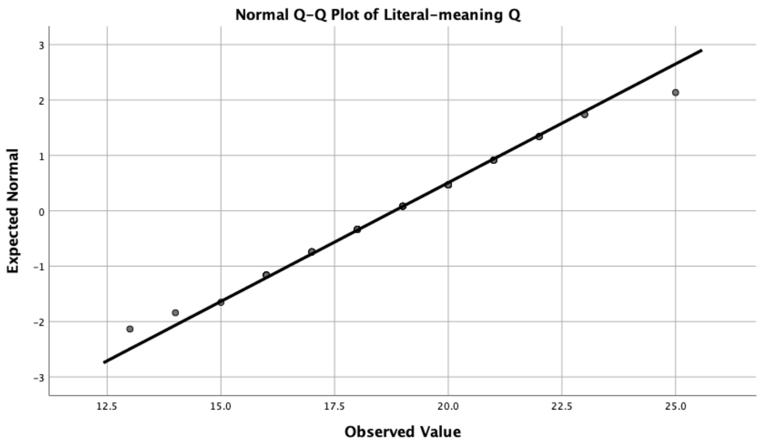


Figure 6.17 Q-Q plot of o literal-meaning scores (L1-Chinese-speaking)

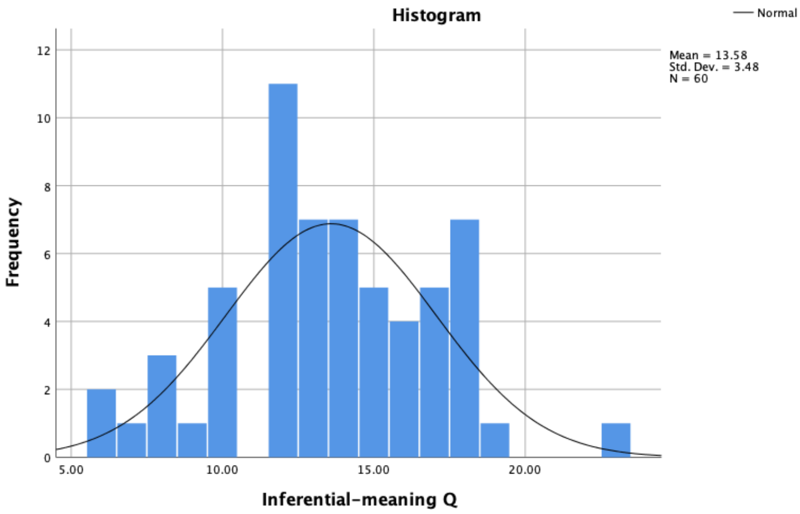


Figure 6.18 Histogram of inferential-meaning scores (L1-Chinese-speaking)

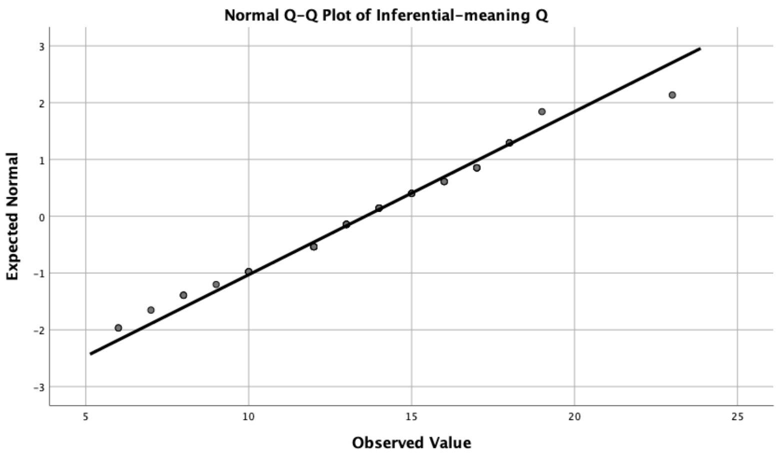


Figure 6.19 Q-Q plot of o literal-meaning scores (L1-Chinese-speaking)

A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the screen group and the print group were equal for both literal-meaning questions (p = .97) and inferential-meaning questions (p = .09). Thus, there was homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance ($p > .05$).

Thus, it was appropriate to use an independent t-test to examine total scores and a one-way MANOVA test to compare separate scores of literal-meaning and inferential-meaning questions in research question 3.

6.7.3.4 Research question 4 – assumption checking

How do subskill scores differ across different text presentation for L1-English- and L1-Chinese-speakers together?

In order to conduct a one-way MANOVA test to answer research question 4, the normality on each subskill were checked numerically by using a Kolmogorov-Smirnov test, and graphically by using a Normal Q-Q Plot. Although the Kolmogorov-Smirnov test suggested that the distribution of the data on all the subskill was not normal, $p < .00$, the data graphically showed no serious outliers or looked fairly skewed (see the following figures). The assumption of normal distribution can be assumed for the data on each subskill.

A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the print and the screen were equal for the following subskills: subskill 3 (p = .50), subskill 4 (p = .91), subskill 5 (p = .25), subskill 7 (p = .06), subskill 8 (p = .14), subskill 10 (p = .81), subskill 11 (p = .54), subskill 12 (p = .09); however, the variances between two groups for the subskill 1 (p = .00), subskill 2 (p = .04), subskill 6 (p = .00) and subskill 9 (p = .00) were not equal.

Therefore, the scores of some subskills did not fulfil the assumption of homogeneity of variance required to conduct a MANOVA test. Thus, the MANOVA was not an appropriate test to use for answering research question 4. Alternatively, multiple independent t-tests can be applied to compare separate scores of the 12 subskills.

6.7.3.5 Research question 5 – assumption checking

How do subskill scores differ across different text presentation for L1-English-speakers only?

Similar to RQ 4, the normality on each subskill for L1-English-speaking participants were checked numerically by using a Kolmogorov-Smirnov test, and graphically by using a Normal Q-Q Plot. The data on all the subskill did not pass the Kolmogorov-Smirnov test, suggesting that the distribution was not normal, $p < .00$. However, graphically, with an exception for subskill 4, the data from the other subskills showed no serious outliers or without looking fairly skewed (see the following figures). For subskill 4, the data showed a degree of negative skew. Therefore, the assumption of normal distribution can be assumed for the data on each subskill except on subskill 4.

A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the print and the screen were equal for the following subskills: subskill 1 ($p = .06$), subskill 4 ($p = .95$), subskill 7 ($p = .06$), subskill 8 ($p = .10$), subskill 9 ($p = .06$), subskill 10 ($p = .68$); however, the variances between two groups for the subskill 2 ($p = .00$), subskill 3 ($p = .03$), subskill 6 ($p = .02$) and subskill 11 ($p = .02$) subskill 12 ($p = .00$) were not equal.

Therefore, the scores of some subskills did not fulfil the assumption of normal distribution or homogeneity of variance required to conduct a MANOVA test. Thus, the MANOVA was not an appropriate test to use for answering research question 4. Alternatively, multiple independent t-tests can be applied to compare separate scores of the subskills: an independent t-test is used for the parametric data; a Mann-Whitney U test (a t-test counterpart) is used for the non-parametric data. A decision was made as follows: for subskill 4, a Mann-Whitney U test was used for subskill 4; independent t-tests were used for the other subskills.

6.7.3.6 Research question 6 – assumption checking

How do subskill scores differ across different text presentation for L1-Chinese-speakers only?

Similar to RQ 5, the Kolmogorov-Smirnov test suggested that the distribution of the data on all the subskill was not normal, $p < .00$. With an exception on subskill

1 and 7, the data from the other subskills graphically showed no serious outliers or without looking fairly skewed. For subskill 1, the data showed a degree of positive skewed; for subskill 7, the data showed a degree of negative skewed. Therefore, the assumption of normal distribution can be assumed for the data on each subskill except on subskill 1 and 7.

A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the print and the screen were equal for the following subskills: subskill 2 ($p = .11$), subskill 3 ($p = .39$), subskill 4 ($p = .57$), subskill 8 ($p = .52$), subskill 10 ($p = .84$), subskill 12 ($p = .45$); however, the variances between two groups for the subskill 1 ($p = .02$), subskill 5 ($p = .01$), subskill 6 ($p = .03$) and subskill 7 ($p = .03$), subskill 9 ($p = .01$), subskill 11 ($p = .03$) were not equal.

Therefore, the scores of some subskills did not fulfil the assumption of normal distribution or homogeneity of variance required to conduct a MANOVA test. Thus, the MANOVA was not an appropriate test to use for answering research question 4. Alternatively, multiple independent t-tests can be applied to compare separate scores of the subskills. A decision was made as follows: for subskill 1 and 7, a Mann-Whitney U test was run; multiple independent t-tests were run for the other subskills.

6.7.3.7 Research question 7 – assumption checking

When reading on screen how do subskill scores differ between L1-English and L1-Chinese participants?

The assumption of normality on 12 subskills were checked numerically and graphically. Numerically, the Kolmogorov-Smirnov and test suggested that the distribution of the data on all the subskill was not normal, $p < .00$. However, having graphically checking, with an exception on subskill 4, the data from the other subskills showed no serious outliers or without looking fairly skewed (see the following figures). For subskill 4, the data showed a degree of positive skewed. Therefore, the assumption of normal distribution can be assumed for the data on each subskill except on subskill 4.

A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the print and the screen were equal for the following subskills: subskill 1 ($p = .64$), subskill 2 ($p = .26$), subskill 4 ($p = .36$), subskill 7 ($p = .64$), subskill 8 ($p = .15$), subskill 9 ($p = .39$), subskill 10 ($p = .82$), subskill 12 ($p = .10$); however, the variances between two groups for the subskill 3 ($p < .00$), subskill 5 ($p < .00$), subskill 6 ($p = .03$) and subskill 11 ($p = .01$) were not equal.

Therefore, the scores of some subskills did not fulfil the assumption of normal distribution or homogeneity of variance required to conduct a MANOVA test. Thus, the MANOVA was not an appropriate test to use for answering research question 4. Alternatively, a decision was made as follows: for subskill 4, a Mann-Whitney U test was run; multiple independent t-tests were run for the other subskills.

6.7.3.8 Research question 8 – assumption checking

When reading on print how do subskill scores differ between L1-English and L1-Chinese participants?

The assumption of normality on 12 subskills were checked numerically and graphically. Numerically, the Kolmogorov-Smirnov and test suggested that the distribution of the data on all the subskill was not normal, $p < .00$. However, having graphically checking, the data from the subskills showed no serious outliers or without looking fairly skewed. Therefore, the assumption of normal distribution can be assumed for the data on each subskill.

A Levene's test was used to check the homogeneity of variance for each dependent variable. The results showed that the variances between the print and the screen were not equal for the following subskills: subskill 6 ($p = .10$) and subskill 12 ($p = .05$); however, the equal variances between two groups for the other subskills were met.

Therefore, although the scores of some subskills did not fulfil the assumption of homogeneity of variance required to conduct a MANOVA test, alternative multiple independent t-test can be used to compare separate subskills.

6.8 Conclusion

This chapter started with a brief description of the main study design, including three parts: RCA, WMCT and a follow-up interview. Then it focused on the methods (Section 6.4, including context, recruitment process, procedures, materials) and the theoretical assessment framework (Section 6.5) for designing and conducting the experimental reading test. A complete reading test was developed on this framework (6.6). Next, it moved to the most important section of this chapter – data analysis (6.7). The first part of data analysis (6.7.1) focused on the test validation by using factor analysis statistical technique. On a basis of the validated test, the remaining parts of data analysis (6.7. 2 and 6.7.3) discussed the appropriate statistical tests that may be run to answer the research questions. The following table (Table 6.7) explains the original decisions and final decisions regarding the statistical analysis planning.

Table 6.7 Statistical analysis planning decisions for each research question

Research questions	Original decisions	Final decisions	Violations of assumptions
RQ 1	Independent t-test and one-way MANOVA	Independent t-test and one-way MANOVA	NS
RQ 2	Independent t-test and one-way MANOVA	Independent t-test and one-way MANOVA	NS
RQ 3	Independent t-test and one-way MANOVA	Mann-Whitney U test and one-way MANOVA	Normality
RQ 4	One-way MANOVA	Multiple independent t-test and Mann-Whitney U test	Normality and homogeneity of variance
RQ 5	One-way MANOVA	Multiple independent t-test and Mann-Whitney U test	Normality and homogeneity of variance

RQ 6	One-way MANOVA	Multiple independent t-test and Mann-Whitney U test	Normality and homogeneity of variance
RQ 7	One-way MANOVA	Multiple independent t-test and Mann-Whitney U test	Normality and homogeneity of variance
RQ 8	One-way MANOVA	Multiple independent t-test	Homogeneity of variance

Chapter 7 Results of Reading Comprehension Assessment

7.1 Research question 1

Do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

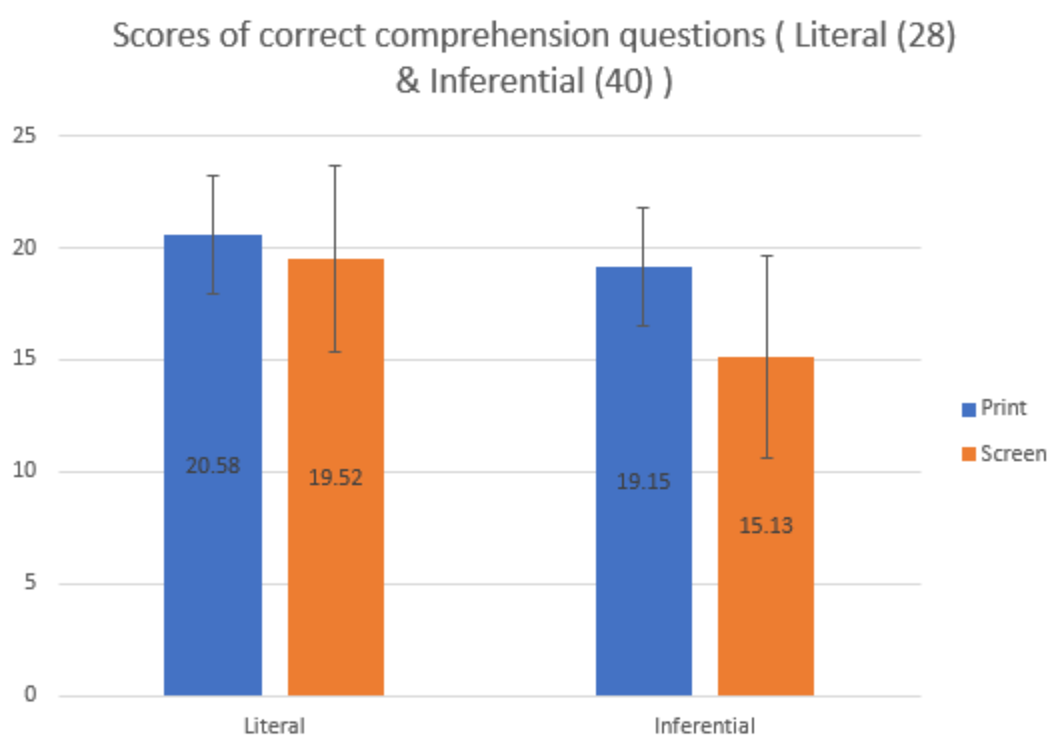


Figure 7.1 Scores for correct comprehension questions

An independent t-test was run to compare the total scores between print and screen. The participants who read on screen obtained a significantly lower mean score (Mean = 34.58, SD = 6.30) than those who read from print (Mean = 39.67, SD = 6.20), $t\ (118) = 4.45$, $p < .05$.

A one-way MANOVA was run to determine the effect of text presentation on reading comprehension performances. Two measures of comprehension performances were assessed: literal-meaning and inferential-meaning question scores. Preliminary assumption checking (section 6.7.3) revealed that data was satisfied. Participants from print and screen scored higher in their literal-meaning questions (M = 20.58, SD = 2.61; M = 19.52, SD = 2.64, respectively) than their inferential-meaning questions (M = 19.15, SD = 4.79; M = 15.13, SD = 4.54, respectively). The differences between the text presentation on the combined dependent variables was statistically significant, using Pillai's trace (as the sample size were equal), V = .16, $F(2, 117) = 11.15$, $p < .05$. Following univariate ANOVAs showed that only inferential-meaning question scores ($F(1, 118) = 22.25$, $p < .000$) were statistically different between the participants from screen and print group while literal-meaning question scores were not significantly different ($F(1, 118) = 12.11$, $p = .280$), using a adjusted significant level of .025 because there were two dependent variables.

Table 7.1 Question type, medium, mean scores comprehension questions answered correctly (Standard Deviations are in parentheses) and *p*-values

Question type	Medium	Mean (SD)	*Sig.*
Literal	Print	20.58 (2.61)	p = .280 (p > .025)
	Screen	19.52 (2.64)	
Inferential	Print	19.15 (4.79)	p < .000 (p < .025)
	Screen	15.13 (4.54)	
Total	Screen	39.67 (6.20)	p = .00 (p < .05)
	Print	34.58 (6.30)	

Overall it can be seen from Table 7.1, that participants performed better on print than on screen. The same trend can also be found in the participants when answering literal-meaning and inferential-meaning questions respectively, although the difference of literal questions was not significant.

In addition, it is surprisingly found the participants hardly used the dictionaries although the researcher provided in-built and physical dictionary for both screen and print participants. Given the low usage overall, the researcher did not compare dictionary usage across conditions and further pursue this analysis.

7.2 Research question 2

For L1-English-speakers only, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

An independent t-test was run to compare the total scores between print and screen from L1-English-speaking participants. The participants who read on screen obtained a significantly lower mean score (Mean = 38.80, SD = 4.55) than those who read from print (Mean = 45.03, SD =3.33), $t(58) = 6.01$, $p < .05$.

A one-way MANOVA was run to determine the effect of text presentation on reading comprehension performances. Two measures of comprehension performances were assessed: literal-meaning and inferential-meaning question scores. Preliminary assumption checking (section 6.7.3) revealed that data was satisfied. Participants from print and screen scored higher in their literal-meaning questions (M = 21.90, SD = 2.29 ; M = 20.53， SD = 2.67 respectively) than their inferential-meaning questions (M = 23.13, SD = 2.76 ; M = 18.26, SD = 2.86, respectively). The differences between the text presentation on the combined dependent variables was statistically significant, using Pillai's trace (as the sample size were equal), $V = .45$, $F(2, 57) = 22.82$, $p < .05$. Following univariate ANOVAs showed that only inferential-meaning question scores ($F(1, 58) = 44.67$, $p < .00$) were statistically different between the participants from screen and print group while literal-meaning question scores were not significantly different ($F(1, 58) = 4.51$, $p = .038$), using a adjusted significant level of .025 because there were two dependent variables.

Table 7.2 Question type, medium, mean scores of comprehension questions answered correctly (Standard Deviations are in parentheses) and *p*-values

Participants	Question type	Medium	Mean (SD)	Sig.
L1-English-speaking	Literal (28)	Print	21.90 (2.29)	p = .038 (p > .025)
		Screen	20.53 (2.67)	
	Inferential (40)	Print	23.13 (2.76)	p = .000

		Screen	18.26 (2.86)	(p < .025)
	Total	Print	45.03 (3.33)	p < .000
		Screen	38.80 (4.55)	(p < .05)

Similar to the results in Table 7.1, L1-English-speakers performed better on print than on screen. When answering inferential-meaning questions, L1-English-speaking participants still achieved significantly higher on print than on screen.

7.3 Research question 3

For L1-Chinese-speaking participants, do the scores on literal-meaning and inferential-meaning questions differ according to different text presentation?

An independent t-test was run to compare the total scores between print and screen from L1-Chinese-speaking participants. The participants who read on screen obtained a significantly lower mean score (Mean = 34.30, SD = 2.77) than those who read from print (Mean = 30.37, SD = 4.85), t (46.11) = 4.78, p < .05.

A one-way MANOVA was run to determine the effect of text presentation on reading comprehension performances. Two measures of comprehension performances were assessed: literal-meaning and inferential-meaning question scores. Preliminary assumption checking (section 6.7.3) revealed that data was satisfied. Participants from print and screen scored higher in their literal-meaning questions (M = 19.27, SD = 2.24 ; M = 18.36， SD = 2.37 respectively) than their inferential-meaning questions (M = 15.17, SD = 2.47; M = 12.00, SD = 3.65, respectively). The differences between the text presentation on the combined dependent variables was statistically significant, using Pillai's trace (as the sample size were equal), V = .23, F(2, 57) = 8.54, p < .05. Following univariate ANOVAs showed that only inferential-meaning question scores (F(1, 58) = 15.46, p = .000) were statistically different between the participants from screen and print group while literal-meaning question scores were not significantly different (F(1, 58) = 2.28, *p* = .140), using a adjusted significant level of .025 because there were two dependent variables.

Table 7.3 Question type, medium, mean scores of comprehension questions answered correctly (Standard Deviations are in parentheses) and *p*-values

Participants	Question type	Medium	Mean (SD)	Sig.
L1-Chinese speaking	Literal (28)	Print	19.27 (2.24)	p = .136 (p > .025)
		Screen	18.36 (2.37)	
	Inferential (40)	Print	15.17 (2.48)	p = .000 (p < .025)
		Screen	12.00 (3.65)	
	Total	Print	34.30 (2.77)	p = .00 p < .05
		Screen	30.37 (4.85)	

Table 7.3 compares the mean scores of literal, inferential and total comprehension questions answered correctly only by L1-Chinese-speaking participants across screen and print text presentation. Results from this table shows L1-Chinese-speakers performed significantly better on print than on screen for overall mean score. Not surprisingly, the significant difference was also found in answering inferential-meaning questions. However, the medium of text presentation had no statistically significant effect on L1-Chinese participants in answering literal-meaning questions. This outcome was clearly similar to L1-English-speaking participants gained.

The following Table 7.4 compares the data from Table 7.2 and Table 7.3, it is obviously shown that L1-English-speakers achieved better results than L1-Chinese-spearkers both on print (L1-English Mean = 39.67, SD = 6.20; L1-

Chinese Mean = 34.30, SD = 2.77) and screen (L1-English Mean = 34.58, SD = 6.30; L1-Chinese Mean = 30.37, SD = 4.85) condition for the total scores. Better performance were also found for L1-English-speakers at literal-meaning and inferential-meaning questions.

Table 7.4 Comparison of L1-English and L1-Chinese-speaking participants on question type, medium, mean scores of comprehension questions answered correctly and *p*-values

Participants	Question type	Medium	Mean (*SD*)	Sig.
L1 English	Literal	Print	20.58 (2.61)	*p* = .028 (*p* > .025, one-tailed test)
		Screen	19.52 (2.64)	
	Inferential	Print	19.15 (4.79)	*p* = .000 (p < .025, one-tailed test)
		Screen	15.13 (4.54)	
	Total	Print	39.67 (6.20)	*p* = 0.00 (*p* < .05, two-tailed test)
		Screen	34.58 (6.30)	

Participants	Question type	Medium	Mean (*SD*)	Sig.
L1 Chinese	Literal	Print	19.27 (2.24)	*p* = .136 (*p* > .025, one-tailed test)
		Screen	18.36 (2.37)	
	Inferential	Print	15.17 (2.48)	*p* = .000 (*p* < .025, one-tailed test)
		Screen	12.00 (3.65)	
	Total	Print	34.30 (2.77)	*p* = .00 (*p* < .05, two-tailed test)
		Screen	30.37 (4.85)	

7.4 Section summary

From Table 7.1, 7.2 and 7.3, it can be seen that the medium of text presentation had a significant effect on reading comprehension performances. On average participants achieved better results when reading on print than reading on screen. However, there was no significant difference in the scores between the screen and print conditions for the literal-meaning level of comprehension. A significant difference was apparent, however, in the inferential-meaning level of comprehension, with higher scores obtained after reading on print than on screen. Although better print performance than screen was clearly seen at literal and inferential-meaning level of comprehension, it is unknown whether the text medium had the same effect on the subskill level of reading. Thus, it is worthwhile to further investigate whether there are any differences on subskills between screen and print conditions.

7.5 Research question 4

How do subskill scores differ across different text presentation for L1-English- and L1-Chinese-speakers together?

The validated RCA was comprised of 12 subscales (discussed in Section 6.7.1). Each subscale represented one reading subskill. Multiple independent t-tests were conducted to compare the subskill scores between print and screen. The literal-meaning questions consisted of five subskills, shown in Table 7.5.

Table 7.5 Subskills, medium, mean scores of literal comprehension questions answered correctly by L1-English and L1-Chinese speakers together (Standard Deviations are in parentheses) and p-values

Types of subskills	Subskills	Medium	Mean (SD)	Sig.
Literal	1.Recognition or recalling of details	Print	4.62 (1.26)	.07
		Screen	4.98	

		(0.12)	
2. Getting main ideas of a paragraph and larger body of text	Print	4.52 (1.05)	.63
	Screen	4.42 (1.21)	
3. Recognition or recalling comparison or cause and effect	Print	4.32 (1.05)	.00
	Screen	3.67 (1.20)	
4. Outlining	Print	3.25 (0.97)	.11
	Screen	2.97 (0.96)	
5.Summarizing	Print	3.88 (1.26)	.00
	Screen	3.25 (1.19)	

Table 7.5 indicates that no difference was found in the following: *Subskill 1 - recognition or recalling of details* (Mean print = 4.62; SD = 1.26; Mean screen = 4.98; SD = 0.12; t(118) = -1.82, p= .07); *subskill 2 - getting main ideas of a paragraph and larger body of text and subskill* (Mean print = 4.52; SD= 1.05; Mean screen = 4.42; SD = 1.21; t(18) = .48, p = .63); *subskill 4 – outlining* (Mean print = 3.25; SD = 0.97; Mean screen= 2.97; SD = .96; t(118) = 1.62, p = .11). This means the medium of text presentation had no effect on these three subskills, which were consistent with the general trend of literal-meaning questions.

However, interestingly, a significant difference was found on the *subskill 3 - recognition/recalling comparison or cause and effect* (Mean print = 4.32; SD = 1.05; Mean screen = 3.67; SD = 1.20; t(118) = 3.15, p = .00) and *subskill 5 – summarizing* (Mean print = 3.88, SD = 1.26; Mean screen = 3.25,SD = 1.19, , t(118) = 2.83, p = .00), meaning, participants performed significantly better on print than on screen on both subskills. Next it is necessary to examine whether the inconsistency was apparent for certain inferential-meaning subskills. Thus, participants achieved better results on print than on screen for subskill 3 – recognition/recalling comparison or cause and effect as well as subskill 5 – summarizing, but no statistical difference was shown between two groups on the other literal-meaning subskills.

The inferential-meaning questions consisted of seven subskills, which is shown in the Table 7.6.

Table 7.6 Subskills, medium, mean scores of inferential comprehension questions answered correctly by L1-English and L1-Chinese speakers together (Standard Deviations are in parentheses) and p-values

Types of subskills	Subskills	Medium	Mean (SD)	Sig.
Inferential	6. Inferring supporting details	Print	2.12 (1.08)	.00
		Screen	3.03 (0.69)	
	7. Inferring main ideas	Print	2.87 (0.98)	.52
		Screen	2.98 (0.99)	
	8. Inferring cause and effect	Print	2.98 (1.13)	.00
		Screen	2.19 (1.00)	
	9. Predicting outcomes	Print	2.08 (0.67)	.00
		Screen	1.02 (0.89)	
		Print	3.21 (1.08)	.00

	10. Judging facts and opinions	Screen	2.63 (0.90)	
	11. Judging strengths and weakness of a position	Print	2.85 (1.18)	.00
		Screen	1.70 (1.15)	
	12. Judging adequacy and validity	Print	3.03 (1.31)	.00
		Screen	1.55 (1.02)	

As can be seen from the table above, participants performed significantly differently across screen and print text presentations with evident trend of better outcome on print for *subskill 8 - inferring cause and effect* (Mean print = 2.98, SD = 1.13; Mean screen = 2.20, SD = 0.90; t(118)= 4.21, p = .00); *subskill 9 - predicting outcomes* (Mean print = 2.08, SD = 0.67; Mean screen = 1.02, SD = 0.89; t(118)= 7.40, p= .00); *subskill 10 - judging facts and opinions* (Mean print = 3.21, SD = 1.08; Mean screen = 2.63, SD = 0.90; t(118)= 3.21, p= .00); *subskill 11 - judging strengths and weakness of a position* (Mean print = 2.85, SD = 1.18; Mean screen = 1.70, SD = 1.15; t(118)= 5.41, p= .00); *subskill 12* – judging adequacy and validity (Mean print = 1.55, SD = 1.02; Mean screen = 1.55, SD = 1.02; t(118) = 6.92, p = .00).

Surprisingly, the data of *subskill 7 – inferring main ideas* indicated a noticeable inconsistency (Mean print = 2.87, SD = 0.98; Mean screen = 2.98, SD = 1.00; t(118) = -0.65, p = .52), compared with the general trend of inferential-meaning questions and the other inferential subskills. In other words, it is unusual to see that the text presentation had no significant effect on the results of subskill 7 with even slightly higher mean score on screen than print. Furthermore, significant higher mean scores on screen than on print were found on *Subskill 6 - inferring supporting details* (Mean print = 2.12, SD = 1.08; Mean screen = 3.03, SD = 0.69; t(118) = -5.57, p = .00). Thus it is necessary to further investigate the factors contributing to the results.

To summarize, the participants (L1-English and L1-Chinese-speaking together) in the print condition performed significantly better than participants from screen condition on the following literal subskills: 3 and 5; and the following inferential subskills: 8, 9, 10, 11, and 12. Nevertheless, on the remaining subskills, participants between print and screen had no significantly different performance.

7.6 Research question 5

How do subskill scores differ across different text presentation for L1-English-speakers only?

For L1-English-speakers, the scores were not significantly different on all five literal-meaning subskills although participants in print condition tended to achieved higher than those in screen condition. The data were shown in Table 7.7, subskill 1 (Mean print = 4.73, SD = 1.17; Mean screen= 4.87, SD = .94; t(58) = .49, p = .63); subskill 2 (Mean print= 4.87, SD = .94; Mean screen = 4.50, SD = 1.28; t(58) = -1.27, p = .21); subskill 3 (Mean print = 4.80, SD= .89; Mean screen= 4.30, SD = 1.21; t(58)= -1.83, p= .07) and subskill 5 (Mean print= 4.20, SD = 1.13; Mean screen= 3.93, SD =1.20; t(58) = -.89 , p = .38).

In addition, a Mann-Whitney U test was run to determine if there were differences in subskill 4. Distributions of the scores on print and screen were not similar, as assessed by visual inspection. Scores for print (mean rank = 34.43) and screen (mean rank = 26.57) were not statistically significantly different, U = 332, z = -1.86, p = .06, using an exact sampling distribution for U (Dineen & Blakesley, 1973).

This was in line with the previous results from RQ 1, which indicated non-significant difference between participants in screen and print condition on literal-meaning questions. Nevertheless, it might also be argued that the p value (.07) of subskill 3 - recognition or recalling comparison or cause and effect was just narrowly missed the significance criteria (.05), if comparing with the other subskills at literal-meaning level. That means L1-English-speaking participants in print condition could possibly achieve better results than those in screen condition. Furthermore, on subskill 5 – summarizing L1-English-speakers in print condition did not perform significantly better than those in screen condition, which indicated opposite results to the overall trend of L1-English and L1-Chinese speakers together.

Table 7.7 Subskills, medium, mean scores of literal comprehension questions answered correctly by L1-English-speakers only (Standard Deviations are in parentheses) and p-values

Types of subskills	Subskills	Medium	Mean (SD)	Sig.
Literal	1.Recognition or recalling of details	Print	4.73 (1.17)	.63
		Screen	4.87 (0.94)	
	2. Getting main ideas of a paragraph and larger body of text	Print	4.87 (0.94)	.21
		Screen	4.50 (1.28)	
	3. Recognition or recalling comparison or cause and effect	Print	4.80 (0.89)	.07
		Screen	4.30 (1.21)	
	4. Outlining	Print	3.30 (0.95)	.63
		Screen	2.93 (0.87)	
	5.Summarizing	Print	4.20 (1.13)	.38
		Screen	3.93 (1.20)	

Table 7.8 Subskills, medium, mean scores of inferential comprehension questions answered correctly by L1-English speakers only (Standard Deviations are in parentheses) and p-values

Types of subskills	Subskills	Medium	Mean (SD)	Sig.
Inferential	6. Inferring supporting details	Print	2.43 (1.28)	.00
		Screen	3.30 (0.75)	
	7. Inferring main ideas	Print	3.63 (0.61)	.51
		Screen	3.50 (0.90)	
	8. Inferring cause and effect	Print	3.67 (0.96)	.00
		Screen	2.57 (0.68)	
	9. Predicting outcomes	Print	2.27 (0.64)	.00
		Screen	1.47 (0.73)	
	10. Judging facts and opinions	Print	3.87 (0.90)	.00
		Screen	2.80 (1.00)	
	11. Judging strengths and weakness of a position	Screen	3.60 (0.97)	.00
		Print	2.50 (0.63)	
	12. Judging adequacy and validity	Screen	3.67 (1.21)	.00
		Print	2.10 (0.76)	

For inferential-level subskills, L1-English-speakers performed in accordance with the overall trend of L1-English and L1- Chinese-speakers together: significantly better results on print than on screen for subskills 8, 9, 10, 11 and 12; whilst no significant difference for subskill 7 – inferring main ideas between L1-English-speaking participants in screen and print condition, (Mean print= 2.12, SD = 1.08, Mean screen = 3.02, SD = 0.67; t(58) = -.67, p = .51). The most striking result to emerge was that participants in screen condition performed significantly better than those in print condition for subskill 6 (Mean print= 2.43, SD = 1.28, Mean screen = 3.30, SD = .75, t(58) = 3.20, p = .00). Thus, an imperative step is to investigate the nature of the reading subskills and explore participants' reading process and strategies.

In summary, the text medium had no significant effect for L1-English-speaking participants in the screen condition and print condition on all literal-meaning subskills. However, participants in the print condition performed significantly better than those in the screen condition on the following inferential subskills: 8, 9, 10, 11, and 12. On inferential subskill 6, oppositely, participants in the screen condition achieved better results than in the print condition.

7.7 Research question 6

How do subskill scores differ across different text presentation for L1-Chinese-speakers only?

Different from L1-English-speaking participants showing no different results on all literal-meaning subskills, L1-Chinese participants achieved significantly higher scores on print than on screen for subskill 3 - recognition or recalling comparison or cause and effect (Mean print= 3.80, SD = .99, Mean screen = 3.20, SD= .94, t(58) = 2.42, p = .02) and subskill 5 - summarizing (Mean print = 3.57, SD = 1.33, Mean screen = 2.57, SD = .68, t(43.14) = 3.67, p = .00). However, no statistical difference was shown on the other subskills. The data was shown in Table 7.9.

Table 7.9 Subskills, medium, mean scores of literal comprehension questions answered correctly by L1-Chinese-speakers only (Standard Deviations are in parentheses) and p-values

Types of subskills	Subskills	Types of text presentation	Mean (SD)	Sig.
Literal	1.Recognition or recalling of details	Print	4.50 (1.05)	.11
		Screen	5.10 (0.88)	
	2. Getting main ideas of a paragraph and larger body of text	Print	4.17 (1.05)	.56
		Screen	4.30 (1.15)	
	3. Recognition or recalling comparison or cause and effect	Print	3.80 (0.99)	.02
		Screen	3.20 (0.94)	
	4. Outlining	Print	3,20 (1.00)	.81
		Screen	3.13 (1.17)	
	5.Summarizing	Print	3.13 (1.33)	.00
		Screen	2.57 (0.68)	

Table 7.10 Subskills, medium, mean scores of inferential comprehension questions answered correctly by L1-Chinese-speakers only (Standard Deviations are in parentheses) and p-values

Types of subskills	Subskills	Medium	Mean (SD)	Sig.
Inferential	6. Inferring supporting details	Print	1.80 (0.71)	.00
		Screen	2.77 (0.48)	
	7. Inferring main ideas	Print	1.10 (0.61)	.01
		Screen	2.58 (0.87)	
	8. Inferring cause and effect	Print	1.85 (0.91)	.05
		Screen	2.30 (0.84)	
	9. Predicting outcomes	Print	0.85 (1.20)	.00
		Screen	1.90 (0.66)	
	10. Judging facts and opinions	Print	2.55 (0.79)	.63
		Screen	2.57 (0.82)	
	11. Judging strengths and weakness of a position	Print	1.15 (1.25)	.00
		Screen	2.10 (0.85)	
	12. Judging adequacy and validity	Print	1.18 (1.07)	.00
		Screen	2.40 (1.10)	

As can be seen from Table 7.10, L1-Chinese-speaking participants in print condition achieved significantly better than on screen condition on subskill 8 - inferring cause and effect (Mean print = 2.30, SD = .84, Mean screen = 1.83, SD

= .95, t(58) = -2.02, p< .05), subskill 9 - predicting outcomes (Mean print= 1.90, SD = .66, Mean screen = .57, SD = .82, t(55.60) = 6.95, p < .00), subskill 11 - judging strengths and weakness of a position (Mean print = 2.10, SD = .85, Mean screen = .90, SD = 1.00, t(56.52) = 5.04, p= .03) and subskill 12 - Judging adequacy and validity (Mean print = 2.40, SD = 1.10, Mean screen = 1.00, SD = .95, t(58) = 5.28, p < .00).

However, the most striking aspect from the data was to find the statistical higher scores on subskill 6 and subskill 7 for L1-Chinese-speaking participants in screen condition than in print condition, *subskill 6* - Inferring supporting details (Mean print= 1.80, SD = .72, M-screen = 2.77, SD = .50, t(52.14)= -6.05, p < .00) and *subskill 7* - Inferring main ideas (Print Mean rank = 25.37, Screen Mean rank = 33.97 , U = 296, z = -2.49, p = .01). The possible explanation to this needs to be traced back to the specific questions at subskill 6 and subskill 7, which will be discussed in next chapter.

These results suggest L1-Chinese-speaking participants in the print condition performed better than participants in the screen condition on literal-meaning subskill 3 – recognition/recalling comparison or cause and effect, and subskill 5 – summarizing, and the following inferential subskills: 8, 9, 11, and 12. On inferential subskill 6 and subskill 7, on the other hand, participants in the screen condition achieved better results than in print condition.

7.8 Section summary

Taking together, the results from question 4 to 6 provided insights to whether the text medium have effect on individual reading subskills for L1-English- and L1-Chinese-speaking participants. First, the general trend of L1-English and L1-Chinese speakers together was significant better performance in print condition than in screen condition on the following literal-meaning subskills 3 - recognition/recalling comparison or cause and effect, and 5 – summarizing, and the following inferential subscales: 8, 9, 10, 11, and 12. Second, the text medium had no significant effect for L1-English-speaking participants in the screen condition and print condition on all literal-meaning subskills but had an effect on the following inferential subskills: 8, 9, 10, 11, and 12. Third, unlike L1-Enlgish-speakers, L1-Chinese-speakers participants in the print condition performed better than participants in the screen condition on literal-meaning subskill 3 subskill 5; and the following inferential subskills: 8, 9, 11, and 12. Last, both L1-English- and L1-Chinese-speaking participants indicated an interesting fact on inferential subskill 6 - inferring supporting details, that better performance was in the screen condition than print condition.

7.9 Research question 7

When reading on screen how do the subskill differ between L1-English- and L1-Chinese-speakers?

L1-English speakers obtained significant higher mean score only for literal-level subskills 3 (Mean English= 4.80, SD= .88, Mean Chinese= 3.83, SD = .96, $t(50.65) = 4.77$, $p < .00$) and subskill 5 (Mean English= 3.93, SD= 1.20, Mean Chinese = 2.57, SD = 1.68, $t(45.81) = 5.42$, $p < .00$). In terms of other literal-level subskills (subskill1, subskill 2, subskill 4), L1-Chinese got lower mean score although the difference between L1-English and L1-Chinese participants was not significant.

With regards to the seven inferential-level subskills, L1-English participants all obtained significantly higher mean score than L1-Chinese participants from subskill 6 (Mean English= 3.30, SD= .75, Mean Chinese = 2.77, SD= .50, $t(50.77) = 3.23$, $p < .00$) to subskill 12 (Mean English = 2.10, SD = .76, Mean Chinese = 1.00, SD = .95, $t(58)= 4.96$, $p < .00$), with an exception for subskill 10, as shown in Table 7.11.

Table 7.11 Scores for subskills of between L1-English and L1-Chinese (Screen)

Types of subskills	Subskills	Participants	Mean	SD	Sig.
Literal	1	L1-English	4.86	.94	.34
		L1-Chinese	5.10	.88	

	2	L1-English	4.50	1.28	.60
		L1-Chinese	4.33	1.15	
	3	L1-English	4.30	1.20	.00
		L1-Chinese	3.03	.81	
	4	L1-English	2.93	.87	.79
		L1-Chinese	3.00	1.05	
	5	L1-English	3.93	1.20	.00
		L1-Chinese	2.57	.68	
Inferential	6	L1-English	3.30	.75	.02
		L1-Chinese	2.77	.50	
	7	L1-English	3.50	.90	.00
		L1-Chinese	2.47	.82	
	8	L1-English	3.67	.96	.00
		L1-Chinese	2.30	.84	
	9	L1-English	1.46	.73	0.03
		L1-Chinese	.57	.82	
	10	L1-English	2.80	1.00	.15
		L1-Chinese	2.47	.78	
	11	L1-English	2.50	.63	.00
		L1-Chinese	.90	1.00	
	12	L1-English	2.10	.94	.00

		L1-Chinese	1.00	.95	
Total		L1-English	38.80	4.54	.00
		L1-Chinese	30.37	4.85	

7.10 Research question 8

When reading on print how do the subskill scores differ between L1-English- and L1-Chinese-speakers?

L1-English-speakers obtained significant higher mean scores only for literal-level subskill 3 (Mean English = 4.80, SD= .89, Mean Chinese= 3.83, SD = .99, t(58) =3.99, p < .00). In addition, L1-English participants obtained significantly higher mean scores than the L1-Chinese participants for all inferential-level subskills, from subskill 6 (Mean English = 2.43, SD= 1.27, Mean Chinese = 1.80, SD= .71, t(45.51) = 2.37, p = .02) to subskill 12 (Mean English = 3.67, SD = 1.21, Mean Chinese = 2.40, SD= 1.10, t(58) = 4.23 , p < .00).

Table 7.12 Scores for subskills of between L1-English and L1-Chinese (Print)

Types of subskills	Subskills	Participants	Mean	SD	Sig.
Literal	1	L1-English	4.73	1.17	.48
		L1-Chinese	4.50	1.36	
	2	L1-English	4.87	.94	.10
		L1-Chinese	4.17	1.05	
	3	L1-English	4.80	.89	.00
		L1-Chinese	3.83	.99	

	4	L1-English	3.30	.95	.63
		L1-Chinese	3.20	.98	
	5	L1-English	4.20	1.27	.05
		L1-Chinese	3.57	1.33	
Inferential	6	L1-English	2.43	1.28	.02
		L1-Chinese	1.80	0.72	
	7	L1-English	3.63	0.62	.00
		L1-Chinese	2.10	0.61	
	8	L1-English	3.67	0.96	.00
		L1-Chinese	2.30	0.84	
	9	L1-English	2.27	.64	.03
		L1-Chinese	1.90	.66	
	10	L1-English	3.87	.90	.00
		L1-Chinese	2.57	.82	
	11	L1-English	3.60	.97	.00
		L1-Chinese	2.10	2.77	
	12	L1-English	3.67	1.21	.00
		L1-Chinese	2.40	1.10	
Total		L1-English	45.03	3.33	.00
		L1-Chinese	34.30	2.77	

7.11 Section summary

Overall, L1-English participants obtained higher scores on both screen and print than L1-Chinese participants. Specifically, L1-English speakers obtained significantly higher scores than L1-Chinese participants on the literal-meaning subskill 3 (recognition or recalling comparison or cause and effect). This effect was also found for subskill 5 (summarizing) but only when reading on screen. With the exception of subskill 10 (Judging facts and opinions), L1- English participants performed better on all inferential-level subskills on both print and screen.

7.12 Conclusion

This chapter presented the results for the research questions 1 – 8 of the RCA. The key findings are:

RQ 1- 3

- The medium of text presentation had a significant effect on reading comprehension performances. When looking at L1-Englslish- and L1-Chinese-speaking participants together, they tended to achieve better results when reading on print than reading on screen.
- There was no significant difference in the scores between the screen and print text presentation for the literal-meaning level of comprehension. However, a significant difference was apparent in the inferential-meaning level of comprehension, with higher scores obtained from print than screen.
- When looking at L1-Englslish- and L1-Chinese-speaking participants separately, the above pattern was consistent.

RQ 4 – 6

- When looking at L1-English and L1-Chinese-speaking participants together, the general pattern was that significant better performances in print than in screen were found in the following literal-meaning subskills:
  subskill 3 – recognition/recalling comparison or cause and effect;
  subskill 5 – summarizing;

and the following inferential subskills:
subskill 8 – inferring cause and effect
subskill 9 – predicting outcomes
subskill 10 – judging facts and opinions
subskill 11 – judging strengths and weakness of a position
subskill 12 – judging adequacy and validity.

- When looking at L1-English-speaking participants only, the text presentation had no significant effect on all the literal-meaning subskills; but the effect was consistent with the above general pattern on the inferential-meaning subskill 8, 9, 10, 11 and 12.
- When looking at L1-Chinese-speaking participants only, the effect was consistent with the general pattern on the literal-meaning and inferential-meaning subskills but with an exception on subskill 10.
- Surprisingly, better performance was found on screen than print on the inferential meaning *subskill 6 – inferring supporting details* for both L1-English- and L1-Chinese-speaking participants.

RQ 7 – 8

- Overall, L1-English-speaking participants obtained higher scores on both screen and print than L1-Chinese-speaking participants.
- L1-English speakers obtained significantly higher scores than L1-Chinese participants on the literal-meaning *subskill 3* – recognition or recalling comparison or cause and effect. This effect was also found for *subskill 5 – summarizing* but only when reading on screen.
- L1- English-speaking participants performed better on all the inferential-level subskills on both print and screen. But the effect on *subskill 10* was only found on screen.

Chapter 8 Part Two: Working Memory Capacity Test

8.1 Introduction

Chapter 6 and 7 have presented the methods and results of the first empirical part – Reading Comprehension Assessment (RCA). This chapter moves to the second empirical part – Working Memory Capacity Test (WMCT). In this chapter, it starts with the research purpose (8.2) of the WMCT and its relation with other parts of the overall design, followed by the research methods (8.3) comprising the participants (8.3.1), the design and procedures (8.3.2), the implementation scoring scale of different WMC measures (8.3.3). Next section moves to the data analysis (8.4), explaining what statistical tests are chosen and the rationales. Following this, results (8.5) will be presented in two parts: descriptive results for the WMC measures (8.5.1) and the answers regarding three research questions (8.5.2 to 8.5.4).

8.2 Purposes

The main purpose of the WMCT was to explore what factors, from the perspective of the working memory system, account for the differences of comprehension performances from print and screen that have been found in previous comprehension test. It should be noted that that the assumptions of working memory system of this thesis is principally on a basis of Baddeley and Hitch's (2000) Multi-component Model of working memory (see Section 4.2). Guided by this model, the WMCT aims to measure the WM capacity in different aspects – phonological short-term memory (verbal WM), complex working memory and visuospatial short-term memory (non-verbal WM), and thus it consists of three memory capacity span tests – Forward Digit Span (FDS), Backward Digit Span (BDS) and Corsi-block span task (Corsi). In order to further explore the relationship between the performances from the RCA and the capacity measured by the WMCT, next section (8.3) will describe the research methodology that has informed the WMCT and provide a justification for its choice.

8.3 Methodology of WMCT

8.3.1 Participants

The final sample of WMCT included 120 participants with 60 L1-English-speaking and L1-Chinese-speaking students. The recruitment process comprised two stages: 1. During September to December 2018, 109 of 112 who previously took the RCT joined the WMCT, as three participants dropped out; 2. During January to March 2019, with the ‘snowball strategy’ 11 first-year undergraduate students (6 L1-English-speaking and 5 L1-Chinese-speaking) took the RCT and WMCT individually under the administration of the researcher. At the end of a series of tests, each participant received a bag of sweets and snacks and a food voucher.

8.3.2 Design

Taking a similar approach to that used in Waterman et. al (2017), a forward digit recall (FDR) task, a backward digit recall (BDR) task and a Corsi-block task (Corsi) were used to measure participants’ working memory capacity. This study used a combination of verbal and non-verbal, simple and complex working memory measures, rather than a pure reading span task (RST), for two considerations. First, as the existing literature on working memory differs from one model to another, this study assumes the existence of different domains and loads within the working memory system: the FDR and Corsi-block aimed to tap simple verbal/phonological and visuospatial sketchpad domains (linked to storage); the BDR aimed to tap complex working memory (linked to processing). Storage and processing are two primary functions controlled by the central executive within the working memory (Baddeley, 2012). Second, the present study needed to use non-reading-based measurements in an attempt to provide domain-independent evidence of working memory to understand reading process across different modes of text presentation. A considerable amount of research has relied on the RST when examining the contribution of working memory to reading comprehension, either in L1 or L2 (e.g., Alptekin & Erçetin, 2009, 2011; Harrington & Sawyer, 1992; Leeser, 2007; Waters & Caplan, 1996), however, one possible issue of using pure RST is probably having measured overlapping construct – reading ability. Thus, it could be problematic to determine that the correlation between the WMC and reading comprehension performances is due to the capacity of WM or the proficiency of reading.

8.3.3 Procedure

Before implementing the working memory tasks, the instructions for each task were explained. The tasks were processing in a non-computerised manner as the following sequences: a forward digit span task (FDR); a backward digit span task (BDR) and Corsi-block task (Corsi). A set of tasks lasted for about 15 - 20 minutes for each participants.

Note that the tasks were conducted in participants' first language – English for L1-English participants and Chinese for L1-Chinese participant in order to obtain their actual WM capacity. Although previous research tends to support the view that working memory is language-dependent (Osaka, Osaka and Groner, 1993; Osaka and Osaka, 1992), this study reflected concerns that different pronunciations in English and Chinese syllables might affect response time and accuracy for Chinese participants (discussed in Section 4.4). If Chinese participants use both L1 and L2 in the test, they may consume excessive cognitive resources and feel fatigue which could have an effect on their test performances.

8.3.4 Working memory measures

FDR.

The Forward digit span (Alloway, 2007) was used to assess the capacity of phonological storage. The experimenter first read aloud a string of digits, and the participant was required to repeat the digits as the same order which they were presented. For example, if the experimenter said "4, 3, 6, 9", the participant had to repeat "4, 3, 6, 9". When the participant repeated, the experimenter kept records on the scoresheet. The internal consistency reliability for this task was calculated as .78.

BDR.

The backward digit span is designed with dual-task (storage and processing) paradigm and used to assess the capacity of complex working memory. The procedures were the same as FDR but the participant had to repeat the digits in reverse order. For example, if the experimenter said “4, 3, 6, 9”, the participant had to repeat as “9, 6, 3, 4”. The internal consistency reliability for this task was calculated as .75.

Corsi-block.

The Corsi-block task is used to assess the visuospatial short-term memory. A board was placed on the table in front of the participant on which nine blocks were randomly spaced. The Duplo (a type of Lego) was used to represent the board and blocks. The experimenter tapped out spatial sequences on the blocks and the participant had to repeat the spatial pattern immediately after presentation. The internal consistency reliability for this task was calculated as .73.

Scoring

The WMCT used a partial-credit unit scoring for each span task, which expressed “the mean proportion of elements within an item that were recalled correctly” (Conway et al, 2005, p.775). For each task, a span methodology was used where the number of to-be-remembered items starting with two and increasing to nine. At each length there were three sets of sequences. Partial-credit unit scoring gave equal credit to partly correct items. For example, if the participant answered three out of three items correctly in a sequence, 1 point would be give; if the participant answered four out of five items correctly, 4/5 points would be given.

This present task did not take the traditional absolute scoring of WM span task (e.g., Daneman & Carpenter, 1980) for reasons connected with reliability. The traditional scoring is as follows: the task begins with an item consisting of two elements and continues until the participant’s accuracy falls below a threshold; once the participant reaches the threshold, the test is stopped and the span score is the last item size recalled with a specific probability (e.g. five of six items). This scoring is commonly used in previous reading and WM research (Alptekin and Ercetin, 2010), however, a problem with these absolute scoring methods, pointed by Conway et al. (2005), is the difficulty of a span item may be vary on different dimensions, thus threatening span reliability.

The composite score of working memory capacity was obtained by converting FDS, BDS and Corsi scores to z-scores and taking their average, as suggested by Waters and Caplan (1996). Z-score is a standardized score which indicates how far away a data value is from the mean but also enables us to compare the scores from different group of data. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean. If a Z-score is 0, it indicates that the data value is identical to the mean score. The present study transformed the raw scores of three different working memory capacity tests into standard scores (z-score) and used the composite score (mean of z-score) for further correlational and regression analysis in SPSS.

8.4 Data analysis

The data analysis has two stages. The first is the descriptive analysis for each span task. The second is correlational analysis and regression analysis with regards to the three research questions of the WMCT that this study posed previously. The descriptive analysis was done directly in SPSS and results are presented in section 8.5. Prior to this, next two sections (8.4.1 and 8.4.2) explains the statistical tests that were used to answer the research questions.

8.4.1 What statistical test to choose?

Research question 1

1a. Is there a relationship between working memory capacity and literal reading comprehension across different text presentation?

1b. Is there a relationship between working memory capacity and inferential reading comprehension across different text presentation?

Research question 1a and 1b aimed to test the relationship between working memory capacity and literal or inferential reading comprehension. Each participant in the present study had a pair of score, one score for working memory capacity and another for reading comprehension. The score for working memory capacity was regarded as one continuous variable; the score for literal/inferential reading comprehension was regarded as another. Thus, to determine the strength and direction of an association between two continuous variables (measured at the interval or ratio level), the Pearson product-moment correlation was suggested to use for these two questions.

Research question 2

2a. To what extent does working memory capacity predict literal comprehension across different text presentation?

2b. To what extent does working memory capacity predict inferential comprehension across different text presentation?

Research question 2a and 2b aimed to examine how the reading comprehension performance might be affected by the working memory capacity. This needed a further regression analysis on the basis of research question 1a and 1b in that the correlation analysis can only indicate the correlation between two variables but not tell which variable could affect another. This is because the Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable but treats them equally (Weisberg, 2005). That said, it was unknown that whether reading comprehension could be explained or predicted by the working memory capacity or vice versa. Further, Pearson correlation coefficient *r* does not imply the degree of change in one variable affects another variable. For example, the Pearson correlation coefficient of +1 does not mean that for every unit increase in one variable affects a unit increase in another. It simply means no variation between the observed data points and best fit line. For this reason, it was uncertain to what extent the scores for reading comprehension would be affected by the working memory capacity when reading on screen and print respectively. Thus, a simple linear regression was suggested to use in answering research question 2a and 2b. The predictive/independent variable was the composite score for working memory capacity and the outcome/dependent variable was the score for literal/inferential reading comprehension.

Research question 3

3a. To what extent does working memory measured by different tasks (FDS, BDS, and Corsi-block) *predict literal reading comprehension across different text presentation?*

3b. To what extent does working memory measured by different tasks (FDS, BDS, and Corsi-block) *predict inferential reading comprehension across different text presentation?*

Research question 3a and 3b are intended to further explore working memory measured by different tasks (FDR, BDR and Corsi-block). The composite score may mask the individual differences between the effects of working memory measured by FDR, BDR and Corsi-block tasks. Each task would have different degree of impact on reading comprehension performance. To explore which span tasks may have significant effects and how those effects could be different, a multiple regression analysis was needed. The scores for FDR, BDR and Corsi-block were considered as three predictive variables and the scores for literal/inferential reading comprehension was the outcome variable.

8.4.2 Checking assumptions for statistical tests

8.4.2.1 Research question 1 – assumption checking

*1a. Is there a relationship between working memory capacity and **literal** reading comprehension across different text presentation?*

1b. Is there a relationship between working memory capacity and inferential reading comprehension across different text presentation?

Three assumptions related to Pearson's correlation have to be met: A. Linearity; B. Normality; C. No significant outlier.

A. Linearity. This means there is a linear relationship between the variables. In the research question 1a, the variables were WMC composite scores and literal reading comprehension scores from participants respectively under screen and print text presentation. In the research question 1b, the variables were WMC composite scores and inferential reading comprehension scores respectively under screen and print. A visual inspection of the following scatterplots suggested the assumption of linearity was met (Figure 8.1, Figure 8.2, Figure 8.3 and Figure 8.4).

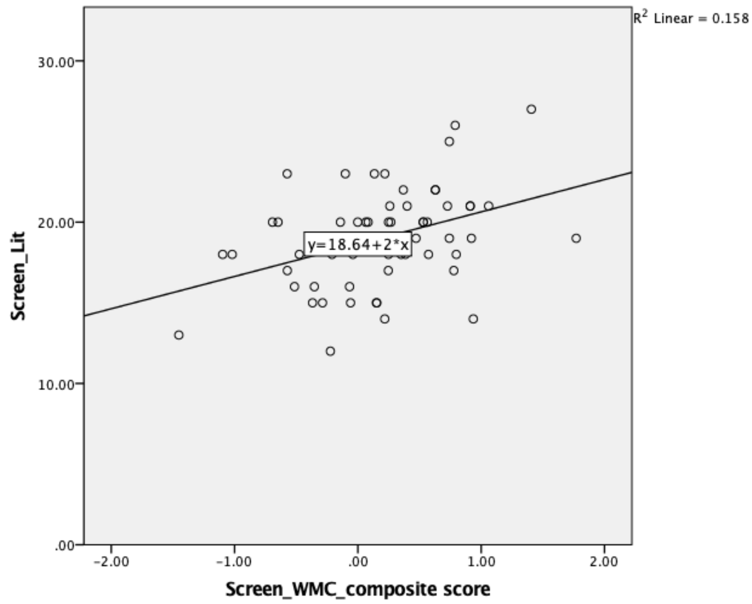


Figure 8.1 Simple scatter of literal reading comprehension scores by WMC composite scores on screen

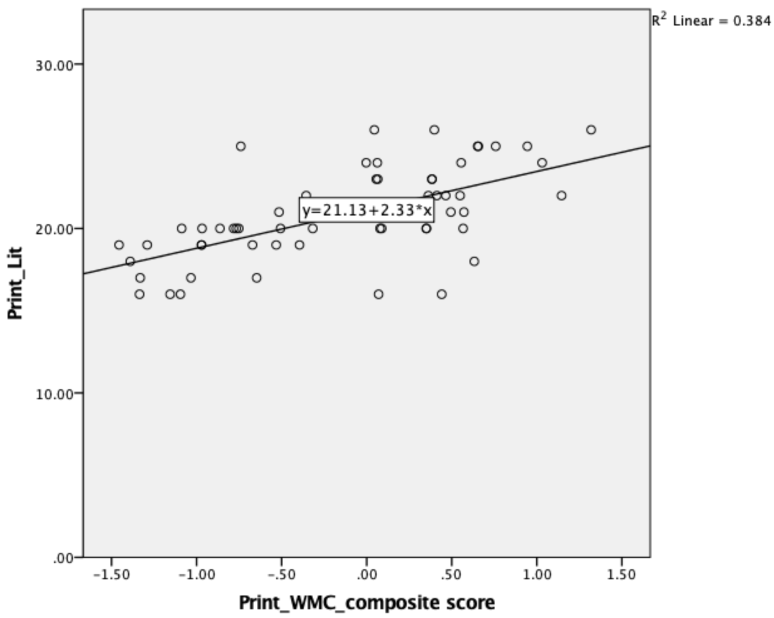


Figure 8.2 Simple scatter of literal reading comprehension scores by WMC composite scores on print

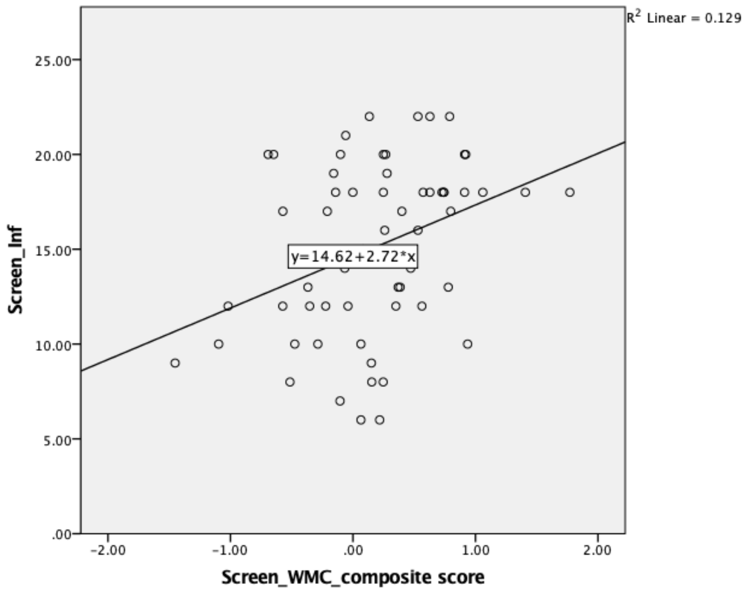


Figure 8.3 Simple scatter of inferential reading comprehension scores by WMC composite scores on screen

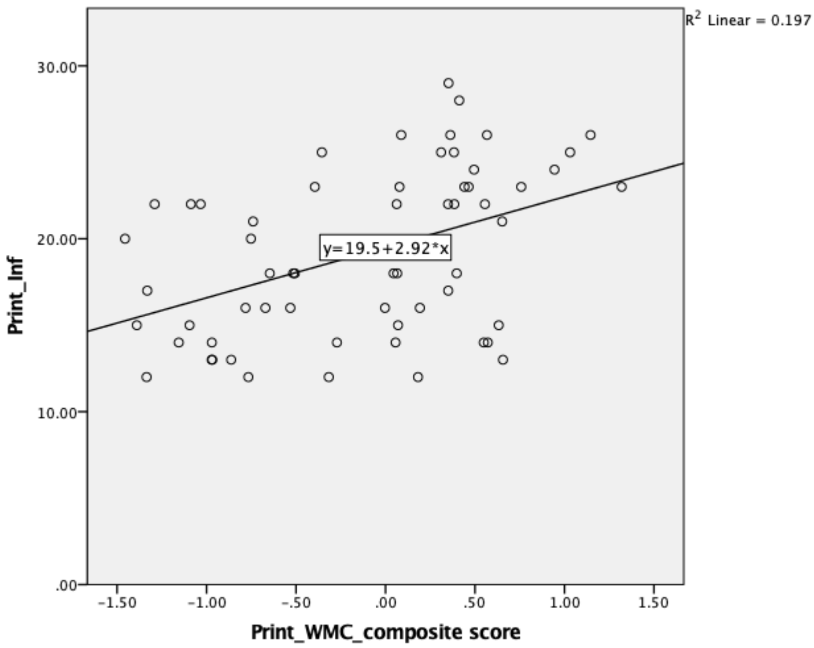


Figure 8.4 Simple scatter of inferential reading comprehension scores by WMC composite scores on print

B. Normality. This means the variables should be normally distributed. With respect to research question 1, the normality of each variable was checked both numerically and graphically. Numerically, the variables of

WMC composite score and literal reading comprehension for participants on screen were normally distributed, as assessed by Shapiro-Wilk's test (P > .05). The variables of WMC composite score and literal reading comprehension for participants on print, the variables of inferential reading comprehension scores for participants on both screen and print were not normally distributed, as assessed by Shapiro-Wilk's test (P < .05). Graphically, a visual inspection of the following Q-Q plot and Histogram (Figure 8.5 to Figure 8.8 in Appendix) indicated all the variables were approximately normally distributed except for the variables of inferential reading comprehension for participants on screen and print. The Q-Q plots and Histogram (Figure 8.9 to Figure 8.10) showed some evidence of skew. Therefore, it is assumed not all the variables in research question 1 were not normally distributed.

In such a situation when the assumption of normality is violated, an alternative way is to run Spearman's run-order correlation which is a non-parametric test. However, given that normality matters only for inferring significance, concerns about the confidence interval can be addressed by using the bootstrap method (Field, 2012). This is because the confidence interval is calculated using a random sampling procedure and thus will be unaffected by the distribution of scores. Therefore, the present study chose to run a Pearson's correlation regardless of the issue of normality.

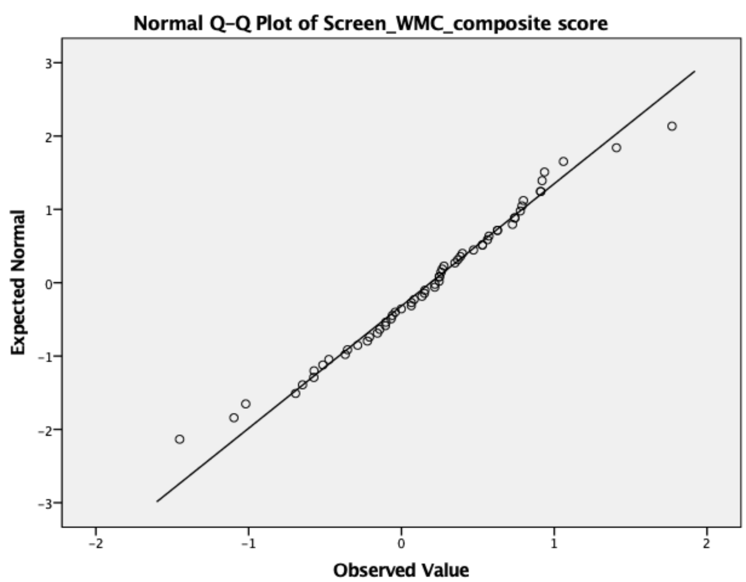


Figure 8.5 Q-Q plot of WMC composite scores for participants on *screen*

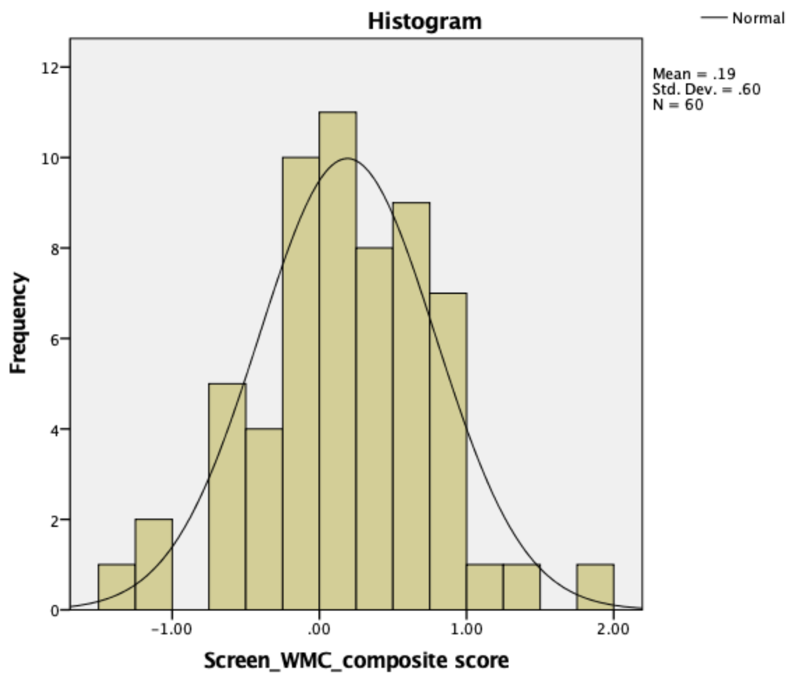


Figure 8.6 Histogram of WMC composite scores for participants on *screen*

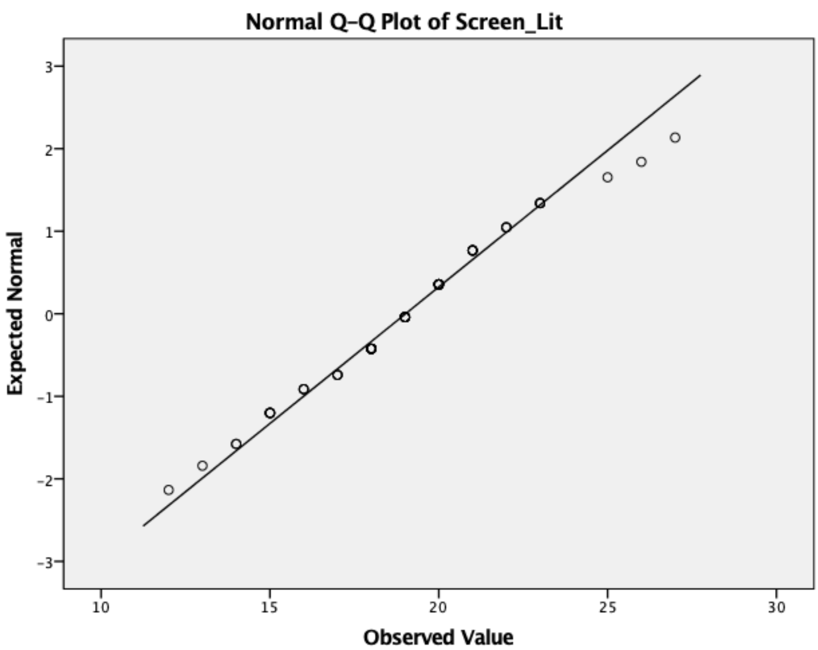


Figure 8.7 Q-Q plot of *literal* reading comprehension scores for participants on *screen*

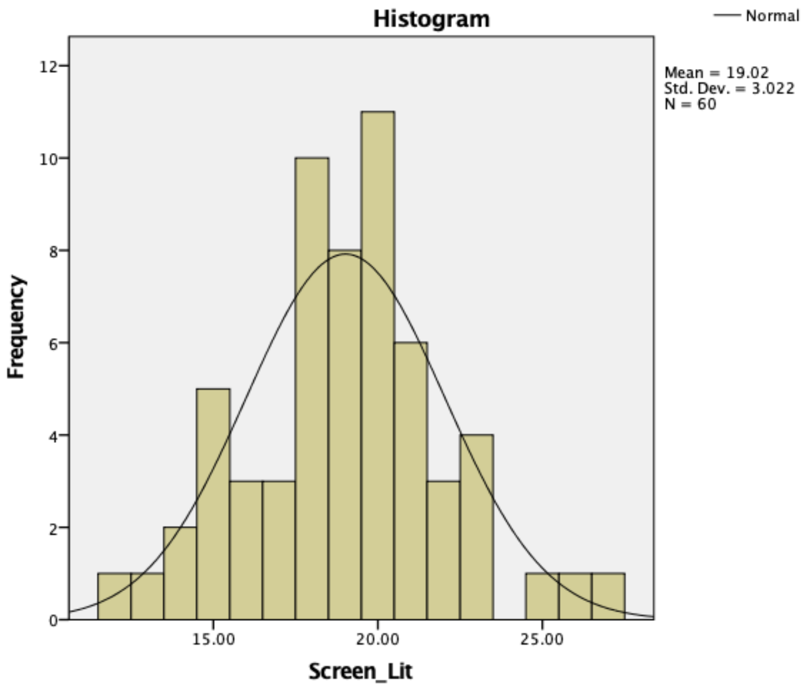


Figure 8.8 Histogram of *literal* reading comprehension scores for participants on *screen*

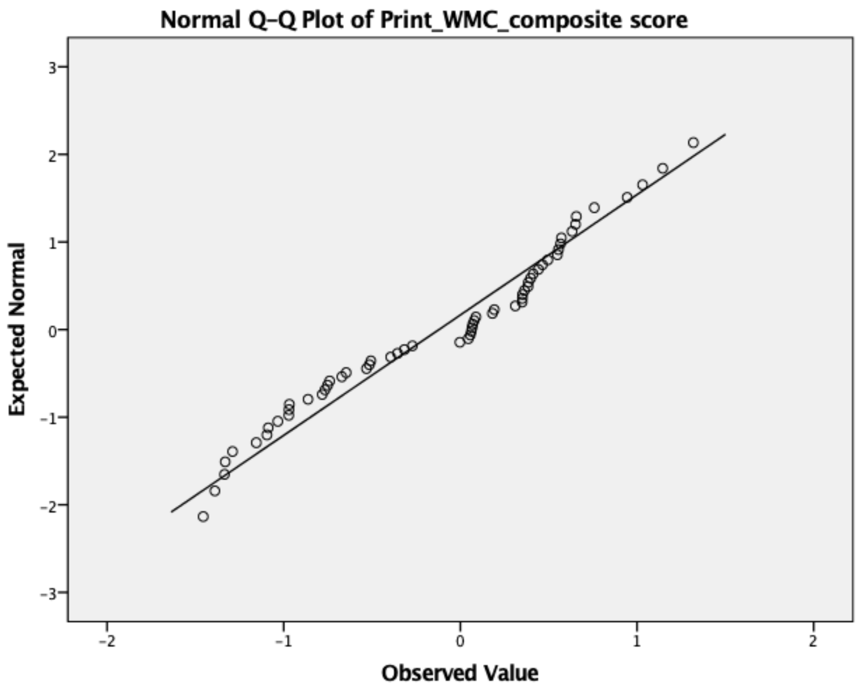


Figure 8.9 Q-Q plot of WMC composite scores for participants on *print*

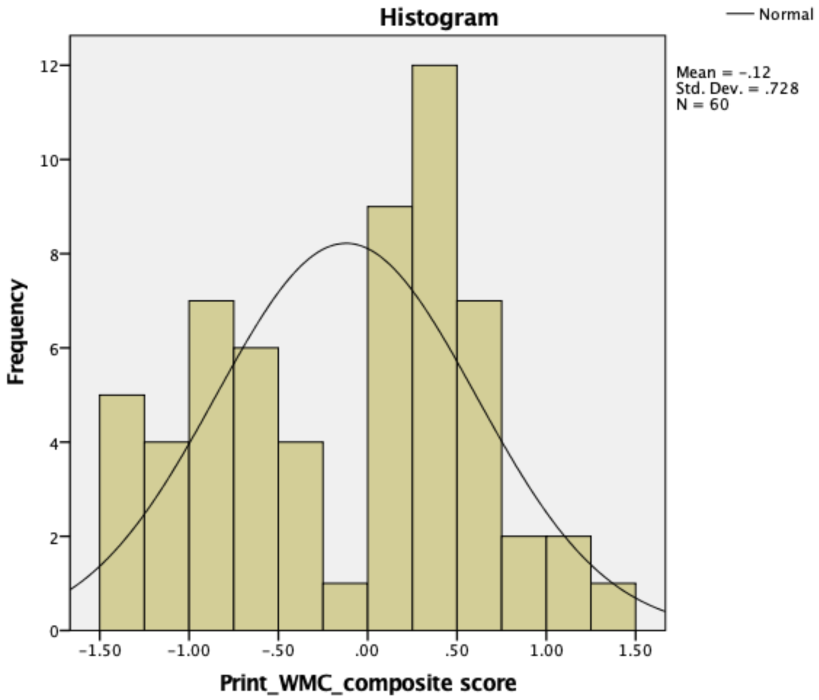


Figure 8.10 Histogram of WMC composite scores for participants on *print*

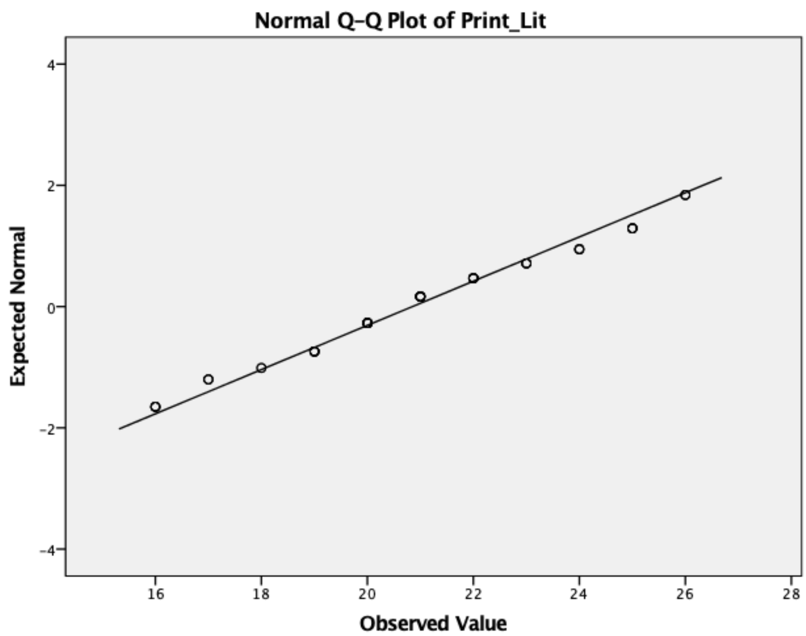


Figure 8.11 Q-Q plot of *literal* reading comprehension scores for participants on *print*

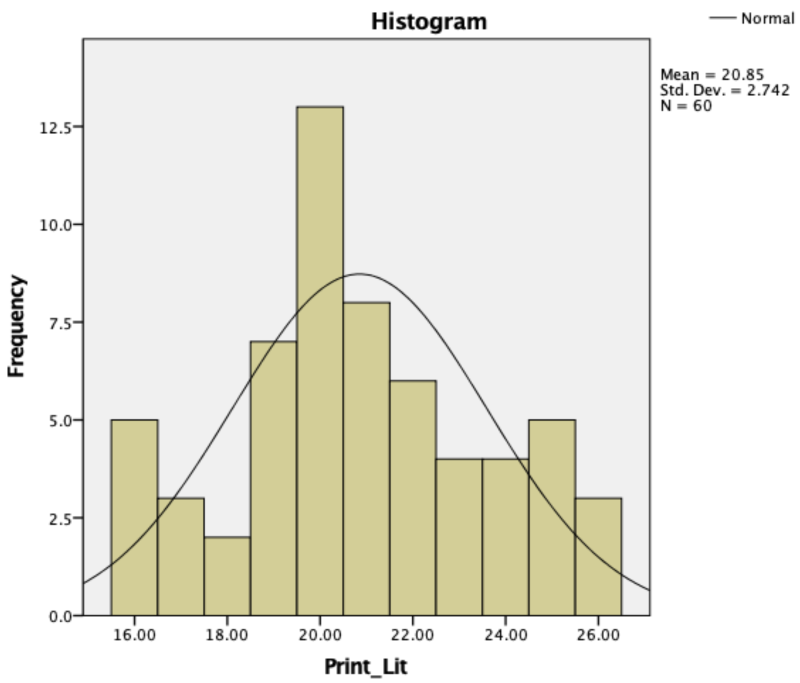


Figure 8.12 Histogram of *literal* reading comprehension scores for participants on *print*

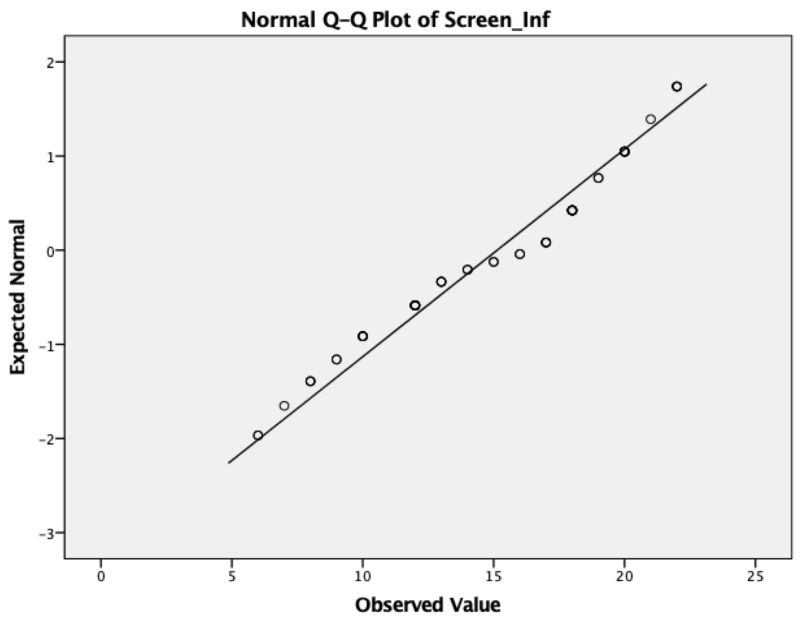


Figure 8.13 Q-Q plot of *inferential* reading comprehension scores for participants on *screen*

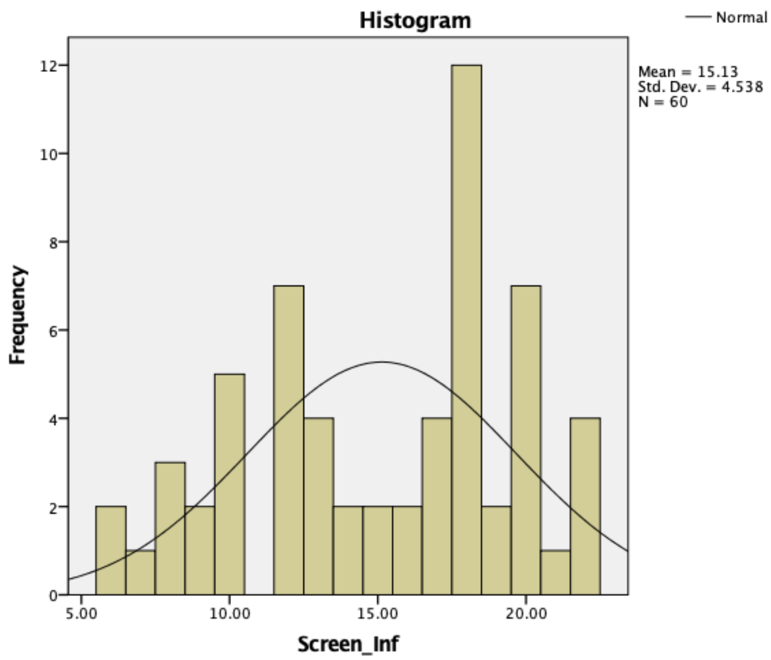


Figure 8.14 Histogram of *inferential* reading comprehension scores for participants on *screen*

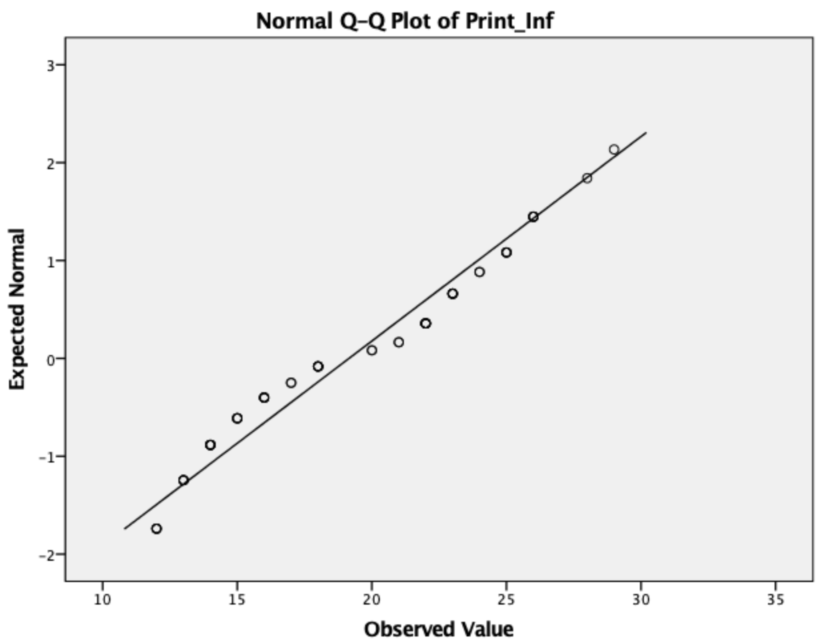


Figure 8.15 Q-Q plot of *inferential* reading comprehension scores for participants on *print*

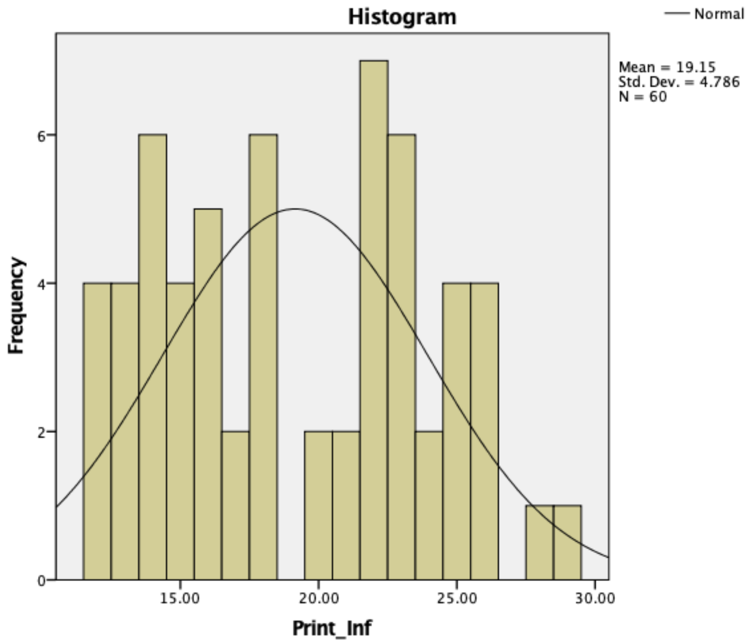


Figure 8.16 Histogram of *inferentia*l reading comprehension scores for participants on *print*

C. No significant outliers. This means running Pearson's correlation coefficient should ensure no outliers or keep them to a minimum because outliers can have an exaggerated influence on the value of r, which leads to Pearson's correlation coefficient not having a value that best represents the data as a whole. The scatterplots above indicated there were no significant outliers in the present study.

8.4.2.2 Research question 2 – assumption checking

2a. To what extent does working memory capacity predict literal reading comprehension across different text presentation?

2b. To what extent does working memory capacity affect inferential reading comprehension across different text presentation?

Five assumptions have to be considered when running simple linear regression. Assumptions A and B (see below) have to be met before running regression analysis; however, assumption C, D and E about impendence, homoscedasticity and normality will be re-examined after analysis when there is a need to correct and generalize the estimated model.

A. Linearity. There needs to be a linear relationship between predictive and outcome variables. The linear relationship between variables was already inspected and established in research question 1 (Figure 8.1 to Figure 8.4).

B. No significant outliers. The scatterplots above (Figure 8.1 to Figure 8.4) indicated no significant outliers.

C. Independent errors. This means the residuals need to be independent i.e. one residual cannot provide any information about any other residual. In the present study design, each participants was independent and unrelated. It is highly unlikely the observed data, such as working memory capacity scores and literal/inferential reading comprehension scores will be related. For this reason, the variables in research question 2 met the assumption of independent errors.

D. Homoscedasticity. This means the residuals at each level of the predictors should have the same variance (Field, 2012, p. 311). This can be checked by inspection of a plot of the unstandardized or standardized residual values against the unstandardized and standardized predicted values (Draper & Smith, 2014, p. 156). This issue was checked after running regression analysis in Section 8.5.2. The standardized residual values plot (Figure 8.17 and 8.18) suggested there was homoscedasticity because the points of the plot exhibited no pattern and were approximately constantly spread.

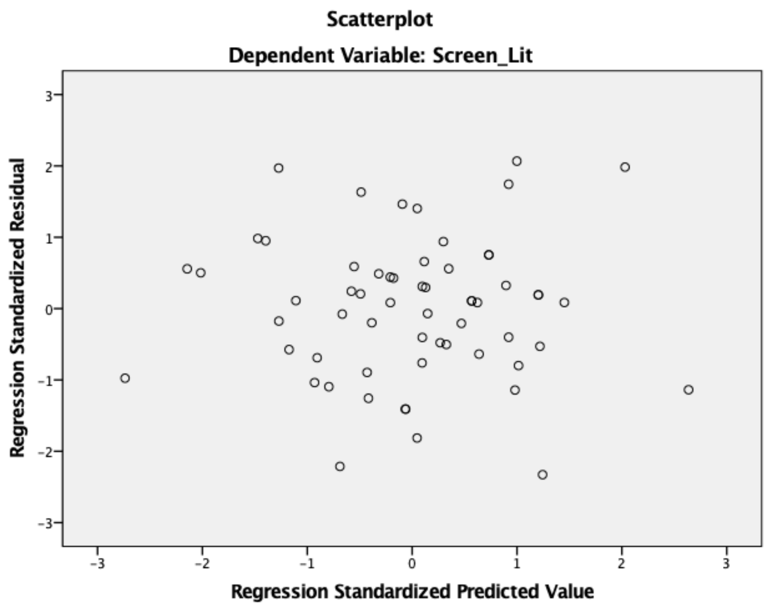


Figure 8.17 Regression Standardized predicted value plot of literal reading comprehension on screen

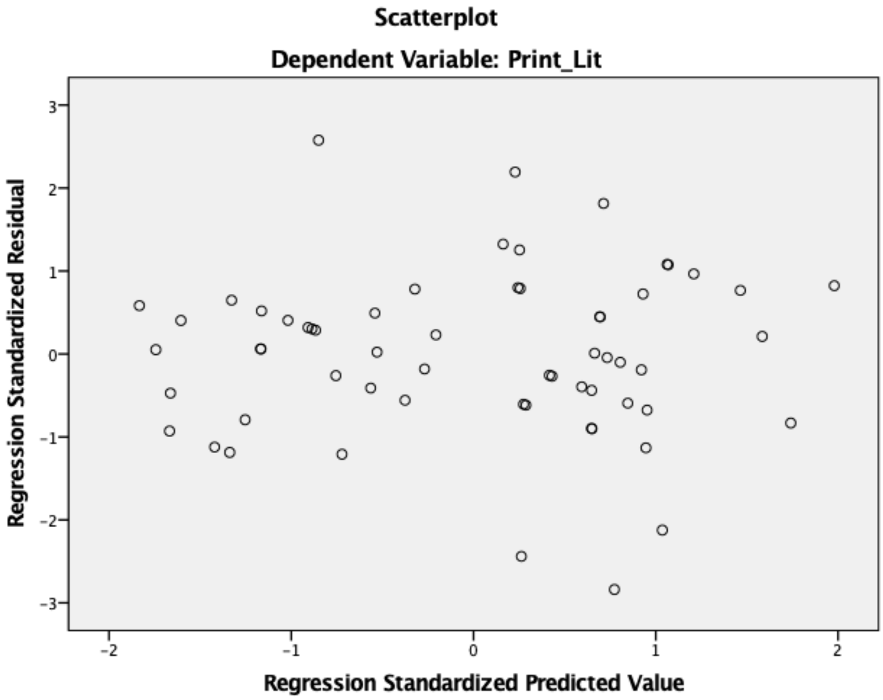


Figure 8.18 Regression Standardized predicted value plot of literal reading comprehension on print

E. Normally distributed errors. The assumption of normal distribution has been discussed in Section 6.7.3, indicating the variable of literal and inferential reading comprehension scores were not normally distributed. However, in regression analysis, predictors do not need to meet this assumption, in particular in large sample size due to central limit theorem (Field, 2012, p. 311). To answer research question 2, it should be concerned about the model parameters (the slope coefficient in simple linear regression) rather than significance test. Therefore, the issue of non-normal distribution will not be problematic in running simple linear regression.

8.4.2.3 Research question 3 – assumption checking

The assumptions of multiple regression analysis for this question were checked after running the procedures of building a regression model in SPSS. Some of the assumptions were checked by inspection of residuals, which can only be obtained once a regression line has been generated. Thus, some assumptions would be re-examined in Section 8.5.4 before reporting the results.

A. Independence of observations. The issue of independence was already discussed in Section 6.7.3, indicating each participant in the current study design was uncorrelated. Further, the value of Durbin-Waston statistics was 1.79, which was very close to 2, meaning no problems with independence.

B. There is a linear relationship between the dependent variable and independent variables collectively but also each of the independent variables. First, the linear relationship between the dependent variable and independent variable collectively was checked using a scatterplot. Second, the linear relationship between the dependent variable and each of the independent variable was checked using partial regression plot.

Take the literal reading comprehension scores on print (dependent variable) and FDR, BDR and Corsi scores (independent variables) as an example. First, in the following Figure 8.19, the residuals almost form a horizontal band, indicating the relationship between literal reading comprehension scores and FDR, BDR and Corsi collectively was likely to be linear. Second, the following Figure 8.20 indicated an approximately linear relationship between the literal reading comprehension scores on print and each score of FDR, BDR and Corsi.

The same procedures would be done to other variables in research question 2 and the issue of linearity would also be reported in Section 8.5.4.

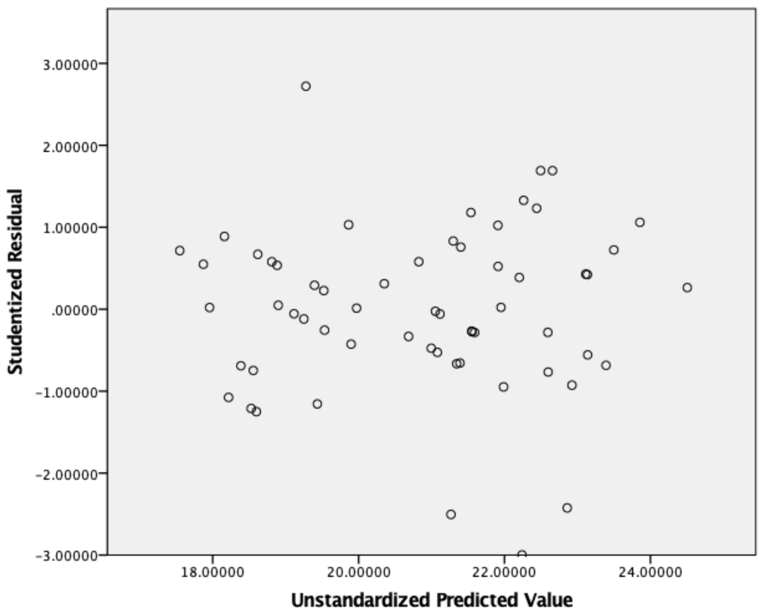


Figure 8.19 Scatterplot of literal reading comprehension scores on print by FDR, BDR and Corsi collectively

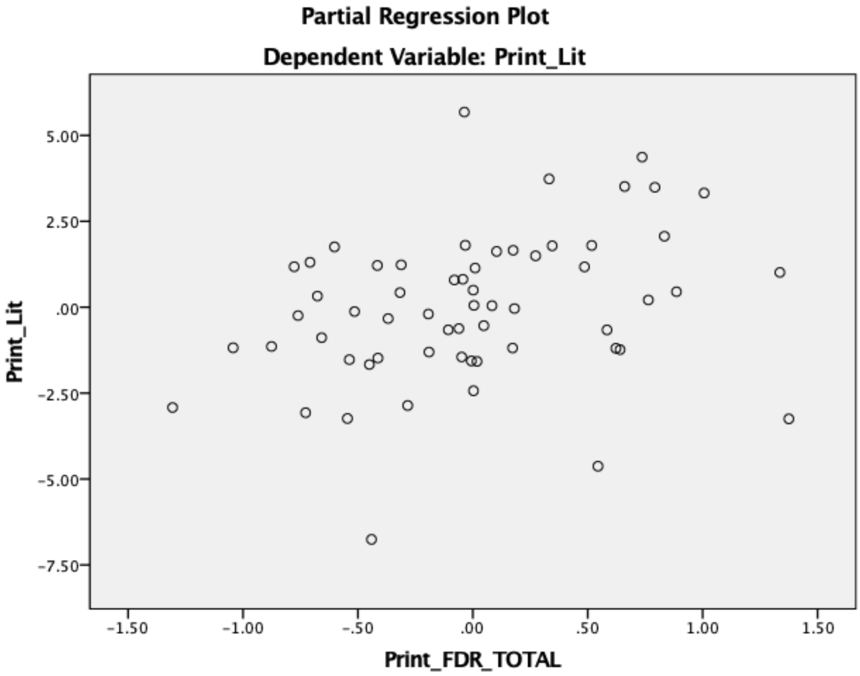


Figure 8.20 Partial regression plot of literal reading comprehension scores on print by each score of FDR

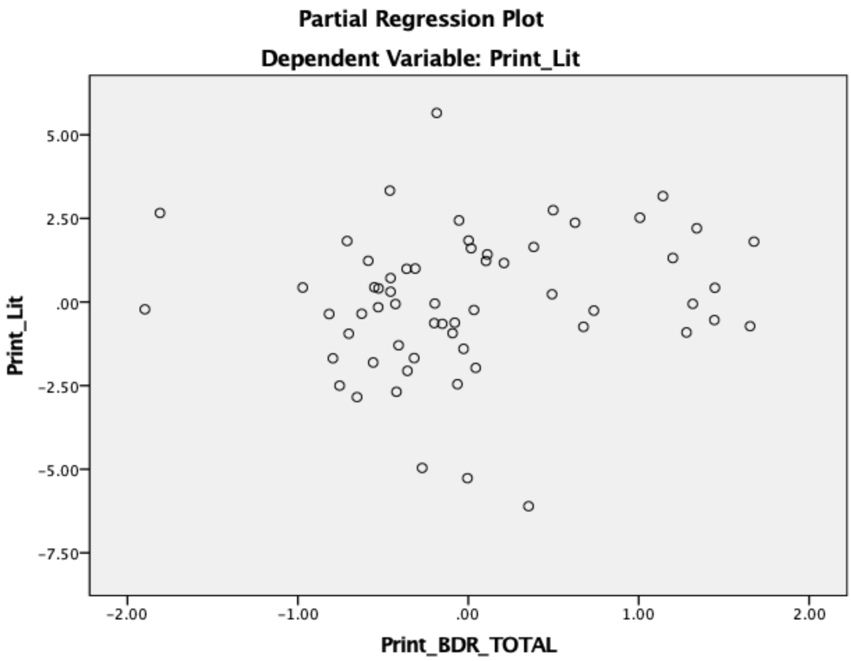


Figure 8.21 Partial regression plot of literal reading comprehension scores on print by each score of BDR

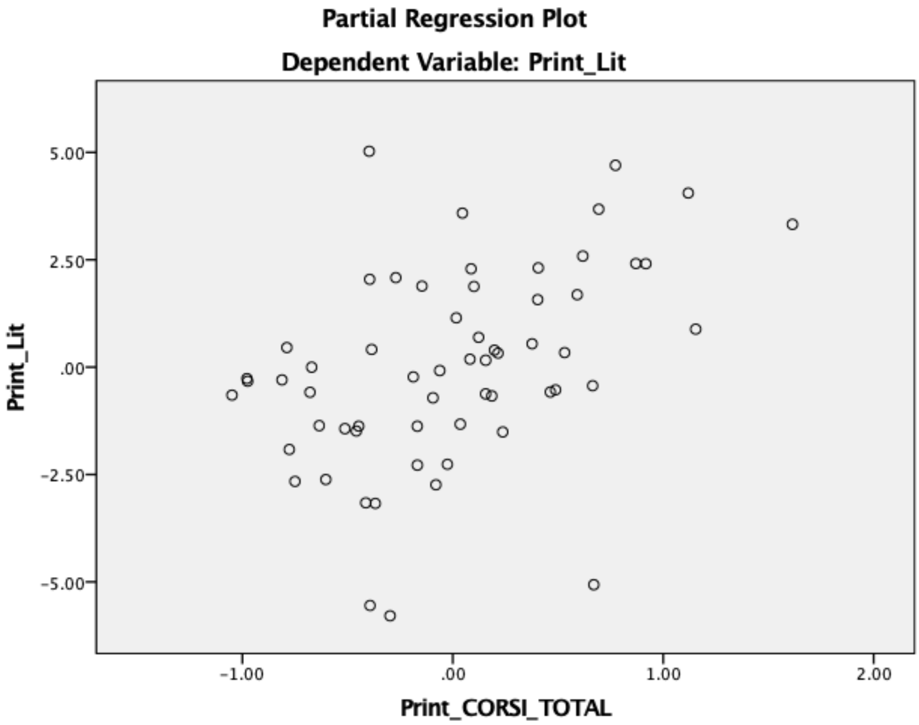


Figure 8.22 Partial regression plot of literal reading comprehension scores on print by each score of Corsi

C. Homoscedasticity. This means the variance is equal for all values of the predicted dependent variable. By looking at the same scatterplot (Figure 8.19), the residuals were evenly spread and exhibited no pattern, suggesting the assumption of homoscedasticity for literal reading comprehension scores on print (dependent variable) and scores on FDR, BDR and Corsi (independent variables) was not violated. The same procedures would be done to other variables in research question 2 and the issue of linearity would also be reported in Section 8.5.4.

8.5 Results

8.5.1 Descriptive statistics

This section gives detail of descriptive statistics for FDR, BDR and Corsi measures and composite scores between screen and print (Table 8.1).

Table 8.1 A summary of descriptive statistics for FDR, BDR and Corsi measures and composite scores between screen and print

		N	M	SD	Min	Max
Screen	FDR	60	18.37	0.57	17.03	19.66
	BDR	60	17.24	0.58	14.54	17.24
	Corsi	60	15.4	0.71	13.87	17.53
	WMC comp.	60	0.12	0.61	-1.45	1.77
Print	FDR	60	18.43	0.65	17.04	19.95
	BDR	60	15.31	0.82	13.7	17.04
	Corsi	60	15.43	0.68	14.03	17.18
	WMC comp.	60	-0.12	0.73	-1.46	1.32

WMC comp: working memory capacity composite score

8.5.2 Research question 1

1a. Is there a relationship between working memory capacity and literal reading comprehension across different text presentation?

A Pearson's product-moment correlation was run to assess the relationship between the working memory capacity and literal reading comprehension across screen and print text presentation. 120 participants were recruited in total with 60 were in each group. Preliminary analysis showed the relationship to be linear with both variables normally distributed (Section 8.4.2), and there were no outliers. With regards to participants under screen text presentation, there was a statistically significant, moderate positive correlation between the literal reading comprehension scores and working memory composite scores ($r = .51$, $p < .00$), the WMC scores measure by FDR ($r = .36$, $p = .01$), by BDR ($r = .36$, $p = .01$) and by Corsi ($r = .32$, $p = .01$). With regards to participants under print text presentation, there was a statistically significant, strong positive correlation between literal reading comprehension scores and working memory composite scores ($r = .62$, $p < .00$) and WMC measured by Corsi $r = .60$, $p < .00$).

Table 8.2 Correlation matrix between literal and inferential scores and WMC measures (Screen)

	Lit-Screen	Lit-Print	FDR	BDR	Corsi	WMC Comp.
Lit-Screen	1	0.53*	0.36*	0.34*	0.32*	0.51*
Inf-Screen		1	0.24	0.56*	0.31*	0.53*
FDR			1	0.6	0.24	0.67*
BDR				1	0.55*	0.34*
Corsi					1	0.75*
WMC Comp.						1

p* < 0.05

Table 8.3 Correlation matrix between literal and inferential scores and WMC measures (Print)

	Lit-Screen	Lit-Print	FDR	BDR	Corsi	WMC Comp.
Lit-Print	1	0.48*	0.36*	0.24*	0.60*	0.62*
Inf-Print		1	0.22	0.75*	0.18	0.44*
FDR			1	0.27	0.45*	0.67*
BDR				1	0.55*	0.78*
Corsi					1	0.60*
WMC Comp.						1

p* < 0.05

Table 8.4 Comparison of correlation of literal and inferential scores and WMC measures between screen and print

	Text presentation	FDR	BDR	Corsi	WMC Comp.
Literal	Screen	0.36*	0.34*	0.32*	0.51*
	Print	0.36*	0.24*	0.60*	0.62*
Inferential	Screen	0.24	0.56*	0.31*	0.53*
	Print	0.22	0.75*	0.18	0.44*

p* < 0.05

1b. Is there a relationship between working memory capacity and inferential reading comprehension across different text presentation?

A Pearson's product-moment correlation was run to assess the relationship between the working memory capacity and inferential reading comprehension across screen and print text presentation. 120 participants were recruited in total with 60 were in each group. Preliminary analysis showed the relationship to be linear but the variable of inferential scores were non-normally distributed. With regards to participants under screen text presentation, working memory composite scores was significantly correlated with inferential reading comprehension scores, r = .53 [.053, .545], p = .014. With regards to participants under print text presentation, working memory composite scores was also significantly correlated with inferential reading comprehension scores, r = .44 [.188, .590], p = .001.

In summary, literal reading comprehension and working memory capacity measured by FDR, BDR and Corsi together were moderately correlated across screen and print text presentation. Inferential reading comprehension was also significantly correlated with working memory capacity measured by FDR, BDR and Corsi together. These results were different from the original hypothesis that the working memory capacity was only correlated with inferential reading comprehension rather than literal comprehension.

8.5.3 Research question 2

2a. To what extent does working memory capacity predict literal reading comprehension across different text presentation?

A linear regression was run to understand the effect of working memory capacity on literal reading comprehension. To assess linearity a scatterplot of working memory capacity composite scores against literal reading comprehension scores across screen and print text presentation was plotted (see section 8.4.2). A visual inspection of these two plots indicated a linear relationship between the independent and dependent variables. There was homoscedasticity and normality of the residuals.

Table 8.5 A summary of simple regression analysis for WMC composite score predicting literal reading comprehension on screen and print text presentation

	Screen					Print				
	R	R^2	B	F	95%CI	R	R^2	B	F	95%CI
WMC Comp	0.50	0.25	2.50	19.49	[1.37, 3.63]	0.62	3.84	2.34	36.15	[1.56, 3.11]

Note. WMC Comp-working memory capacity composite score

With regards to *literal* reading comprehension performed on *screen* (Table 8.5), the prediction equation was: *Screen-literal* reading comprehension = 18.64 + 2.50*WMC. The composite scores of working memory capacities statistically predicted the literal reading comprehension scores, $F(1, 58) = 19.49$, $p = .000$, accounting for 25.0% of the variance in literal reading comprehension with adjusted $R^2 = 23.9\%$, a medium size effect according to Cohen (1988). An extra score of working memory capacity leads to a 2.50 points, 95% CI [1.37, 3.63] increase in literal reading comprehension on screen.

With regards to *literal* reading comprehension performed on *print,* the prediction equation was: *Print-literal* reading comprehension = 21.1 + 2.34*WMC. The composite scores of working memory capacities statistically predicted the literal reading comprehension scores on print, $F(1, 58) = 36.15$, $p = .000$, accounting for 38.4% of the variance in literal reading comprehension with adjusted R^2 = 37.3%, a medium size effect according to Cohen (1988). An extra score of working memory capacity leads to a 2.34 points, 95%CI [1.56, 3.11] increase in literal reading comprehension on print.

In conclusion, working memory capacity composite score significantly predicted literal reading comprehension score on both screen and print text presentation. When comparing the two models, the beta of literal comprehension scores on screen was slightly bigger than print. It is implied that, regarding literal reading comprehension, the working memory capacity may have slightly greater impact when participants read on screen than on print.

2b. To what extent does working memory capacity predict inferential reading comprehension across different text presentation?

A linear regression was run to understand the effect of working memory capacity on inferential reading comprehension. To assess linearity a scatterplot of working memory capacity composite scores against inferential reading comprehension scores across screen and print text presentation was plotted (see section 8.4.2). A visual inspection of these two plots indicated a linear relationship between the independent and dependent variables. There was homoscedasticity and normality of the residuals.

Table 8.6 A summary of simple regression analysis for WMC composite score predicting inferential reading comprehension on screen and print text presentation

	Screen					Print				
	R	R^2	B	F	95%CI	R	R^2	B	F	95%CI
WMC Comp	0.50	0.28	3.94	22.34	[2.28, 5.62]	0.45	0.20	2.92	14.24	[1.37, 4.46]

With regards to *inferential* reading comprehension performed on *screen* (Table 8.6), the prediction equation was: *Screen-inferential* reading comprehension = 14.64 + 3.94*WMC. The composite scores of working memory capacities statistically predicted the inferential reading comprehension scores, F(1, 58) = 22.34, *p* = .000, accounting for 28.0% of the variance in inferential reading comprehension with adjusted R^2 = 26.6%, a medium size effect according to Cohen (1988). An extra score of working memory capacity leads to a 3.94 points, 95% CI [2.28, 5.62] increase in inferential reading comprehension on screen.

With regards to *inferential* reading comprehension performed on *print,* the prediction equation was: *Print-inferential* reading comprehension = 19.5 + 2.92*WMC. The composite scores of working memory capacities statistically predicted the inferential reading comprehension scores on print, F(1, 58) = 14.24, *p* = .000, accounting for 20.0% of the variance in inferential reading comprehension with adjusted R^2 = 18.3%, a medium size effect according to Cohen (1988). An extra score of working memory capacity leads to a 2.92 points, 95%CI [1.37, 4.46] increase in inferential reading comprehension on print.

8.5.4 Research question 3

3a. To what extent does working memory measured by different tasks (FDS, BDS and Corsi-block) predict literal reading comprehension across different text presentation?

A hierarchical multiple regression was run to determine if the addition of FDR and then of FDR and BDR, and finally of FDR, BDR and Corsi working memory capacity measures improved the literal comprehension scores across screen and print text presentation. There was linearity as assessed by partial regression plots and a plot of studentized residuals against the predicted values. There was independence of residuals, as assessed by a Durbin-Watson statistic of 1.62. There was homoscedasticity, as assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values. There was no evidence of multicollinearity, as assessed by tolerance values greater than 0.1. There were no studentized deleted residuals greater than ±3 standard deviations, no leverage values greater than 0.2, and values for Cook's distance above 1. There assumption of normality was met, as assessed by Q-Q Plot.

NB. The assumptions for other regression analysis in this research question (literal reading comprehension on print, inferential reading comprehension on screen and print) were also met.

Table 8.7 Hierarchical multiple regression predicting literal reading comprehension on screen from scores on FDR, BDR and Corsi

Literal reading comprehension scores on screen									
	Step 1			Step 2			Step 3		
Variable	B	β	p	B	β	p	B	β	p
Constant	19.08		.000	18.51		.000	18.58		.000
FDR	1.15	.36	.005	1.09	0.34	.006	.94	.30	.017
BDR				1.3	0.32	.009	1.14	.28	.019
Corsi							.59	.20	.107
R^2	.13			0.23			.26		
F	8.40			8.35			6.62		
ΔR^2	.13			.10			.04		
ΔF	8.40		.005	7.37		.009	2.69		.107

Note. N=60, *p< .05, B-unstandardized coefficient, β-standardized coefficient, FDR-forward digit recall, BDR-backward digit recall, Corsi-block test, ΔF-the increase of F, ΔR^2-the increase of R^2

Regarding literal reading comprehension scores on screen (see Table 8.8), the model of step 1 (addition of FDR) was significant, R^2 = .13 , F(1, 58) = 8.40 , *p* = .005. The addition of BDR (step 2) made a significant contribution , ΔR^2 = .10, F(2, 57) = 8.35, with FDR a predictor (β = 1.09, p = .006) and BDR another predictor (β = 1.30, p = .009) of literal reading comprehension on screen. However, the step 3 with addition of Corsi led to non-statistically significant increase in R^2 of .04, F(3, 56) = 6.62 , *p* > .05. In short, for literal reading comprehension scores on screen, the scores of FDR and BDR working memory measures were significant predictors but Corsi-block measures did not make significant contribution. In addition, the standardized beta value for FDR and BDR were virtually identical (.34 and .32 respectively) indicating that both variables had a comparable degree of importance in the scores of literal reading comprehension on screen.

Table 8.8 Hierarchical multiple regression predicting literal reading comprehension on print from scores on FDR, BDR and Corsi

	Literal reading comprehension scores on print								
	Step 1			Step 2			Step 3		
Variable	B	β	*p*	B	β	*p*	B	β	*p*
Constant	20.78		.000	21.05		.000	20.91		.000
FDR	1.24	.48	.000	1.22	.48	.000	.72	.28	.018
BDR				.60	.23	.047	.30	.11	.298
Corsi							1.22	.44	.001
R^2	.23			.29			.42		
F	17.64			11.35			62.62		
ΔR^2	.23			.05			.14		
ΔF	17.64	.48	.000	4.11		.047	13.46		.001

Note. N=60, B-unstandardized coefficient, β-standardized coefficient, FDR-forward digit recall, BDR-backward digit recall, Corsi-corsi-block test, ΔF-the increase of F, ΔR^2-the increase of R^2

Regarding literal reading comprehension scores on print (see Table 5.5.3), the model of step 1 (addition of FDR) was significant, R^2 = .23 , F(1, 58) = 17.64 , *p* = .000. The addition of BDR (step 2) made a significant contribution , ΔR^2 = .05, F(2, 57) = 11.35, p = .047 (note that this was only marginally significant). The step 3 with addition of Corsi-block led to a statistically significant increase in R^2 of .14, F(3, 56) = 62.62, p = .001. The final model (step 3) of FDR, BDR and Corsi-block was not all statistically significant, R^2 = .42, with FDR a predictor (β = .28, p = .000) and Corsi-block another predictor (β = .44, p = .001), however, BDR was not a significant contributor to the (β = .11, p > .05) literal reading comprehension scores on screen. In short, only the FDR and Corsi-block working memory measures uniquely predicted literal reading comprehension performance on print. In addition, the impact of Corsi-block working memory measure was greater than that of FDR measure for literal reading comprehension on print.

3b. To what extent does working memory measured by different tasks (FDS, BDS and Corsi-block) predict inferential reading comprehension across different text presentation?

A hierarchical multiple regression was run to determine if the addition of FDR and then of FDR and BDR, and finally of FDR, BDR and Corsi-block working memory capacity measures improved the inferential comprehension scores across screen and print text presentation.

With regards to inferential reading comprehension scores on screen (see Table 8.9), the model of step 1 (addition of FDR) was non-significant, R^2 = .06 , F(1, 58) = 3.59 , *p* > .05. The addition of BDR (step 2) made a significant contribution, ΔR^2 = .30, F(2, 57) = 16.18, p = .000. The step 3 with addition of Corsi-block led to a non-statistically significant increase in R^2 of .03, F(3, 56) = 11.77, p > .05. Therefore, the variable of BDR was the only significant predictor was BDR working memory measure, β = .55, p = .000, which played a vital important role in the inferential reading comprehension on screen. However, the other two variables FDR and Corsi-block were not significant contributors to the inferential reading comprehension scores on screen.

Table 8.9 Hierarchical multiple regression predicting inferential reading comprehension on screen from scores on FDR, BDR and Corsi

	Literal reading comprehension scores on screen								
	Step 1			Step 2			Step 3		
Variable	B	β	*p*	B	β	*p*	B	β	*p*
Constant	15.19		.000	13.71		.000	13.79		.000
FDR	1.18	.24	.063	1.01	0.21	.056	.82	.17	.012
BDR				3.36	0.55	.000	3.18	.52	.000
Corsi							.73	.16	.140
R^2	.06			.36			.39		
F	3.59			16.18			11.77		
ΔR^2	.06			.30			0.03		
ΔF	3.59		.063	27.15		.000	2.25		.140

Note. N=60, B-unstandardized coefficient, β-standardized coefficient, FDR-forward digit recall, BDR-backward digit recall, Corsi-block test, ΔF-the increase of F, ΔR^2-the increase of R^2

Table 8.10 Hierarchical multiple regression predicting inferential reading comprehension on print from scores on FDR, BDR and Corsi

	Inferential reading comprehension scores on screen								
	Step 1			Step 2			Step 3		
Variable	B	β	*p*	B	β	*p*	B	β	*p*
Constant	19.15		.00	20.67		.00	20.70		.00
FDR	.10	.02	.87	.01	.00	.98	.10	.02	.84
BDR				3.46	.75	.00	3.52	.76	.00
Corsi							- .21	-.04	.68
R^2	.00			.56			.56		
F	.03			36.16			23.77		
ΔR^2	.00			.56			.00		
ΔF	.03		.87	72.14		.00	.17		.68

Note. N=60, B-unstandardized coefficient, β-standardized coefficient, FDR-forward digit recall, BDR-backward digit recall, Corsi-block test, ΔF-the increase of F, ΔR^2-the increase of R^2

With regards to inferential reading comprehension scores on print (see Table 8.10), the model of step 1 (addition of FDR) was non-significant, R^2 = .00 , F(1, 58) = .03 , *p* > .05. The addition of BDR (step 2) made a significant contribution , ΔR^2 = .56, F(2, 57) = 36.16, p = .000. The step 3 with addition of Corsi-block led to a non-statistically significant increase, F(3, 56) = 23.77, p > .05. Therefore, the variable of BDR was the only significant predictor to inferential reading comprehension scores on print, β = .75, p = .00. This showed the same trend as inferential reading comprehension on screen, indicating complex working memory tapped by the measure of BDR had an important impact on inferential reading comprehension on print text presentation.

8.5.5 Conclusion

In conclusion, multiple regression was run to examine how different working memory tasks predict literal and inferential reading comprehension across screen and print text presentation. A hierarchical multiple regression method was used to determine if the measure of FDR, BDR, Corsi-block could predict the literal/inferential comprehension scores respectively.

When participants performed the RCA on screen, the working memory measures of FDR and BDR significantly predicted the literal reading comprehension scores with virtually equal important contributions; the Corsi-block measure however made no significant prediction. This suggests that both storage (tapped by FDR) and processing (tapped by BDR) aspects of working memory played important role in literal reading comprehension when reading on screen.

The Corsi-block measure did however make a significant prediction to literal reading comprehension scores in the print reading condition. The effect of Corsi-block performance was even greater than the measure of FDR. Further, the measure of BDR had a marginally significant effect on literal reading comprehension. This indicates that, first, answering literal reading comprehension questions seems to be more dependent on simple working memory (mainly responsible maintaining information) than on complex working memory (responsible for both storage and process information). Second, reading on print text presentation for literal comprehension questions could make better use of working memory related to visual-spatial domain (tapped by Corsi-block) than on screen.

Lastly, the inferential reading comprehension across screen and print could be only predicted by the working memory measure of BDR. That means, when performing inferential reading comprehension questions neither on screen or on print, complex working memory came into effect and thus participants not only had to maintain (storage) but also manipulate (processing) information.

Chapter 9 Part Three: Interview

9.1 Introduction

The aim of the interview is to provide confirmatory and supplementary qualitative data for the previous quantitative results. After the RCA and the WMCT, follow-up interviews of eight participants (four L1-English speakers and four L1-Chinese speakers) were conducted in order to gain in-depth understanding of how the participants read the academic text on screen and print and provide some explanations for the differences. The aim of the interview was addressed by two research questions:

1. What are the reading process and strategy that the participants used during the RCA in the different subskills?

2. What are the factors contributing to participants' screen or print reading preferences?

9.2 Methods

9.2.1 Participants

Eight participants, four English speaking and four Chinese speaking students, were recruited based on their responses at the begin of the RCA. Among the four Chinese interviewees, Lin and Zhao were from the participants completing the reading test on print; Yu and Wang were from those reading on screen. Similarly, among the four English interviewees, Emma and Eric were from the participants reading print; Mia and Mike were from those reading on screen. Pseudonyms were used to refer to them. All the participants were first year undergraduate students studying in Business.

Chinese interviewees, Lin and Yu, had been studying and living in the UK for two months; Zhao had attended a 10 weeks pre-sessional academic language course and Wang had completed a one-year foundation study at the university language centre before they officially started the university course. In addition, Wang spent two-years studying at high school in the UK before he went to university. All of the Chinese participants had studied English for more than ten years and their English reading proficiency (indicated by IELTS score in reading section) reached the university's criteria (minimum of 6) for non-English-speakers. The Chinese participants reported their reading time for academic materials ranged from 10 to 12 hours on paper and the time spent on screen for academic materials was much less than on paper (below 5 hours). Details are shown below.

The English interviewees all began their 1st year undergraduate study two months before the test. Emma, Eric and Mike's first language were English only. Mia was born in a bilingual family and grew up in the UK so her first language was English mixed with Spanish. The four English participants spent a similar amount of time reading on printed English academic materials (eight to eleven hours) and on-screen materials (less than 5 hours) as their Chinese counterparts. Details are shown below.

Table 9.1 Background information of four L1-Chinese-speaking interviewees

	Lin	Yu	Zhao	Wang
Age	18	19	19	20
First language	Chinese	Chinese	Chinese	Chinese
Level of study in a UK university	1st year	1st year	1st year	1st year
Length of study in a UK university	2 months	2 months	4.5 months	14 months
Length of study in an English-speaking country	2 months	2 months	4.5 months	3 yrs and 2 months
Length of English learning	12 yrs	10 yrs	11 yrs	10 yrs
Hours of English academic reading on print per week	10	11	12	10
Hours of English academic reading on screen per week	2	0.5	2	5

English reading proficiency (IELTS reading score)	6	6	6	7
Self-assessment of L1 (Chinese)reading proficiency	good	very good	good	good

Table 9.2 Background information of four L1-English-speaking interviewees

	Emma	Mia	Eric	Mike
Age	18	19	19	19
First language	English	English	English	English
Level of study in a UK university	1st year	1st year	1st year	1st year
Length of study in a UK university	2 months	2 months	2 months	2 months
Length of study in an English-speaking country	18 yrs	19 yrs	19 yrs	19 yrs
Length of English learning	Native	Native	Native	Native
Hours of English academic reading on print per week	11	10	8	8
Hours of English academic reading on screen per week	3	2.5	4	5
English reading proficiency (IELTS reading score)	NA	NA	NA	NA
Self-assessment of L1 (English) reading proficiency	very good	good	very good	very good

9.2.2 Research design and data collection

The semi-structed interview had two parts with regarding to the two main research questions. Within each part, 6 – 8 sub questions were asked. Part one was related to the reading process and strategy use while the participants undertook the RCA, aiming at understanding how the participants answered some specific comprehension questions. Additionally, the sub questions in part one were slightly different according to whether the participants did the test on screen or on print. Part two was related to participants' general reading behaviour regarding screen and print in the academic study, aiming at exploring the factors that contribute to their screen or print reading preferences.

Interview data were collected in two months, March and May in 2019. The interviews were conducted twice with each individual participant (each lasts for 30 minutes). The first interview was conducted immediately after the participant completed the RCA in a quiet classroom; the second interview was conducted one or two days after the first interview in a coffee bar. The reason was that the first part of the interview was similar to an retrospective interview, looking back to participants' cognitive reading process and strategy use while undertaking the reading test. Thus, combining the RCA and the first part of interview can help participants better recall their reading process.

Interviews were conducted in English for L1-English-speakers and L1-Chinese-speakers felt free to use Chinese or English if some Chinese concerned their English proficiency. Note-taking and audio recording were used with the interviewee's permission to collect the interview data.

9.2.3 Data analysis

This section outlines the qualitative analysis in this study including its purposes, nature and procedures. The main purpose of the analysis was to understand and interpret the differences of reading comprehension that had been found in literal and inferential levels between screen and print in previous RCA. The interview data analysis was conducted using both deductive and inductive approaches (Reichertz, 2013, Hennink et al., 2011). Data from the first part of the interview, focusing on participants' specific thinking process and strategy use on the questions involving the same subskill, were predetermined by research questions and collected purposefully according to each comprehension questions. Thus, the first part of the data analysis was essentially a within-case analysis in a deductive approach. Having completed the within-case analysis, the second part interview data – exploring factors contributing to participants'' screen and print preferences, were conducted in cross-case analysis (Cohen et al., 2018). Thus, the second part data analysis was more bottom-up and conducted using an inductive approach.

Drawing on Braun and Clarke's (2006) guidelines for data analysis, this thesis followed the following phases: a. familiarizing with the data; b. generating initial codes; c. searching for themes; d. reviewing, defining and naming themes; e. producing the report.

Familiarizing with the data. The recorded interview data were stored with the researcher's scanned notes in a password-protected encrypted drive. Having stored and organized, the data was transcribed verbatim over approximately three months. The data for each participant was transcribed and saved separately in word document with noted dates and times. Then the entire transcript was read and re-read with noting down initial list of ideas.

Generating initial code. Coding is a process of reducing the data to "meaningful segments and assigning names for the segments" (Creswell and Poth, 2018, pp.183-190). The coding stage is a recursive, non-linear and long-term process and "reading and interpretation are the starting points for meaningful analysis" (Bazeley , 2009, p.7). Following this guideline, the transcripts was re-read and printed with writing initial codes using no more than five words. The first part of coding which was related to the reading process during comprehension test was more "theory-driven"; the second part of the coding which was related to the factors contributing the screen print reading preferences was more "data-driven". For example, bearing the specific questions in mind, the entire data was coded and highlighted on how participants read the text and approach the comprehension question answers (Figure 9.1).

Data extract	Coded for
I first read the question on the separate sheet and quickly go back to the text from the beginning. Then I found the word 'decisions' in the first paragraph, so I copied the following whole sentence as the answer. I think this question is not hard because I can find the answers in one sentence.	1. locating the answers 2. direct word matching 3. from question to text (Chinese)

Figure 9.1 Data extract with codes applied

Searching for themes. The following phase was to conveying the codes into themes. Essentially, this phase is to "sort different codes into broader level of themes" and combine a list of codes to form an overarching theme (Braun and Clarke, 2006, p.89). The codes were rearranged into themes and identified the relationship between codes and themes by using mind-maps. For example, the codes were aligned according to the emergent themes from the comprehension questions involving different subskills in analysing the first part of interview data. By different subskills, themes were included correspondingly, indicating strategy use by the Chinese and English participants.

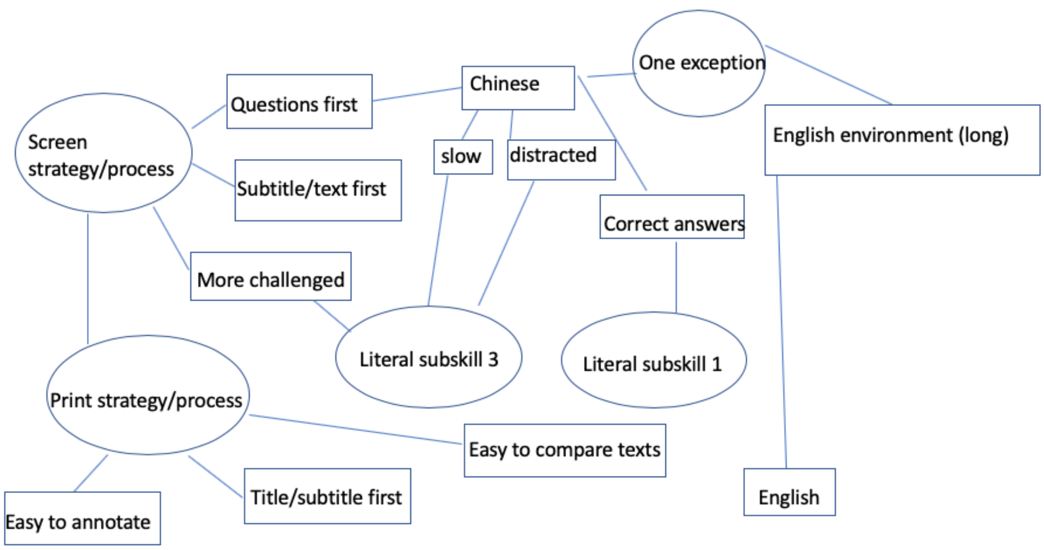


Figure 9.2 Initial thematic map, showing examples of subskill 1 and 3

Reviewing, defining and naming themes. According to Braun and Clarke (2006), reviewing themes aims to identify themes within the data which are coherent and meaningful but also clear and identifiable; defining means identifying the essence of each theme and determining what aspect of the data each theme captures. The collated data extracted for each theme was reviewed and organized into a coherent and internally consistent account with meaningful summary (Figure 9.5). Meanwhile, a summary of what is of interest of the data and why was noted down. For each individual theme, a detailed analysis of how each theme (e.g. patterns in subskill 3) can fit into the broader theme in relation to the research question (e.g. reading process and strategy use) was conducted and the overlaps between themes were removed. Such a structure was helping to explain the differences among the subskills that were shown in previous RCA (discuss further in Chapter 10).

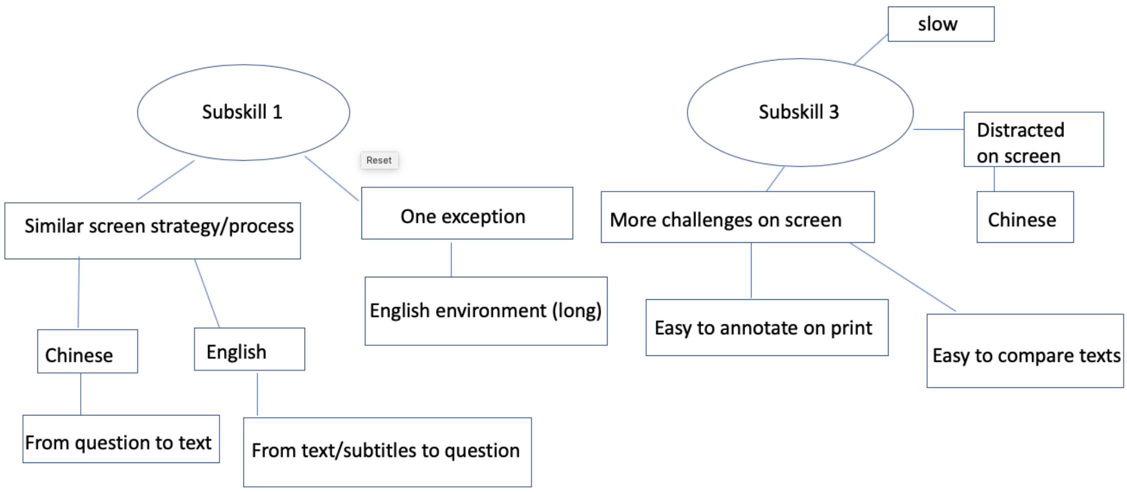


Figure 9.3 Developed thematic map, showing examples of subskill 1 & 3

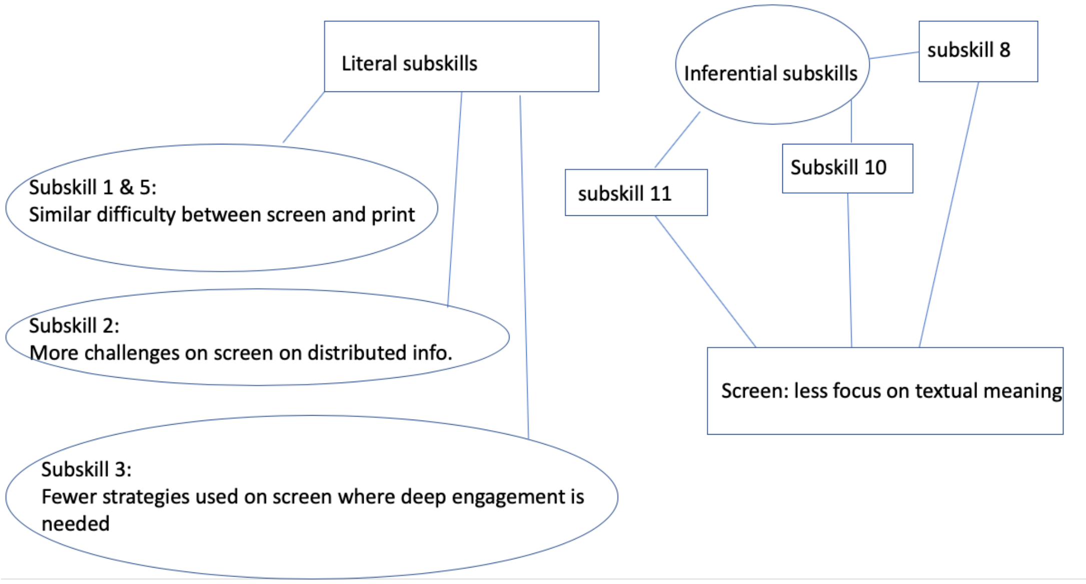


Figure 9.4 Final thematic map, showing four themes on reading process and strategy between screen and print

Producing the report. A report was finally produced with four themes on reading process and strategy during the reading test (9.3.1) and five factors contributing to screen and print reading preferences (9.3.2). Each theme is linked back to main research questions and related with data extracts. Additionally, Chinese interviews were free to use all their language repertoire in the interviews according to their preference. For this, the data of participants' answers were translated by the researcher from Chinese into English for coding purposes. A literal and balanced approach was used to retain the most important features of source language (Chinese) and target translation (English) (Hervey and Higgins, 2002).

9.3 Results

9.3.1 Reading process and strategy during the RCA

In this section, the results from the qualitative part of this mixed methods study are presented that the participants experienced more cognitive challenges reading on screen than reading on paper. This conclusion was consistent with the quantitative results from the RCA with higher scores attained on paper than on screen. Even so, regarding each open-ended question within literal and inferential level of comprehension, interviewees reported their difficulties and steps they took to overcome during taking the RCA.

9.3.1.1 Theme one

Participants identified similar levels of difficulty between screen and print but used different strategies in answering literal-level questions testing the subskill of *Recognition/recalling detail and outlining* in the subskill.

Recognition and recalling detail

When looking for answers to questions testing the subskill of *Recognition and recalling detail,* interviewees either reading on screen or print said that they could locate information at ease because the answers were relatively straightforward once some targeted words had been spotted in the text. However, L1-Chinese participant Yu and L1-English participant Eric indicated that different strategies were used in approaching the same type of question. L1-Chinese Yu tended to read the questions first and then immediately search for answers by one-to-one word matching. On the other hand, L1-English participant Eric revealed a tendency to scan the whole text before answering questions and pay attention to the headings and figures in the meantime. When the same questions were asked for the other two Chinese speakers, Lin and Zhao, and three English students, differences in strategy use were also found. Excerpt 1 (Yu-screen) and Excerpt 2 (Eric-print) explained:

Excerpt 1

Subskill 1: Recognition and recalling detail

Question 1: What factors will affect people making business decision in an organization?

Researcher: How did you get this answer to question 1? For example, what is the first step when you look for the answers?

Yu-(screen): “ I first read the question on the separate sheet and quickly go back to the text from the beginning. Then I found the word ‘decisions’ in the first paragraph, so I copied the following whole sentence as the answer. I think this question is not hard because I can find the answers in one sentence.

Excerpt 2

Subskill 1: Recognition and recalling detail Question 2: Please give two examples of different types of marketing factors mentioned in the text. Researcher: Can you briefly describe how you find the answers to this question? Eric-(print): well, to be honest, I didn't look at the questions before I read the text. I tend to quickly flip the pages to see the length of the text. I looked at the question and circled the key word "factor" and then turned back to the text. Some important terms such as "macro, micro environment …" were highlighted by the author so they popped up at first sight. I looked at the heading of this section and confirmed the word 'factor', which exactly what I'm looking for. I also noticed there is a Venn diagram in that 'factor' section. I literally did not read the words in the text but found the answers in that figure.

Nevertheless, another Chinese-speaker, Wang, who had been studied in the UK for three years showed a tendency for scanning the text with a focus on the subtitles and illustrations. When undertaking comprehension questions, he made use of memory to recall subtitles and meaning from sections rather than rely on word matching from the beginning. Thus, the difference between strategy uses may be due to the different first language but also the length of study in an English-speaking environment.

Excerpt 3

Subskill 1: Recognition and recalling detail Question 1: What factors will affect people making business decision in an organization? Researcher: Did you read this text before or after you looked at this question?

Wang-screen: Of course I read the text first. But I read it super quickly on the laptop, just for looking at the headings, or the pictures if it attracts me. This is my reading habit, I think. Researcher: Why do think reading the subtitles are important? Does it help you to answer this question? Wang-screen: … I think, sometimes it does sometimes it doesn't. But, this is my reading habit, you know. One good thing I feel about doing so is, I can have a rough idea about the text… when I read the questions, I have a clue where are the answers about, like approximately in which section.

Outlining

Unlike the recognizing/recalling details, outlining-type question need readers to extract answers from the a group of sentences instead of identifying single word. Six Interviewees reading on screen and print suggested they were able to find complete answers with confidence. This is because, once they had correctly located the target information by word-matching, readers then read the relevant sentences to extract answers. Nevertheless, different approaches were also found between L1-Chinese and L1-English. That is , L1-Chinese speakers emphasized the strategy of elaborating syntactic meaning while L1-English speakers focused on searching and linking information beyond syntactic level. For example, having difficulty in comprehending a long sentence on screen, Yu and Wang Wang re-read the sentence and split it into small parts on mind. Another Chinese interviewee, Lin (print), adopted a similar approach. Lin read slowly the long sentence, underlined phrases and translated parts of the sentence. Conversely, L1-Englsih speaker Emma (print) did not read intensively one sentence but paid attention to the relationship beyond sentence.

Excerpt 4

Subskill 2 outlining Question 14： List two consequences of economic recession mentioned in the text.

Text: "… during periods of recession, consumers are likely to postpone major purchases such as washing machines or new carpets due to uncertainty about employment security, and (by the same token) businesses will cut back on capital expenditure for such items as new factories or machinery."

Researcher: …You found this sentence is too long to understand, how did you overcome this issue?

Wang-screen: "… I reread the sentence and broke this sentence into smaller parts in my mind. If I failed I read it again. "

Lin-print: " I felt like the answers are in this sentence but I only found one - consequence - postpone major purchase. I was confused with ' uncertainty about employment security, capital expenditure'… so I translated from 'due to uncertainty…or machinery' into Chinese. I feel that only switching into Chinese helps me to understand.

Excerpt 4 shows differences related to reading strategies between reading on screen and print were not particularly prominent in the interview data for these two types of comprehension questions. This seemed in line with the quantitative results from reading test with similar scores achieved on screen and print.

Nevertheless, differences in reading strategy were found between L1-Chinese and L1-English participants. Chinese readers both on screen and print primarily relied on lower-level reading skills when searching information and elaborating ideas – word recognition, syntactic parsing and translation. Unlike L1-Chinese readers, L1-English seemed to rely on the higher-level reading skills – drawing on discourse knowledge to building textual representation. For example, Emma (print) paid attention to the linking words such as *due to, by the same token* when seeking answers (excerpt 5).

Excerpt 5

Subskill 4 outlining

Question 14： List two consequences of economic recession mentioned in the text.

Text: "… during periods of recession, consumers are likely to postpone major purchases such…, and (by the same token) businesses will cut back on capital expenditure…."   Researcher: I noticed you've underlined due to, and by the same token on the text, do you think these words are important?   Emma-print: "Yes. I thinks so. When I was reading this paragraph these kinds of words help me sort out the connection between the phrases or sentences."   Researcher: So do you think the connection between phrases or sentences affect your understanding about the text?   Emma-print: To be honest, I wasn't sure if it affects my understanding. I just felt it easy to follow the text with these words.

9.3.1.2 Theme two

Participants experienced more challenges on screen than print in answering literal-level questions which involved memorizing/summarizing distributed information in subskill – getting main idea and summary.

Getting the main idea

Interviewees reading on screen experienced difficulties in getting the main idea, in particular when they were required to identify the main idea for a section containing several paragraphs. L1-Chinese participants Yu and Wang both expressed their reluctance to answer this type of question because they had to spend an amount of time reading every line of the paragraphs but ended up with getting lost in the text. If coming across a long paragraph on the laptop screen, Yu said he seemed to easily forgot what had been already said when reading toward the end. If coming across several paragraphs, Wang added that he tended to scroll the mouse back and forth several times to make connection of each paragraph. Furthermore, Mia (screen) mentioned a scrolling issue that affected her reading on screen. That is, while scrolling the cursor might not stop at the exact place which was desired, in particular when reading several paragraphs across pages. Excerpt 6 and Excerpt 7 illustrated this phenomenon:

Excerpt 6

Subskill 2 Getting the main idea of a paragraph and large body of the text Question 6 What is the main idea of the section “cultural environment”? Researcher: I noticed that you only highlighted this heading “cultural environment” but you didn’t leave any other marks on screen. Why is that ? Yu-screen: well, let me think … highlighting the heading was because I reminded myself I have to read this section carefully. I don’t know why I have no other marks later, maybe I forgot. But to be honest, I don’t get used to underline on the PDF text. Researcher: So you just read through the section and remember ideas in your mind? Yu-screen: I guess so… I scrolled down quickly and … in the end I picked up the first sentence and wrote down the heading of that section as main idea. I didn’t want to read every sentence in detail on laptop because it’s too long and time consuming… I can’t remember that much.”

Excerpt 7

Subskill 2 Getting the main idea of a paragraph and large body of the text Question 6 What is the main idea of the section “cultural environment”? Researcher: When you read this section on the laptop, did you have any problems like scrolling the mouse, highlighting the text or other similar issue? Mia-screen: well… I literally read very fast so I scrolled quickly. For example, I saw this question and went back to search for those relevant paragraphs. I’d like to read the first paragraph and then the last. Research: Do you mean you skipped the middle paragraphs? Mia-screen: No, I didn’t skip them. I like to read the beginning and conclusion first. The issue came when I jumped to the conclusion by scrolling. The annoying thing is that I have no idea where I’m supposed to stop. In most cases, the text didn’t stop exactly where you want. That’s why I repeated reading some paragraphs. ”

Similar issues were also reported when talking about answering summary-type question. Yet different reading strategies were found for readers on print text. To illustrate, L1-Chinese participant Zhao took advantages of the printed materials to make notes instead of relying on the cognitive resources to memorize. She said,

Excerpt 8

> Zhao- print: "…for this question I have to summarize these paragraphs. I wrote down some key words on the margin for the first paragraph in case I forgot what is this part about. Then I continued to read next and also made notes on the side. I draw a mind map to understand because some concepts are new to me, like intervention, recession, stockpiling … At last, I put the notes together as a summary for the section."

The quantitative result of the reading test seemed to be inconsistent with the interview data. That is, participants on screen performed as well as print in getting main ideas but better on print than on screen on summary questions, although interviewees reported they felt more challenged and less confident on screen in doing both types of questions. One possible explanation for this would be that main ideas could be directly found in the original text in subtitles, headings, first sentence regardless of reading on screen and print, even though reader themselves assumed they would have more difficulties on screen. Another possible explanation for the fact that participants on print significantly performed better than those on screen in summary questions is because a summary required readers to give answers in their own words on a basis of overall understanding. A summary needs a deeper engagement and interaction with the text and yet readers reading on screen seemed to have reduced their level of interaction with the text. This was also evidenced from the interviewees when asked about their reading process in answering questions in the subskill—Recognition or recalling comparison or cause and effect. This will be reported in detail in next section.

9.3.1.3 Theme three

Chinese participants used fewer strategies reading on screen than print in answering literal-level questions which needed deep engagement with the text in the subskill – Recognition/recalling comparison or cause and effect.

Chinese interviewees reading on screen revealed that they encountered more challenges when answering why-questions than what-questions. They spent more time in elaborating relevant ideas but have certain level of uncertainty about the textual meaning. This was typical in a situations where new terms and concepts appeared in the text.

Excerpt 9

Subskill 3 Recognition/recalling comparison or cause and effect. Question 9: Given that some macro factors are common for all firms, why will they affect firms differently? Text: P31 "In some cases there will be overlap between the micro environment and the macro environment. For example … This has certainly been the case with major fruit-importing companies operating in Central America. On the other hand…" Researcher: Why did you leave this question blank? Did you find it too hard to answer? Yu-screen: er… I didn't remember why I left it blank. I think I didn't find the answers at first so I decided to continue completing the other questions. Let me think why.. Oh, yes. I think I did find where the relevant parts in the text but, to be truth, I didn't know the word subsidiary. And 'macro' was also a new concept to me. I looked back and tried to remember the explanations…I didn't understand why it mentioned the example of fruit-importing companies here…and what were relationship between them?" Researcher: So did you continue reading the text even though you did not understand this paragraph?

Yu-screen: well, it seemed as it is. I just continued reading the rest… you know, as I said before, I would easily forget what I've read on laptop.

However, another Chinese interviewee who read on print used different reading strategies for answering the same type of question. Lin (print) said she circled the key terms in this paragraph and wrote them aside in order to find their relationship.. Arrows, question marks or other symbols were utilized to help visualize the textual meaning.

Excerpt 10

Researcher: I noticed you made a lot of notes and drawings on paper, how is this helping you to read?

Lin-print: I like circling words or make notes while I'm reading. Otherwise, I would lost the track. You know what, I sometimes fall asleep when reading long text. So I developed this habit to follow the text.

Researcher: What paragraph or question is this mind map for? What does this symbol mean?

Lin-print: oh, this was for question 9… the arrow to the left means due to (reasons), the arrow to the right means lead to (consequences).

In comparison to the interviewee reading on screen, participants reading on printed text made physical attributes of reading on paper as opposed to reading on screen. Readers on paper can better engage in the text; however, participant reading on screen tended to rely on the limited cognitive resources to understand text. When the cognitive resources are over-burdened, difficulties arise from reading comprehension. This might be one of the reasons for Chinese participants reading on screen attained lower scores than those on print in the questions of the subskill recognition/recalling comparison or cause and effect. The reasons for this will be discussed in Chapter 10.

9.3.1.4 Theme four

Participants reading on screen tended to focus less on the textual meaning than those reading on paper in answering inferential-level question which involved accessing background knowledge in the subskill of inferring supporting details/main ideas, inferring case and effect and predicting outcomes.

Inferring supporting details

Two interviewees reading on screen said that it was difficult to focus on the screen for long time. Mike (screen) and Wang (screen) tended to access to prior knowledge to support their understanding of the texts. Once readers failed to locate direct answers in the text, they switched to top-down reading strategy such as activating topic-related background knowledge and discarded bottom-up reading strategy such as reading slowly or syntactic parsing。

Excerpt 11

Subskill 6 Inferring supporting details

Question 24: Can you infer what measures the government took to recover the economy in 2008 recession?

Possible answers: Rescuing bank; cutting interests rates; investing in large purchase; investing infrastructure

Researcher: What would you do next when you realized you can’t find answers to this question?

Wang (screen): well, I just think of what I’ve read from another article on similar topic. Those ideas occurred to me. Actually, I guess the answers might be hidden somewhere in the text. But I don’t bother to look back because I have some measures in my mind.

Researcher: Why didn’t you want to go back the text?

Wang-screen: I just felt like it is kind of time-consuming to look for those details in such a long document. Scrolling the mouse is not as flexible as flipping pages, in my experience.

Nevertheless, another three interviewees reading on print said they were inclined to work on the physical text even when having gaps in answering questions. For example,

Excerpt 12

> Subskill 6 Inferring supporting details
>
> Question 24: Can you infer what measures the government took to recover the economy in 2008 recession?
>
> Researcher: What would you do when you realized you can't find answers to this question directly from the text?
>
> Eric-print: well, I kept on reading and searching relevant parts more carefully.
>
> Researcher: Would you like to answer it using the knowledge that you've learnt or read from other related recourses?
>
> Eric-print: Well, probably I will. But I think that sort of knowledge helps me to understand the text but maybe not reliable for me to answer the questions here. If I can understand the text contents, I'd better stick to the text information.

One puzzling issue arose when comparing with quantitative results in the questions of this subskill: participants reading on screen achieved higher scores than those on print regarding the questions underlying by the subskill inferring supporting detail. One explanation could be that inferring supporting details did not need recalling a set of details from the text but relying on making inferences outside of the text. Less focus on the textual meaning may free up more cognitive resources that can be utilized to activate background knowledge.

9.3.2 Factors contributing to screen reading preference

9.3.2.1 Reading purposes

Two main purposes were identified from the interview data in the academic reading activities: read to learn or read to write. Reading to learn refers to reading to acquire new subject knowledge, for instances, reading for preparing lectures or attending seminars/workshops. Reading to write refers to reading to write assignments.

A general trend emergent from the interview data was that participants tended to engage in reading print when intensive and careful reading was required. Four Chinese and three English interviewees mentioned they preferred to read on hard copy textbooks when they were assigned specific reading tasks. However, reading preferences towards paper or screen, according to the identified reading purposes, were different.

Reading to learn

When reading to learn new knowledge, for instances, preparing a lecture or attending seminar, five interviewees showed a tendency towards paper over screen. Three Chinese students indicated that they would borrow books from library and make a copy of the required chapters so that they could make notes on it. Two English students added, they searched e-books online, scanned the contents on screen but printed out the important chapters to read. One of the possible reasons was that, sustained reading can be better achieved on paper than on screen when the chapter contained over ten pages. For example, Lin said, " We are usually required to read one or two chapters before attending a lecture and one chapter usually contains at least ten, or even over thirty pages. It takes me 2 or 3 hours to finish one chapter. I can't concentrate my mind if I read on screen. I like to flip the pages."

Nevertheless, another Chinese interviewee, Wang, was not a typical case for the above pattern. He emphasized that when reading printed textbooks for preparing lectures and seminars, an amount of time was also spent in reading on screen. Reading on screen, as he pointed out, was not limited to the computer/laptop screen. He would keep electronic devices aside in order to look up unknown words in a translation APP installed in the phone, or to google a new term/concept in Chinese on iPad. The extract below showed,

Except 13:

Wang: "I like to use my phone to look up unknown words. It's quick and easy. Sometimes even [if] I know the meaning of that single word in Chinese, I still

don't get its meaning in that sentence. I'll then google the terms or the phrases, for example, an economic theory, in Chinese on my iPad. I'll click and read the websites to get more information about it."

Aside from textbooks, journal articles are another type of primary academic materials students need to read. Views on reading articles on screen or on paper are various for L1-English and L1-Chinese interviewees. English students seemed to have higher tendency towards reading on screen than on paper; conversely, Chinese students preferred to read print than on screen. To illustrate, two English students Mia and Mike agreed that when preparing for workshops they would read briefly the assigned articles on computer first but read carefully the sections most relevant to the reading tasks. Unlike Mia and Mike, Chinese students Lin and Yu responded they would print out every article that were required to read without an initial examination of them. The extract below shows different opinions from Mike (English) and Lin (Chinese):

Except 14:

Mike: "…the materials given by the tutors are literally PDF or available online. Of course, the first thing is to open the file. I'll have a quick look at the abstract on laptop. I keep reading on screen until I get some ideas to answer the reading tasks given by the tutor. I won't print out everything because I won't read the rest in detail unless I think it is important."

Lin: " I basically read everything on paper. Different from my home country, teachers here send additional journal articles through emails or upload on Minerva. But I absolutely need to print them out because we have group discussion in the workshop. I have to look up unknown words and underlie important parts beforehand and look at my notes while discussing."

From the above comments, the different reading patterns between English and Chinese students in general academic reading was consistent with the interviewee data about the reading process during comprehension test. That is, English students appeared to adopt a top-down and global reading strategy whereas Chinese students tended to use a bottom-up and careful reading strategy. Furthermore, this provided some evidences to explain the quantitative results. Chinese students were inclined to place emphasis on decoding words and sentences so that little working memory capacity was left over to comprehend overall texts (Kuzborska, 2015).

In summary, for the purpose of read to learn new knowledge, two L1-English and three L1-Chinese interviewees had a higher preference to reading printed textbooks over screen due to the length of texts. Nevertheless, one Chinese student believed the importance of additional reading on screen would facilitate understanding textbooks. In terms of reading journal articles for the purpose of attending workshops, English students indicated a higher preference towards screen with initial and optional reading than Chinese students who aimed for full comprehension and careful reading on paper.

One exception was, Wang, who had studied one year foundation course and two years A-level course in the UK, he showed a reduced preference towards print for another purpose of reading: reading to write assignments.

Reading to write

When reading to write assignments, a Chinese interviewee, Wang, seemed to have higher preference to read journal articles on screen. Wang explained, "I may as well not print it because I just read the most relevant paragraphs for my writing. You always have to use a lot of references in your assignment and they are mostly available in online resources. So there is no need to read all of them but to read selectively."

Another English student, Emma, suggested a similar reading pattern. She commented that most of her time was committed to searching and reading paper online because textbooks were less useful than journal articles when writing assignments. After scanning titles, abstracts and conclusions of an amount of paper, she adopted deep engagement with one single text. As she said, "I'm gonna read intensively one full article on paper if I think it is super important for my writing. Otherwise I make notes where necessary and save them on my laptop."

From Wang and Emma, it can be seen that when students read across multiple resources for writing assignment, rather than seek full comprehension of single text for preparing workshops, they are more willing to read on screen due to availability and convenience of electronic references.

However, there were also some exceptions. Lin and Eric seemed not to follow the above reading pattern. They tended to scan and skim on paper, as well as annotated each text so that they could compare and combine useful references for assignment writing. As Lin suggested, “we have to discuss intensively in our assignment so I like to make notes on different references and compare the ideas.” This can be explained by another factor contributing to screen or print preference: annotation.

9.3.2.2 The transfer of annotations

Annotation plays a vital role for some students in choosing whether to read hard copy or on screen. Five out of eight interviewees avoided to read on laptop/computer screen due to annotation issue. Interviewees emphasized the its importance in that they believed annotations can facilitate textual understanding and enhance memory during reading process.

Reading on screen may constrain the types of annotations, as the tools computer software provided were limited. A Chinese student, Lin, who was firmly supportive of paper-based reading, found it hard to make complex annotation on screen, such as noting confusing ideas with question marks, marking beside the statement, putting key information in graphics and charts. She indicated, “I like to draw (arrows) and mark something on the margin of the paper, but it is too difficult to do on laptop.” Without making annotations, she would easily forgot what had been read, “I have no clues to recall the previous contents and may have to read it again.” As a result, fewer annotations and less interaction between the reader and text would lead to more distractions and less engagement in reading.

Unlike Lin, another Chinese student, Wang, had a strong preference for reading and annotating on screen. As Wang explained, “Actually I develop a habit of opening a separate Word document when I’m reading. I highlight key words on screen once in a while, click add-note button, type my own ideas in the pop-up box, and finally paste quotations from the original text.” Further, compiling a list of quotes to form an argument seemed to be more efficient for Wang to do on screen than on paper. He emphasized, "for writing assignment it is very helpful to do so as you can review your notes at one glance from several resources."

Comparing Lin and Wang's different reading patterns, the explanations may be threefold. First, Wang, who had studied one year foundation course in the UK university, had longer exposure to academic reading in English-speaking country and had adjusted himself to online reading environment. He could flexibly transfer some paper-based annotation skills to screen-based reading. Second, Wang's annotations were not as complex as Lin's so that he could easily employ the paper-like annotations (such as highlighting, underling and commenting) on screen. Last, Wang's preference towards screen could be attributed to next factor: familiarity with computer software.

9.3.2.3 Familiarity with computer software

The familiarity with computer software, in this study, refers to the ability of making use of the tools the software provided (e.g. Word, Adobe) to facilitate reading activity. This appeared to be an important reason explaining students' tendency towards hard copy or digital screen. All the interviewees believed that they were part of a generation who grew up with advanced information technology so they were good at using computers. However, when students were asked about whether they had used software tools (such as highlighting) to facilitate reading, four indicated they knew but hardly used in the academic study. Another two students, Lin and Zhao stated they had little awareness of these tools as they had never thought about learning academic materials on computer. For example, Lin explained, at home country, she browsed English websites to gain travel information but seldom read electronic books in English.

Except 15:

> Lin: "Before I came to study in the UK, I didn't even have a chance to read a real English textbook. It's hard for me. I sometimes borrowed English novels, I mean, the book, from the school library… I didn't often read electronic books, I guess, because I didn't have much time in high school. Second, because our high school teachers, in China, didn't allow us to bring any electrical devices, phones, iPad, e-readers to school."

Zhao: “I thought I was very good at using electronic devices as I surfed the internet quite often … But I did not get used to using computer in terms of ‘deep learning’ … By that I mean highlighting the texts, making comments, drawing flow chart, tree map and so on. But I can’t do that directly on laptop. It may be feasible in some software, Word, maybe, but I don’t know how to do it. Actually, no one teach me it before … more importantly, I don’t have the patience to switch between documents on laptop.”

Additionally, Yu, mentioned he would use online dictionary and translation websites to help understanding academic materials, although this function was not provided by common reading software.

Unlike Chinese students, English students seemed to have more exposure to screen reading environment and thus more familiarized with the tools of software. Three out of four students said they had experience in reading books on digital devices and also had an awareness of basic annotation tools on screen. Mike, with a stronger preference to read on laptop, could be a typical example. He explained, “I personally quite enjoy reading on computers because the e-books had sounds and animations”. Regarding software usage, Mike indicated “he learnt a little bit from the ICT course in high school.”

Comparison between Chinese and English students does not mean the familiarity with reading software would increase students’ tendency to read on screen. It only implies that lacking awareness and practice in utilizing software can make Chinese students more likely to rely on the reading habits which they are used to. In an education system, such as in China, where paper-based exams are dominant and electronic devices are deemed as entertainment tools, school teachers would focus little on teaching students how to use computer for academic purpose.

It can be argued that employing complex annotations on screen was not as convenient as on paper; this inconvenience would contribute to lower tendency towards digital materials. This issue will be discussed in next chapter.

9.3.2.4 Second Language proficiency

Language was a fourth factor contributing to preference towards screen. Two out four Chinese interviewees indicated if they read in their first language they were more likely to read on screen. When asked about reading Chinese on screen, Lin and Zhao said they could read faster and catch main ideas more easily than reading English. Also, Lin, a Chinese student who had just studied in the university for two months, experienced more uncertainty and insecurity when reading English on computer than on hardcopy. She explained,

Except 16:

Lin: "I just felt kinda, 'not studying seriously' if I read an English article on computer. I'm not sure about author's opinions so I have to read slowly … to be honest, I've never read an authentic English paper before I came here. I really felt struggled at the beginning … I have to look up new words and write down its translation … but when I read Chinese I don't have that problem. So I don't mind reading Chinese on whatever screen. Actually I read on my phone every day."

However, another Chinese student, Yu, said language did not affect his lower preference to screen over paper in academic context. He explained,

Except 17:

"When I find an English book difficult to read on screen, if it's in Chinese, it won't change anything. The difficulty to read is not just about what language it's written. It's about the content … I've ever read the same English book in the version of Chinese translation, it's about economic policy. Guess what , it's even harder to understand because of the weird translation."

Similar to Yu, Wang also indicated language was not a concern for him in choosing the reading medium. The ability of switching between two languages made no difference for him to read in English or Chinese. Wang himself explained, Except 18:

> "it seemed to me the first journal article I read was here, in English, in my foundation year. Oh… yes. There was a situation when we had a task, like a business case study, I took a Chinese company as an example. Then I searched the websites and read news, reports or anything about that company in Chinese. Both languages seemed fine to me."

Turning to the English interviewees, three out of four considered language proficiency as an important factor in deciding reading medium. Taking Mike as an example, who had a higher preference toward reading on screen than hard copy, he was less likely to read on screen if the article was in his second language (Spanish). First concern was his Spanish language proficiency, " I don't think I can read Spanish as efficiently as English (first language) cuz my Spanish is not super good. I might read very slowly." Another reason was the fact that English articles were more available online in academic context, "even if the authors are Spanish, the articles are still in English." Further, he added, "for, kind of like, entertainment or casual reading, like newspaper or websites browsing, I still prefer to use my own language."

In summary, second language proficiency would affect their preference to reading on screen for the majority students. Two Chinese students tended to increase their preference to reading on screen if it is in the first language. Three English students tended to decrease their preference to reading on screen if it is in the second language. This emergent pattern was in consistent with previous quantitative results, suggesting that language was a significant predictor in reading comprehension between screen and paper. This may also help to explain why English participants achieved better scores on screen than Chinese participants.

9.3.2.5 Eyestrain and physical discomfort

Eyestrain and physical discomfort are another major concern for students in deciding reading print or on screen. Five of eight interviewees reported that their unwillingness to read on laptop/computer was due to eyestrain issue. Compared to reading hard copy, readers tended to suffer visual fatigue and blurred vision more frequently when engaging prolonged reading on screen. Eric, an English interviewee, stated that he preferred to read on hard copy because of his vision problem, “I believe staring at the screen for a few hours absolutely worsens my eyesight.” Mia added, “I felt I read slowly on my laptop and my eyes are more tired. When I read two documents at one time on screen, the size of each document shrinks, which makes my eyes very uncomfortable. Perhaps it’s a personal issue ... my laptop screen is not big enough.”

Eyestrain may not be only resulted from the text size but also from the screen lighting sources. For example, one Chinese student, Zhao, stated, “I can’t keep staring at any screen over one hour, like computers, laptops or iPad … I feel the screen is too bright for me to read texts … sometimes it makes my eyes itchy and red.”

Nevertheless, only one English interviewee, Mike, argued that the eyestrain was not a problem exclusively resulted from screen. Mike explained,

Except 19:

> “… your eyes definitely get tired if you read for a long period. It’s not an issue about choosing screen or paper. It depends how long you read. If you read on paper for two hours without break, your eyes must be very painful. But if you read for ten minutes on screen, you will be fine.’’

Furthermore, Emma pointed out that the main issue for her to avoid reading on screen was not eyestrain but physical discomfort. This involved in stiffness or pain on the neck, wrist or parts of body. To illustrate, she commented, “… because I need to scroll the text frequently and my hand has to keep that position all the time – resting my wrist on the mouse.” In contrast, kinaesthetic movements from paper-based reading, such as holding the materials, touching the paper, and flipping the pages, writing notes make reading activity more tangible. As Lin said, “I like feeling the paper while I’m studying”. Hence kinaesthetic movement from reading print might facilitate sustained reading in the academic study.

9.4 Conclusion

This chapter presented the findings from qualitative interview. Regarding the reading process and strategy during the reading test, four themes were identified：

- Participants identified similar levels of difficulty between screen and print but used different strategies in answering literal-level questions testing the subskill of *Recognition/recalling detail and outlining* in the subskill.
- Participants experienced more challenges on screen than print in answering literal-level questions which involved memorizing/summarizing distributed information in subskill – getting main idea and summary.
- Chinese participants used fewer strategies reading on screen than print in answering literal-level questions which needed deep engagement with the text in the subskill – Recognition/recalling comparison or cause and effect.
- Participants reading on screen tended to focus less on the textual meaning than those reading on paper in answering inferential-level question which involved accessing background knowledge in the subskill of inferring supporting details/main ideas, inferring case and effect and predicting outcomes.

Regarding the factors contributing to screen reading preference, five factors were identified: reading purposes, the transfer of annotations, the familiarity of software, the second language proficiency and eyestrain and physical discomfort.

Main study: Discussion

Chapter 10 Discussion on Reading Comprehension Assessment

10.1 Introduction

This chapter discusses the results from the RCA. It first highlights the main findings from the RCA then explains the reasons. Section 10.2 and 10.3 explain why similar performances were achieved at the literal-level comprehension but better performances were found on print at the inferential-level comprehension. Then, section 10.4 focuses on the differences between L1-English-speakers and L1-Chinese-speakers. The last section 10.5 responses to the hypothesis that have raised in Section 2.7, regarding the hybrid framework of reading comprehension assessment.

The main findings from the RCA in the current study are:

1. For literal and inferential-level comprehension questions together, participants achieved better performances on printed texts than onscreen texts.
2. For the literal-meaning comprehension questions only, no significant difference was found for the participants.
3. For the inferential-meaning comprehension questions only, a significant difference was found with higher scores obtained on print than on screen.
4. The above findings were consistent with both L1-English and L1-Chinese-speaking participants.

The present research considered the length of the texts was first main factor contributing to the poorer performance on screen for literal and inferential comprehend questions together. The text used in this research consisted of 7 pages with 3500 words. When longer texts were involved, two main issues related to navigation occurred. One issue was, a reader had to scroll frequently between portions of texts to access textual information. Still, the frequent scrolling may interrupt continuous reading process, increase cognitive demands and thus negatively influence textual recall (Wästlund, 2007). Another navigational issue concerned with reading a split sentence across pages on screen. Sentence splitting made readers return to the previous more often than reading on paper, implying that jumping backwards and forwards became more frequently. This was supported by evidences from eye-tracking research, suggesting that readers may need to check back and re-read earlier sections to trace an argument if a long sentence was split (Kretzschmar et al., 2013). Therefore, the current study adds another piece of empirical evidence to the recent postulation that when reading lengthy text on computer screen, the navigational issues would disrupt continuous reading process so as to impede overall comprehension of the text.

The second consideration resulting in lower scores on screen was related with the frequency of task switching (Mayes et al. 2001; Noyes and Garland 2008). In this study, participants in the screen condition had to switch between the text document on screen and answer sheet with comprehension questions on paper. Switching from one task to another would disrupt reading process and hampered informing cognitive map of the text structure (Hou et al., 2017).

10.2 Why are similar performances at the literal-level comprehension?

The present study indicated that when answering literal level comprehension questions, reading on paper was similar to reading on computer. A first glance of the finding appeared to be inconsistent with some existing research (Clinton, 2019; Singer & Alexander, 2017; Singer Trakhman et al., 2018), arguing that both literal level and inferential level of comprehension were worse from screen than paper. There are two possible reasons. First, studies reviewed by Clinton (2019) only included participants who were reading in their native language; nevertheless, the current study included a population with intermediate-to-advanced/advanced reading skills who were reading in their second language. Second, it should be noted that the types of literal measures from the empirical studies reviewed by Clinton (2019) were different in nature from the current study. For example, the literal measures in two studies (Singer & Alexander, 2017; Singer Trakhman et al., 2018) of the systematic review (Clinton, 2019) focused on memorization of the texts rather than understanding of the texts. The literal level comprehension questions in current study involved lower-level skills, such as recalling of main ideas and supporting details, but also included higher-level skills, such as identifying causal relationship of texts and summarization. Therefore, it is worth emphasizing the subskills included in the current literal measurements were: Recognition or recalling of details; Getting main ideas of a paragraph and larger body of text; Recognition/Recalling comparison or cause and effect; Outlining; Summarizing.

The first reason accounting for the non-significant results is concerned with the time allocation. While many experimental tests allocated a time limit to reading (e.g. Singer and Alexander, 2017; Wastlund et al., 2005), the present study allowed sufficient time for the participants to complete reading test. Unlike Singer and Alexander (2017), participants in the current study had no time constraints and could always read back to the original text. This might facilitate onscreen readers to better retrieve textual information and find explicitly stated answers to the literal comprehension questions. In addition, the present findings were in accordance with previous research by Hou et al., (2017) which was also conducted with no time limit, suggesting that when given sufficient time participants on screen could perform as well as on those on paper. This implies that when textual information is accessible for readers to answer literal-level questions, recognition and recalling may not an issue once they can locate the specific information from the text.

Second, the nature of the literal-level comprehension tasks contributed to similar results. The literal-level comprehension questions focused on testing the subskills of recognition/recalling details, getting main ideas from a paragraph or a section and outlining and information to those questions were explicitly stated in a local area of the text. The process of undertaking literal-level reading tasks was in essence about searching and locating a piece of textual information. As a result, the reading outcomes may be similar between screen and paper for university-level students when the original text is accessible and time is sufficient. It may be argued that in some research (Li, Chen, & Yang, 2013; Payne & Reader, 2006) screens made it difficult for readers to construct a spatial representation of the text and in turn, impaired navigational performances and reading comprehension. Indeed, navigational issues may still exist, resulting in slower reading speed and more reading time, but it may not have a significant impact on the performances of literal level comprehension questions. This is in line with Daniel and Woody's (2012) research, showing that students performed similarly on the quiz but took longer to do so in the electronic conditions. In this regard, the medium of print or screen exerted little impact for readers when understanding literal-level comprehension.

Lastly, concerning the results on individual subskills within literal level comprehension, it is worth highlighting that the results on the literal-level subskills of *Recognition/recalling comparison or cause and effect*, and *summarizing* were significantly better on print than on screen. Participants performed better on print for these two types of questions, which showed an inconsistent pattern with other literal-level questions. The most plausible reason for this is, that answering these two types of subskill questions requires connecting distributed information from more than one single paragraph. For example, identifying cause and effect relation requires readers to look for the discourse markers and understand the relationship among sentences or paragraphs; summarizing requires readers to extract distributed information across paragraphs. This process may place an excessive load on cognitive resources because readers on screen need to hold the previous information in mind and scroll down to look for new relevant information. On the contrary, when reading on paper, readers can access and compare information on two separate pages simultaneously. Evidence from the follow-up interview supported this view.

10.3 Why was inferential-level performance better when reading from print, compared with reading from screen?

Readers may understand literal meaning of the text and find the correct information to literal comprehension questions similarly between screen and paper, without significant interference from the text presentation. However, when answering inferential comprehension questions, an effect of text presentation was apparent. Reading to answer an inferential comprehension question was likely to require more efforts and therefore consume more cognitive resources. Unlike answering literal comprehension questions, answering inferential comprehension questions requires the reader to make connections within the text, or between the text and background knowledge (Cain, Oakhill, & Bryant, 2004; Clinton & Van den Broek, 2012). Answers need to go beyond individual ideas stated in the text (Stevens et al., 2015) and integrate them with readers' own ideas. To this end, the subskills included in the current inferential measurements were: Inferring supporting details; Inferring main ideas; Inferring cause and effect; Predicting outcomes; Judging facts and opinions; Judging strengths and weakness of a position; Judging adequacy and validity.

There are three possible explanations why participants performed significantly better on printed text than onscreen text for inferential-level comprehension. They can be attributed to three categories: 1. the attributes of text medium (physically); 2. cognitive load and working memory capacity of individuals (psychology); 3 visual fatigue and tiredness (ergonomically).

First, the attributes of text medium refers to the negative effect of navigational issues (Magen et al., 2013) on screen (Hou et al., 2017) and positive effect of tangible interaction with paper. Research has demonstrated that, the navigational issues, such as scrolling, give rise to the instability of spatial layouts of a page and this mobilization of fluid text, in turn, interrupts the reading process (Cataldo, Oakhill, 2000; Kerr & Symons, 2006; Le Bigot, Passerault, Olive, 2009; Wastlund, 2007). Such interruptions would make it more difficult for readers to reconstruct a coherent propositional model of textual information (Kintsch, 1998). Without coherent textual information, readers might fail to understand the implicit meaning of the texts, resulting in poorer performances for inferential-level comprehension questions on screen.

The tangibility of the physical paper enabled readers to actively engage in the text, which contributed to build cognitive structure of the text. Empirical research has indicated the importance of the tangibility of the physical paper. For instance, a considerable an amount of work (Mangen, Walgermo, and Brønnick, 2013; Mangen, 2008; Mangen & Schilhab, 2012) investigated the role of haptic engagement in the reading process, such as the sense of touch and the movement of hands. In this regard, the present study supports the view that there are positive effects of haptic interaction with the physical text, but from a different angle. Findings from the interview data in the present study indicated the convenience and importance of making annotations on physical texts. When reading and learning from a complex and lengthy text on paper, readers were able to interact with the text, by means of annotations, such as highlighting, underlining and making notes on the margin with ease of a pen. It was such active interactions with the text on physical paper that provided readers with fixed cues to construct cognitive structure and metal representation of the text. On the contrary, reading on screen weakened the extent of interaction between readers and text because onscreen texts seemed intangible and mediated (Mangen & Schilhab, 2012). Further, onscreen environment appeared unfriendly to annotate and therefore annotative strategies were more available to use for readers when text presented on paper than on screen. In turns, the annotative strategies such as text-highlighting can effectively improve the accuracy of the questions that required inferential processing, which was evidenced in a study by Ben-Yehudah & Eshet-Alkalai (2018) when comparing comprehension performances under with – and without – highlighting conditions. On the other hand, it should also be considered that the inconvenience of onscreen annotations was due to the participants' unfamiliarity with the software and this is difficult to control in the study. Therefore, the interaction with the tangible paper affords readers richer engagement with the text compared to onscreen texts, which enhanced information encoding (Hou et al., 2017) and inferential comprehension.

Second, findings from the WMCT in this study can also provide some explanations for the participants' better performance on paper for the inferential comprehension. It was found that undertaking inferential comprehension tasks either on print or on paper, complex working memory, responsible for both storing and processing information, came into effect. This implied that participants not only had to maintain (storage) but also manipulate (processing) information. On this ground, the act of annotation on paper was more convenient and natural than that on screen, and then readers on paper have more cognitive resources available for processing of the text, consequently, with a better comprehension result. The frequent annotations on paper, such as drawing symbols on margins or highlighting key words (evidences from interview data), would help to reduce the cognitive load, which might be consumed by maintaining textual information, and therefore leave more working memory capacity for processing the text. This was in line with the view of Hou, Rashid and Lee (2017), pointing out the haptic coordination may offload the visual sensory processing and allow more cognitive capacity for comprehension. Unlike reading on paper, readers may consume more working memory resources to maintain what had been read due to possible distractions on screen (e.g. task switching, scrolling, jumping and navigating in a long text). The more working memory resources was consumed by storing textual meaning, the less could be utilized to integrate knowledge in the long-term memory. Thus, when limited working memory capacity was occupied by loads of incoming textual information and at the same time may be consumed by other distractions, readers on screen may find difficulties in constructing and updating situational model of the text, resulting in disadvantages outcomes for inferential comprehension.

Third, some ergonomic issues related to digital screen contributed to lower scores for the inferential comprehension. Factors such as text size, screen resolution, backlighting and luminance contrast might contribute to physical differences for screen versus print reading (Lee, Ko, Shen, & Chao, 2011) For example, the lighting sources may result in visual fatigue (Mangen et al., 2013) especially when reading lengthy texts as in this study. Other issues include refresh rate and fluctuating light which may also result in poorer screen reading comprehension (Garland & Noyes, 2004). These ergonomic issues, in turns, made readers distracted due to increased fatigue and tiredness and consumed cognitive resources in the working memory capacity (Lenhards et al., 2017). This was also evidenced from the interview data. One participant suggested that it was easy to get lost in the text and forget previous textual meaning due to eye blurring when constantly searching for information on screen. Therefore, the degraded effects on the visual processing had negative implications for inferential level comprehension.

10.4 Why are there some differences between L1-English-speakers and L1-Chinese-speakers?

To recap, the section first highlights the findings in the RCA with regards to L1-English and L1-Chinese participants (Section 7.9 and 7.10).

1. In terms of total scores of literal and inferential-level questions, L1-English-speaking students obtained higher scores on both print and screen.

2. L1-Chinese performed equally well as L1-Enlgish participants on three literal-level skills including Recognition or recalling of details, Getting main ideas of a paragraph and larger body of text, and Outlining; but worse on Recognition/Recalling comparison or cause and effect on both screen and print. This effect was also found for subskill 5 (Summarizing) but only when reading on screen.

3. On all inferential-level subskills with an exception of subskill 10 (Judging facts and opinions), L1- English participants performed better on both print and screen.

First, the primary reason that L1-Chinese participants scored lower than their L1-English counterparts may lie in the automaticity with decoding. For L1-Chinese-speaking students who learn and use English as a second language, decoding is very likely to be less accurate or slower than their L1-English-speaking counterparts. Due to limited English reading experience and exposure (Stanovich and West, 1989), L1-Chinese students may take more efforts to decode a word, in the sense that conscious thought has to be devoted to it. As a result, fewer cognitive resources in working memory will be available for comprehension process. Furthermore, comprehension requires that the decoded words should be present simultaneously in working memory, so that relations among them can be processed (Kirby, 2008). If decoding is inadequately fluent, then the key information will have decayed by the time later information is decoded. Thus, this may explain why L1-Chinese-speaking students performed disadvantageously than L1-English-speaking students on the total scores and the majority of inferential-level questions on either screen or print.

Second, when looking into the individual subskills, L1-Chinese performed equally well as L1-Enlgish participants on three out of five including *Recognition or recalling of details*, *Getting main ideas of a paragraph and larger body of text*, and *Outlining;* but worse on *Recognition/Recalling comparison or cause and effect*, *Summarizing* and six out seven inferential-level skills (*Inferring supporting details, Inferring main ideas, Inferring cause and effect, Predicting outcomes, Judging strengths and weakness of a position, Judging adequacy and validity*). The reasons for this pattern possibility depends on the cognitive process and strategy use. Chinese students were found in the interview to be more prone to take a bottom-up approach in reading process (discussed in Section 9.2). For example, they got used to reading English from word to word and line to line. Even when having difficulties in understanding a text, they tended to attribute their deficiency into linguistic resources, such as unknown words (vocabulary) or lengthy and complicated sentences (grammar). This indicated that an amount of cognitive resources and strategies would be devoted into the dealing with detail and text-based information. As a result, Chinese students could perform equally well as their English counterparts in some questions that focused on surface-meaning of a text at a local level, such as *Recognition or recalling of details*, *Getting main ideas of a paragraph and larger body of text*, and *Outlining.* However, L1-Emglish-speaking students still were advantageous in the subskill *Recognition/Recalling comparison or cause and effect.* The thesis speculates that it is because even the information is explicit in the text readers are required to connect the scattered information at a global level. This may provide some evidence for the ideas of 'Low-level Inference' or 'Text-based inference' by definition of Applegate, Quinn and Applegate's (2002) and Kinstch (1998) rather than simply regarding them as one type of literal comprehension. In addition, It should be noticed that the conclusions and explanations about individual aspects of inference are tentative, given the lower Chronbach's alpha of these question subsets.

10.5 Is the hybrid framework of reading comprehension assessment applicable and valid?

As discussed in Chapter 2, Section 2.7, this thesis has raised three queries of the hybrid framework of reading comprehension assessment：

1. What evidence is there for the division between literal and inferential reading comprehension?

2. Do the multiple levels of comprehension and the subskills involved in each level really exist?

3. Do the multiple levels of comprehension and the subskills involved in each level follow a linear sequential order?

The hypothesis were:

1. The levels of literal comprehension and inferential comprehension exist.

2. The multiple levels of comprehension and the subskills involved in each level exist.

3. The multiple levels of comprehension and the subskills involved in each level follow a hierarchical order.

The results of this study confirm the Hypothesis 1 and 2. The empirical evidences are:

1. The statistical test (ANOVA) in the pilot study suggests there is a significant difference between literal-meaning and inferential-meaning questions.

2. The component factor analysis (PCA) in the main study suggests 12 factors remained from the 14 subskills in the researcher-developed RCA. Although some of the subskills in Barrett's Taxonomy may not applicable to higher education context, there is still some evidences showing the existence of subskills.

However, the research results disconfirm Hypothesis 3. Instead, this thesis argues that,

3. The multiple levels of comprehension and the subskills involved in each level do not follow a hierarchical order. Instead, the multiple levels of comprehension are interactive and the subskills are integrated.

There is also some informative qualitative evidence. The interview data suggests when students extract surface meaning of text they are still likely to rely on the schema acquired from their own cultural contexts to make sense of the meaning of words and sentences. For example, the logical connection between two ideas in the text could be automatic or explicit because of the shared cultural or topic-related background, but students may still struggle with literal understanding of a sentence, in particular for L1-Chinese-speakers, due to multiple meaning of a word. This was evidenced from interviewee Zhao, "It seemed that I knew every word in this sentence and I can even translate it into Chinese but the sentence still didn't make any sense to me. I'm quite struggled with that…".

In conclusion, the hybrid framework of reading comprehension assessment is applicable and valid to measure reading comprehension for university students. However, it should be noted that Barrett's Taxonomy assumes a hierarchical relationship between the levels of comprehension probably because it aims for classroom teaching and learning context. Yet, CI Model provides a better framework to examine the process of reading comprehension.

Chapter 11 Discussion on Working Memory Capacity Test

The main reasons for including WM in this thesis are twofold: 1) to look for theoretical explanations for the previous findings on reading comprehension from a cognitive psychology perspective (e.g. Multi-component Model of working memory and the Levels of Processing Model); 2) to investigate the role of WM and the possible subcomponents (e.g. central executive, phonological loop and visuospatial sketchpad) within WM may account for the comprehension performances at literal and inferential level across screen and print medium. To achieve the first purpose, Section 11.1 and Section 11.2 discuss the findings from the overall comprehension performances (total scores of literal and inferential comprehension questions) in light of the theories and models from WM. Section 11.1 considers why better performance was achieved on print than on screen (described in Section 7.1). Section 11.2 discusses why L1-English-speakers showed better performance than their L1-Chinese counterparts (Section 7.7 and 7.8).

To achieve the second purpose, Section 11.3 examines the role of WM in reading comprehension across different medium by discussing WMC measured using a composite score of span tasks. The next two sections discuss how the possible subcomponents within working memory may account for the findings on comprehension performances by discussing the results from WMC measured by separate span tasks (described in Section 8.4). Section 11.4 focus on how the WMC correlates with literal comprehension performances across two medium. Section 11.5 focus on how WMC correlates with inferential comprehension performances.

11.1 Why is reading comprehension performance better when reading on paper compared with reading on screen?

The RCA data showed that both L1-English and L1-Chinese speaking participants performed better on paper than on screen. The follow-up interview provides a plausible explanation for this, namely that when reading on screen it is more difficult to remember and retrieve textual information and easier to lose the track of the text than when reading on paper. These results could be considered from the Multi-store Model (Atkinson and Shiffrin, 1968) and the

Levels of Processing Model point of view (Craik & Lockhart, 1972; Craik & Tulving, 1975). First the multi-store model suggests that human minds receive different types of information – visual, audial and haptic; the sources of information influence the processing of the information. The materiality of the tangible paper may enable readers to have haptic interaction and richer sensory engagement with the text, which may enhance information encoding and comprehension. In contrast, the intangibility of screen may make readers detached and mediated (Mangen & Schilhab, 2012) from the physical medium, which may impair navigational abilities (Hou et al, 2017). This could explain why participants when reading on screen felt more challenged in retrieving and recalling, such as searching and locating a piece of information, in a lengthy text.

Second, working memory has a very limited capacity, meaning, when new textual information comes into readers' mind it replaces previous information. In the screen condition, participants could potentially stay at a shallow or visual level and complete the tasks with only a superficial level of encoding, such as sub vocal rehearsal. In the print condition, participants could potentially read the text and complete the tasks via semantic processing, involving a deeper level of processing. Various forms of annotations (underlining, highlighting, translation and drawing), which are reported by the participants on paper, give rise to the opportunities for elaborative rehearsal. By deep processing with the texts via elaborative rehearsal (e.g. the reading strategies and the haptic movement), participants on paper are likely to better understand the meaning of words, transfer the newly textual information into the long-term memory and also retrieve from long-term memory to integrate with the incoming texts. The haptic movement and the act of annotations may help readers to make deeper analysis and semantic connection with their existing knowledge.

In conclusion, the general finding from the RCA that participants achieved better performances on paper than on screen could be explained from the theoretical perspective of working memory. One explanation is that the materiality of print medium affords readers haptic interaction and rich sensory engagement which could enhance encoding and comprehension. Another explanation is that the print medium gives rise to opportunities for deep processing with the text via elaborative rehearsal (e.g. annotations), which could facilitate information transferring into long-term memory.

11.2 Why do L1-English-speakers have better reading comprehension performances than L1-Chinese-speakers across different medium?

Results from the RCA (see Section 7.9 and 7.10) found that L1-English-speakers had a better overall performance across both screen and print medium than L1-Chinese-speakers. These results could be explained using the notion of the episodic buffer of working memory (Baddeley, 2007). The episodic buffer is hypothesized to provide a link between the long-term stored knowledge and the central executive so that the stored knowledge can be utilized to supplement phonological loop or visual sketchpad. In other words, the episodic buffer can facilitate other components of working memory system and improves information recall by assessing to long-term knowledge about grammar, structure or semantics of the language. Regarding the current study, L1-English-speakers may have more advantageous grammatical and semantic knowledge about their first language, which is stored in the long-term memory, than L1-Chinese-speakers who have learnt and used English as a second language. As a result, L1-English-speakers could better access to their long-term stored knowledge and possibly recall more information than their L1-Chinese-speaking counterparts with the help of the episodic buffer. For example, L1-English-speakers could recall a chunk or several phrases, or a paragraph of meaningful units; conversely, L1-Chinese-speaking students might recall a few unrelated words due to a lack of meaningful connection between the central executive and long-term memory.

Furthermore, findings concerning the reading strategies used by L1-English and L1-Chinese-speaking students (discussed in Chapter 9) might provide some evidence for this assumption. L1-English-speakers tended to take a top-down reading strategy and made use of the textual cues to locate target information (e.g. subheadings or main idea of paragraphs) whereas the L1-Chinese-speakers were prone to using a bottom-up reading strategy and recalled single word or phrases to locate textual information. To illustrate, according to one Chinese-speaking student Yu, "… while answering the reading question, I first tried to match the key words that appeared in the question and the original text"; by contrast, "… I think about which paragraph(s) of the text might be related to the question", said by a L1-English-speaking student Eric. This may imply that L1-English-speakers tended to recall information by meaningful chunks whereas the L1-Chinese recalled by discrete words. Therefore, it is tentative that L1-Chinese-speakers' disadvantageous long-term knowledge of English language (e.g. semantic knowledge or structural knowledge) may hinder creating an effective connection with other components in the working memory system (e.g. central executive or phonological loop) through the episodic buffer, leading to a less successful performance on reading comprehension compared with L1-English-speaking counterparts.

However, the explanations should be considered cautiously as the participants in this interview were not randomized and only represent a small subset of the wider population.

11.3 How is WMC related to reading comprehension performance across different medium?

The first research question, related to the WMCT , indicated that WMC measured by the composite score of three span tasks (FDS, BDS and Corsi) significantly correlated with reading comprehension performance across two medium at both literal and inferential levels. More importantly, WMC can better predict reading comprehension on screen than on paper at both literal and inferential level. This implies that WM plays an important role in literal and inferential reading comprehension level across two medium. Readers may be more dependent on WM when they read on screen than on paper at either literal or inferential level. In order to comprehend a text on screen, readers have to be equipped with advanced language proficiency but also sufficient capacity in working memory.

However, the current findings contradict those of Alptekin and Ercetin (2010) concerning WMC only having an important role in inferential reading comprehension but not in literal reading comprehension. Yet the results are essentially consistent if we unpack the notions of WM and the measures that were used in both research. The WM span task in Alptekin and Ercetin (2010) was reading span test which essentially measured the complex WMC and tapped the central executive component. Yet the WM span tasks in this study measured both simple and complex WM and potentially tapped three different subcomponents, according to Baddeley's model of working memory. The findings (shown in Section 8.5) suggested that the complex WMC only had a meaningful relationship between inferential comprehension but no connection with literal comprehension on paper. Therefore, it is necessary to go beneath WMC measured by the composite score to consider more specifically the separate WMC measured by different span tasks. The next section will continue to discuss this issue.

11.4 How is WMC related to literal comprehension across different medium?

The most important finding, related to the second research question in WMCT, indicated that the simple span tasks (which tapped phonological and visuospatial simple working memory) can better predict the performances on print literal reading comprehension than complex span tasks. In contrast, both simple span tasks (tapped phonological simple working memory) and complex span tasks made equal contributions to onscreen literal reading comprehension. This implies that answering literal reading comprehension questions on paper consumes cognitive resources primarily for maintaining information; in contrast, answering literal comprehension questions on screen consumes cognitive resources for both maintaining and processing information.

The most possible explanation is that reading on screen, even when completing comprehension questions at a literal level, tends to be more cognitive loaded than print reading. Readers tend to confront different issues and distractions that might occur during the reading process on screen, for example, navigational issues within the text such as scrolling pages or a jumped cursor (evidence from interview data in Chapter 9). These navigational issues are likely to impede the process of continuous reading, by imposing instability and interruption which could lead to potentially negative effects (Baccino, 2004; Piolat, Roussey, & Thunin, 1997). To tackle such issues, readers would need to control the central executive by allocating their attention and sparing residual cognitive resources to manipulate the information from different sources (e.g. computer screen, software operations, text size or physical distractions). Therefore, when reading on screen for literal reading comprehension, although the intrinsic nature of the literal comprehension questions does not require such deep processing as inferential questions, the central executive component exerts an important function and complex working memory is involved.

Another interesting finding related to the role of WMC in literal reading comprehension concerns with the visuospatial working memory. It was found that the visuospatial working memory had a meaningful contribution for literal reading comprehension on print but no connection with onscreen literal comprehension. This implies that the simple working memory in the visuospatial domain, which is responsible for maintaining information, plays an important role in reading comprehension when undertaking literal comprehension questions on paper. This is in accord with previous empirical and theoretical research showing that having a good spatial mental representation about the physical text layout facilitates reading comprehension (Cataldo & Oakhill, 2000; Kintsch, 1998; Piolat et al., 1997). To this effect, the visuospatial working memory exerts an effect and supports readers' construction of spatial representation of the text. In order to respond appropriately to the literal comprehension questions, participants in this study are very likely to scan and skim the text by making use of the spatial fixed cues, such as towards the right corner, at the bottom, right margin of the page, or any other annotative symbols, to access and retrieve the target information. On the other hand, the insignificant correlation between visuospatial working memory and onscreen reading comprehension seems to confirm previous speculations accounting for less successful onscreen comprehension performances (Ekludh, 1992; Piolat et al., 1997; Kerr and Symons, 2006). It was speculated that readers in the screen condition were restricted to feel the spatial extension and physical dimension of the text (Mangen, 2006 & 2010; Sellen & Harper, 2002); or it is assumed that “difficulties in reading from computer may be due to interrupted mental maps of the text” (Kerr and Symons, 2006, p.5). As a result, readers' comprehension would be negatively affected due to the restricted visuospatial working memory. Therefore, it is not unreasonable to assume that the visuospatial working memory may facilitate literal reading comprehension on paper by constructing mental structure of a text whereas the instability of the screen text might restrict the visuospatial working memory.

11.5 How is WMC related to inferential comprehension across different medium?

The findings, related to the third research questions of WMCT, have shown that only the central executive component measured by BDS can predict inferential comprehension performances both on screen and print. In contrast, the phonological and visuospatial simple working memory had little meaningful contributions when readers undertake inferential comprehension tasks. These results are consistent with previous research regarding the medium of print (Alptekin and Ercetin, 2010; Walter, 2004 & 2007) but also have revealed the same pattern to the medium of screen. The consistency is that inferential reading comprehension is related with WMC measured by complex span tasks rather than simple span tasks, meaning, central executive component rather than phonological or visuospatial simple working memory are involved in the process of inferential reading comprehension. The shared resources underlying the central executive for information storage and processing are considered to be limited and work as a trade-off relationship. To this effect, the limited capacity is a major determinant of the success in completing inferential comprehension tasks. The inferential comprehension tasks in the current study involved in higher-level inferences or knowledge-based inferences, which required participants to extract the deeper meaning of the text by going beyond the text-base into constructing a situational representation of the text. Participants on screen may consume greater cognitive resources to maintain and process textual information due to ergonomic distractions or navigational challenges than their print counterparts. As a result, minimal resources were remained to construct textual units to integrate a situational representation of the text, according to CI model discussed in chapter 2. That explains why participants achieved inferior performances on screen for the inferential comprehension questions.

Nevertheless, the above assumptions do not easily explain an exceptional finding from inferential comprehension. That is, for one type of inferential questions – inferring supporting details, the performances on screen were superior than on print. This thesis speculates this is because this type of inference making – an example question “How do you predict the consequences of …?”, is less text-based but more knowledge-based. Readers are not necessarily required to integrate a situational representation of the whole text but link to a particular domain of background knowledge in the long-term memory. In turns, the detached and mediated medium of screen could enable readers to free up more cognitive resources in the working memory to connect their knowledge pool; in contrast, the readers on paper might be still predominately occupied by the textual information.

11.6 Conclusion

To achieve the dual purposes that have been set at the beginning of this chapter 1) looking for theoretical explanations from cognitive psychology for the RCA results; 2) investigating the role of working memory and its possible underlying subcomponents, this thesis has discussed the findings obtained from reading comprehension and working memory capacity tests. Regarding the first purpose, from the perspective of the Multi-store Model and the Levels of Processing Model, better performances on paper than on screen are attributed to richer sensory haptic movements and deeper elaboration with the text, which in general facilitates reading comprehension. Regarding the second purpose, WMC measured by the composite scores of simple and complex span tasks makes a meaningful contribution to reading comprehension on paper but also on screen. The contribution to screen is even greater than on paper, indicating, reading comprehension on screen is more dependent on WMC.

Furthermore, WMC measured by separate span tasks has shown consistency at the inferential level but also an inconsistency at the literal level between screen and paper. The consistency is that only the WMC measured by complex span tasks has a meaningful contribution to reading comprehension across two medium. The contribution to screen is greater than on paper. It is implicated that the central executive component underlying working memory plays the most important role in comprehending deeper level of reading tasks. Completing the deeper level of reading tasks is likely to be more challenging than on paper most possibly because fewer cognitive resources remain in the central executive in the face of navigational and distraction issues when reading on screen. On the other hand, the inconsistency lies in the different contributions from the WMC measured by separate span tasks to literal comprehension. The central executive component measured by complex span tasks has been found to have a meaningful connection with screen comprehension but little with print. The implication is that even though the tasks on paper may not tap the central executive component or complex working memory, they are still most likely to be involved when reading on screen. Another interesting inconsistency concerns with the visuospatial working memory subcomponents. It has a meaningful contribution to print reading comprehension paper but does not appear to have a strong connection with screen. This gives supportive evidences to previous assumptions that readers achieved better performances on paper than screen because tactile print could provide fixed and spatial cues to build a mental map of the text such as text layouts and structures (Mangen et al., 2013).

Chapter 12 Conclusion

12.1 Summary of the findings

This concluding chapter first summarises the thesis findings with respect to three separate empirical parts (12.1). The remainder of this chapter is divided to four sections. Section 12.2 offers an articulations of the contribution, followed by the implications of the findings for theory, methodology, pedagogy, and technology (12.3), the limitations of this study (12.4) and future research (12.5).

The aim of this thesis was to examine the possible differences of reading comprehension performances between screen and print text presentations for the first-year Anglophone university students and explore the reasons attributing to those differences. To address these issues, this thesis has conducted the empirical study in three parts: 1. RCA – Reading Comprehension Assessment; 2. WMCT – Working Memory Capacity Test; 3. A follow-up semi-structure interview.

The main findings of the first part RCA were:

RQ 1 - 3

- The medium of text presentation had a significant effect on reading comprehension performances. When looking at L1-Englslish- and L1-Chinese-speaking participants together, they tended to achieve better results when reading on print than reading on screen.
- There was no significant difference in the scores between the screen and print text presentation for the literal-meaning level of comprehension. However, a significant difference was apparent in the inferential-meaning level of comprehension, with higher scores obtained from print than screen.
- When looking at L1-Englslish- and L1-Chinese-speaking participants separately, the above pattern was consistent.

RQ 4 – 6

- When looking at L1-English and L1-Chinese-speaking participants together, the general pattern was that significant better performances in print than in screen were found in the following literal-meaning subskills: *subskill 3 – recognition/recalling comparison or cause and effect;*

subskill 5 – summarizing;

and the following inferential subskills:

subskill 8 – inferring cause and effect

subskill 9 – predicting outcomes

subskill 10 – judging facts and opinions

subskill 11 – judging strengths and weakness of a position

subskill 12 – judging adequacy and validity.

- When looking at L1-English-speaking participants only, the text presentation had no significant effect on all the literal-meaning subskills; but the effect was consistent with the above general pattern on the inferential-meaning subskill 8, 9, 10, 11 and 12.
- When looking at L1-Chinese-speaking participants only, the effect was consistent with the general pattern on the literal-meaning and inferential-meaning subskills but with an exception on subskill 10 (judging facts and opinions).
- Surprisingly, better performance was found on screen than print on the inferential meaning *subskill 6 – inferring supporting details* for both L1-English- and L1-Chinese-speaking participants.

RQ 7 – 8

- Overall, L1-English-speaking participants obtained higher scores on both screen and print than L1-Chinese-speaking participants.
- L1-English speakers obtained significantly higher scores than L1-Chinese participants on the literal-meaning *subskill 3* – recognition or recalling comparison or cause and effect. This effect was also found for *subskill 5 – summarizing* but only when reading on screen.
- L1- English-speaking participants performed better on all the inferential-level subskills on both print and screen. But the effect on *subskill 10* (judging facts and opinions) was only found on screen.

The main findings of the second part WMCT were:

- The working memory capacity measured by composite scores both storage (tapped by FDR) and processing (tapped by BDR) working memory played important role in literal reading comprehension when reading on screen. Further, the simple storage working memory related to visual-spatial domain did not affect reading on screen for literal comprehension questions.

- The measure of Corsi-block made no contribution to literal reading comprehension on screen but had a significant prediction on print. The effect of Corsi-block was even greater than the measure of FDR.
- Answering literal reading comprehension questions seemed dependent more on simple working memory (mainly responsible maintaining information) than on complex working memory (responsible for both storage and process information).
- Reading on print text presentation for literal comprehension questions could make better use of working memory related to visual-spatial domain (tapped by Corsi-block) than on screen.
- When performing the inferential reading comprehension questions neither on screen or on print, complex processing working memory came into effect and thus participants not only had to maintain (storage) but also manipulate (processing) information.

The main findings from the interview were:

- Participants identified similar levels of difficulty between screen and print but used different strategies in answering literal-level questions testing the subskill of *Recognition/recalling detail and outlining* in the subskill.
- Participants experienced more challenges on screen than print in answering literal-level questions which involved memorizing/summarizing information which was distributed in the text.
- Chinese participants used fewer strategies reading on screen than print in answering literal-level questions which needed deep engagement with the text information such as in the subskill - Recognition/recalling comparison or cause and effect.
- Participants reading on screen tended to focus less on the textual meaning than those reading on paper in answering inferential-level question which involved accessing background knowledge, such as in the subskill of inferring supporting details/main ideas, inferring case and effect and predicting outcomes.

- There were five factors attributing to participants' screen/print reading preferences: reading purposes, the transfer of annotations, the familiarity of software, the second language proficiency and eyestrain and physical discomfort.
- The participants had a higher preference to reading printed textbooks over screen due to the length of texts when reading to learn (new knowledge). English participants indicated a higher preference towards screen with initial and optional reading than Chinese students who aimed for full comprehension and careful reading on paper when attending workshops.
- When participants need to access multiple resources (e.g. journal articles, book chapters or reports) for writing assignment, rather than read particularly assigned texts for preparing workshops, they are more willing to read on screen due to availability and convenience of electronic references.
- According to Information-processing model, deeper process of information and the interaction between the reader and the text can play a substantial role in comprehending a text. Not surprisingly, participants who preferred paper-based reading indicated the main reasons were more distractions without annotations on screen and the inconveniences to make complex annotations on screen (e.g. question marks, marking beside the statement, putting key information in graphics and charts).
- The preference from print seemed to be changed towards screen for the participant with longer exposure to academic reading in English-speaking country and online reading environment (one Chinese).
- It was believed the participants were good at using computers as young generations but in fact they were less familiar with the software tools to facilitate reading in the academic study.
- Chinese participants had a lower awareness of utilising computer software to study academic English reading materials such as electronic textbooks than their English counterparts.

- Second language proficiency could affect the preference to reading on screen or print. Chinese students tended to increase their preference to reading on screen if it is in the first language. English students tended to decrease their preference to reading on screen if it is in the second language.
- Eyestrain and physical discomfort were major concerns for participants (five out of eight) in deciding reading print over on screen. Others considered visual fatigue was an issue exclusively resulted from screen but from the period of reading time.

12.2 Contributions

This thesis has made theoretical, methodological and pedagogical contributions. Theoretically, this study has attempted to established a hybrid theoretical framework (Section 6.5) for reading comprehension assessment for academic reading comprehension assessment in higher education (the level of comprehension) and discussed whether this framework can be adopted and extended to reading comprehension measurement in digital context (Section 10.5). The fundamental frameworks, Barrett's taxonomy and Construction-Integration Model, provide an analytical and practical model for classroom learning and teaching as well as a theoretical framework in researching and understanding the nature of reading.

Another major theoretical contribution pertains to the theoretical models used in the research of reading when extending to a digital environment. As indicated in Chapter 1, one justification for this research is "[there is] limited understanding of how particular attributes of the context, the text or the learner might interact with the medium to enhance or inhibit comprehension" (Singer and Alexander, 2017, p. 2). This thesis bridges this gap and suggests a multi-dimensional framework, taken into account of different attributes of the context, the text and the reader, to investigate the issues in digital reading environment, which is summarized in Section 12.3.1.

Methodologically, as described in Section 6.6, this thesis has made a contribution of creating a statistically valid and reliable reading comprehension measurement for university-level students and providing detailed test specifications and subskills. Much of previous research adopted neither a standardized or a self-designed comprehension test as reading instrument without statistically checking the validity or reliability. This thesis has provided transparent test specifications for designing the comprehension questions and used factor analysis statistical technique to improve the test validity, which may be extended to the reading research on different populations and contexts (e.g. university students in different disciplines or study levels, L1-English or L2-English-speaking students).

Another methodological contribution is related to WMC and reading research. Previous research on reading from screen often regards WMC as a contributor accounting for the differences between reading by difference medium (Single and Alexander, 2017; Magen et al., 2010) but offer limited empirical evidence. Also, reading-based working memory capacity tests are utilized as main research instruments to investigate the relationship with reading comprehension. By using the Multi-component Model of working memory (Baddeley and Hitch, 2000), this study goes beneath the general concept about limited WMC and unpacks different constructs of the WM system (e.g. verbal and visuospatial, storage and processing) to explore how the different aspects of WM may interact with the medium. Interestingly, a meaningful correlation was found between the visuospatial working memory and literal comprehension on print. This provides some evidence for the assumption that having a good spatial mental representation about the physical text layout facilitates reading comprehension (Cataldo & Oakhill, 2000; Kintsch, 1998; Piolat et al., 1997).

Pedagogically, this thesis provides some recommendations for university-level students in relation to the usage of different medium in their academic reading activities and for educators in relation to how to support their L2-English-speaking students to improve academic reading abilities in an era of digital environment, summarized in Section 12.3.3.

12.3 Implications

12.3.1 The multi-dimensional theoretical framework

This thesis attempted to establish a multi-dimensional framework about what and how the attributes from the context, text and reader interact with the medium so as to facilitate or inhibit reading comprehension.

12.3.1.1 Context

Context affects how the reader approaches the reading medium in two ways. Context, for this study, is defined as two levels: the target reading context and the original context (e.g. learning experience, cultural and educational backgrounds) of the reader. The target reading context determines the reading purpose and the purpose affects how a reader approaches the reading medium. In an academic reading context, there are dual purposes for reading (on screen or print): 1. reading to learn (new information and knowledge); 2. reading to write (a report, an essay or an assignment). Reading to learn new knowledge often involves academic activities such as preparing for a lecture or attending a seminar. For such situations, the benefits of tangible paper and the convenience of annotation with a pen may help students to deeply engage in reading material. Reading to write often involves activities such as searching and scanning journal articles or looking for wider references from an online database. For such situations, the availability and convenience of electronic references allows the readers to read across and filter multiple resources more efficiently, by contrast to taking time and money to print out reading materials.

The original context of the reader influences their understanding of the target reading expectations and such understanding again affects how a reader approach the reading medium. Previous research (Margolin et al., 2013) tends to assume the younger generations from primary students to university students are familiar with the technology due to the frequent use of computers or other forms of e-readers in daily life. However, this might not true for the students in this study who use English as second language and pursue higher education in an English-speaking educational systems. When first-year university students enter a new academic discipline, they tend to bring with them their different cultural backgrounds and perspectives from previous educational systems and learning communities. New Chinese students often bring experience of reading from print rather than from screens. Chinese students might also bring the assumptions about 'English reading as learning lexical features' from school-level reading activities rather than 'English reading as learning new discipline knowledge' for higher education. As a result, students tend to choose the best ways that they are used to and helped them to achieve success in their previous education context (Grabe & Stoller, 2011). Evidence from the interview suggests the frequent reading strategies included reading word by word, analysing sentence structure and grammar and translating texts. These reading strategies might be more easily achieved by reading on print. Furthermore, being educated in a different culture and literacy environment, Chinese students may have different perceptions of how academic reading should be done (e.g. reading digitally is for entertainment purpose; highlighting is necessary while reading; de-emphasizing the value of digital reading for systematic and serious learning) and how academic texts should be constructed and interpreted in a new academic context. Therefore, these different understandings make Chinese students fail to appropriate the conflict between prior assumptions and unstated expectations of the new academic community, causing challenges and difficulties in understanding academic text appropriately and increasing the reluctance towards screen reading.

12.3.1.2 Text

This study highlights the important attributes of the texts that interact with the reading medium include the text type, the text length and the nature of reading tasks. Previously one review article (Clinton, 2019) concluded that for informational and lengthy text, better performance is more often found from the print reading than screen reading. The current study aligns with this result and extends it to include a university-level and English as a second language population, reading a static disciplinary e-textbook. One possible reason for this finding is that reading a lengthy text on screen (over two or three pages) can increase the frequency of scrolling, and the frequent scrolling may interrupt continuous reading process and burden cognitive load, which in turns negatively influences textual information retrieval and recall (Wästlund, 2007). When university students read to learn new concepts from sustained engagement in a textbook, reading from screen is very likely to be detrimental for extraction of information, continual construction, integration, and updating of concepts (Biancarosa & Snow, 2006). On screen reading is likely to require a different set of strategies from read on paper.

In terms of the nature of the task – literal or inferential, this study reveals that there were no differences by medium on literal measures, but differences did emerge on inferential measures. This finding was contrary to the expectation that the benefit of paper for reading performance would be noted for both literal and inferential measures (Clinton, 2019),given that literal measures are based on memory of the text and reading from screen may interfere with encoding specific details relative to paper (Single and Alexander, 2017; Singer Trakhman et al, 2018). When retrieving the explicitly stated details to answer literal questions, participants from screen and print indicated similar reading processes and used similar reading strategies to locate information (a cycle of highlighting key words in the question, looking back at the original text and finding correspondent words in relevant sections). When unpacking the literal measures, the only difference exists on literal questions which require using information is distributed across paragraphs (e.g. recognition or recalling comparison or cause and effect and summarizing). For stated information that is locally located in a sentence or a paragraph (e.g. recalling and recognition details), the performances reading different medium appears to have little difference.

The nature of the reading task interacts with the WM by different medium and thus affects comprehension performances. The difficulty of the reading tasks influences the degree of interaction with the text and the involvement from working memory. In this study, although participants in the paper and print conditions had similar performances on literal comprehension tasks, for the participants in the paper condition, only simple working memory (storage) was likely to be involved whereas for the participants in the screen condition, both simple and complex working memory (storage and processing) were involved. This conclusion is based on the finding that the measure of FDR and Corsi-block made contributions to print literal performances but both FDR and BDR made contributions to screen literal performances. This means the medium of text could influence the degree of involvement from working memory but may not necessarily affect comprehension performances due to the task difficulty. This finding questions the conclusion from Margolin (2013) that onscreen reading and print reading are essentially equivalent because performances across two medium shared no significant difference. Therefore, discussing the equivalence across different medium should take different attributes of the text into consideration (text type, text length, the nature and difficulty of the task) because comparing the scores of comprehension performances solely may conceal the underlying differences.

12.3.1.3 Reader

The attributes of the readers that interacted with the reading medium include individual differences such as age, level of study, language proficiency, motivations and preferences towards reading etc. However, this study is particularly interested in a cognitive aspect of the reader – working memory. Readers' working memory capacity was found to have an important role in reading comprehension across screen and print medium. However, the extent to which effects on comprehension performance was influenced by the level of comprehension tapped by different reading tasks. The working memory capacity test suggested that onscreen reading comprehension at both literal and inferential levels tended to be more cognitively demanding. To this effect, one may assume that the reading on screen is more difficult and consequently comprehension performances on screen would be less successful than on paper. This assumption has been confirmed by previous research (e.g. Magen et al., 2013), which treats reading comprehension as a global construct. However, there is a slightly different picture when examining reading comprehension at two levels. Performance at literal comprehension level was not significantly different between screen and print; and this pattern was found consistently with both English and Chinese participants. That means, participants could perform equally well on both medium for literal reading comprehension tasks in spite of cognitive challenge. The literal comprehension questions designed in this research primarily involved using stated explicitly information or low-level inferences within the text. These type of tasks are essentially dependent on the level of language proficiency and surface readability features such as decoding and syntactic parsing and consequently, and do not impose much cognitive load on WMC (Sweller, 1994). Thus, when undertaking the comprehension tasks that are shallowed to complete, the medium has little effect on readers' comprehension performances for proficient readers.

Furthermore, this study confirms previous findings from print reading that when undertaking inferential reading tasks, complex working memory, rather than the phonological and visuospatial simple working memory, comes into effect (Alptekin and Ercetin, 2010; Walter, 2004 & 2007). The current study has extended the findings to the medium of screen. This means reading processes for inferential understanding on screen and print are similar because they both require maintaining and processing information. Nevertheless, participants on screen need to require more cognitive resources to maintain and manipulate textual information due to greater ergonomic distractions or navigational challenges than their print counterparts. As a result, minimal resources remain to construct textual units and integrate a situational model of the text. An implication of this would be reducing the mismatch between reading from screen and print might help free up cognitive resources that are available for maintaining and processing information.

12.3.2 Methodological implications for researchers

Researchers should be transparent and explicit in defining and choosing the reading measures. The types of inferences and the conception about inferential comprehension cannot reach an agreement in the present literature. This is probably one of the most important reasons that leads to the varying definitions of comprehension measurements and the inconsistent conclusions of the vast empirical research into the difference between reading in the print and on screen. For example, in some research the comprehension measurement assesses 'summarizing' and the researchers are inclined to define it as one kind of 'inference' as it involves 'deep processing' (Chen et al, 2014); however, in some other research (e.g. Hou et al, 2017), assessing 'summarizing' is merely about recalling and listing the factual information mentioned in the original text. Clearly, the latter measurement is literal comprehension oriented. On this ground, the different standards of reading measurement contribute to the inconsistent results. In other words, the researchers use different benchmarks to assess the same issue. Therefore, it is implied here that the researchers should make the definitions and frameworks of reading measurement clear and explicit in the empirical research. In this way, future researchers could identify the underlying frameworks and compare reading measurements from relevant research on the same basis.

Researchers should pay attention to the balance of the nature of reading measures when conducting comparative study between screen and print. For school-level populations, low-level and literal-level comprehension questions seem to be dominant comprehension questions. However, over-emphasizing on low-level comprehension might miss the opportunity to tap into the readers' higher-level of comprehension.

Researchers should give more information on the validity and reliability of the reading measures. For standardized test, the test specification can be less obvious and this increases the difficulty for other researchers in interpreting the general results. For self-designed test, the reading constructs are relatively explicit but the validity and reliability is seldom reported.

12.3.3 Pedagogical implications for learners and educators

Learners

- University students with advanced English language proficiency should make the most of both print and screen reading medium according to different reading purposes and reading tasks. For example, when reading a short text or selected paraphs with the purpose for searching for literal information, both medium might be effective; when reading lengthy texts with the purpose of intensive engagement, reading on print might be better. When searching and scanning for relevant references from wider online database to writing an essay, reading on screen could be more efficient to compare with multiple resources on a particular topic and filter reading materials.
- University newcomers, many of whom learn and use English as a second or an additional language, entering into Anglophone higher education community should be aware of the expectations on reading from the target context and accommodate the assumptions they bring from previous educational system and learning communities. Understanding the differences about the beliefs and values from original and target academic

community can help to better engage in reading activities either on screen and print.

Educators

- *EAP (teaching English for academic purpose) teachers*. The transfer of effective reading strategies from printed to digital platform is not always straightforward. This is because not all the reading strategies effective in print can be equally and successfully transferred to digital platforms for readers. The use of such reading strategies in digital context, such as the basic annotative features of note-taking, text-highlighting, probably needs explicit instruction and extensive practice (Ben-Yehudah & Eshet-Alkalai, 2018) before they are widely applied in the authentic academic reading. Therefore, those lecturers, teachers and course developers who promote "digital literacy" should provide sufficient training regarding reading digitally for students at all levels and ensure technologies can be incorporated appropriately into higher education system and utilized effectively in reading devices (Yue et al., 2015). One important implication from this study for academic training courses is that teachers should consider various cultural and educational backgrounds of the students, in particular those who use English as a second language, regarding digital reading experience. The interview findings of the current study suggested a lower recognition of digital devices among Chinese students for "serious" academic learning whereas a higher reputation for entertainment purpose. This led to an unfamiliarity with reading software for those students who had never studied in an English-speaking context. Comparatively, students who were educated in an English-speaking educational system had relatively higher awareness of computer-assisted learning and better usage of digital reading software. One should note that this finding cannot represent all leaners in all contexts. This only implies that it is imperative for future trainers take careful considerations of students' background differences so as to provide appropriate support and improve digital reading ability.
- *School-level language teachers*. L1-Chinese-speaking students' tendency to adopt a bottom-up approach in academic reading practice reflects the impacts of school-level English teaching methods in China. Li,

Wu & Wang (2007) pointed out the emphasis on linguistic elements such as vocabulary, grammar, and syntax in English-teaching in Chinese context. Influenced by such teaching methods, students are prone to develop a habit of reading English word by word. However, the findings of the current research indicates this has long-term disadvantages (e.g. time-consuming in reading lengthy texts or without grasping main points) for Chinese students in academic study. Indeed, whether the bottom-up or top-down strategy is true for students on screen or print reading still needs more evidences from eye-tracking research. Thus, this thesis suggest that language teachers, in particular those who teach English as a second language at school-level in China, should re-consider the current teaching methods and beliefs in reading classroom. Teachers should raise the students' awareness of the importance of cognitive reading strategies and higher-level skills in reading process. For example, teaching discourse knowledge, such as text structures and cohesive markers would improve effective reading comprehension, as shown in some recent research for university-level Chinese students. It also should be noted that whether this implication are applicable for teachers in other contexts of teaching and learning English as a second language is in need of investigation.

12.3.4 Technological implications for software developers

The Levels of Information Processing Model suggests that deeper processing of information can help readers to better comprehend a text. That means, the interaction between the reader and the text plays a substantial role the process of reading. Software developers should consider how to develop user friendly software to support students to read academic texts. Academic text has its own features which are different from entertainment texts. The variables of length, vocabulary, complex sentence structure and grammar should be taken into account when presenting the text on screen.

Software developers should improve the technical features to reduce the mismatch between the medium of screen and the traditional medium of print. This would allow the availability and convivence readers to make annotations on and transfer reading strategies more easily.

Software designers should be concerned with the structure and presentation of texts that are displayed on digital devices (e.g. laptops or tablets) to ensure their meaning is clear to readers. The structure refers to as the visual organization of table of contents (e-book), the list of subtitles (e-journal), the text and illustrations; the presentation refers to as the issues about typeface, type size and the use of space between lines and paragraphs. In thinking about the reading needs of specific reader group as university students in higher education, the text design might be influenced by particular constrains. In terms of the structure, taking reading academic journal articles as an example, the heading and sub-headings play a key role for readers to navigate within a lengthy article, so the designers are concerned with ensuring the list of sub-headings are visibly hierarchically presented while reading and each sub-heading is hyper-linked with its contents (forward and backward). In addition, the text and the illustrations (figures and tables) are ensured to presented on the same page or double-page spread (Waller, 2012) to lessen the disruptions by extra page top and bottom space. This may help, to some degree, to reduce cognitive cost in locating target information by decreasing the chances of scrolling and page turning.

In terms of the presentation, designers should take the typeface, type size and the use of space into account. For example, reading a textbook requires students learn brand new discipline knowledge from an information-loaded text, sustained reading is easy to cause readers' fatigue and distractions so designers are concerned with choice of typeface (e.g., italic, bold or colours) to increase letter distinctive and make text comfortable and easy to read on digital devices. Another issue to be considered is with the changes in type size and line spaces. For instances, readers often need to accommodate the text by increasing or decreasing the scale (as with fixed-layouts PDFs) due to the original character set. When increasing the scale, fewer words may appear on a line, leading the vertical margin spacing to be reduced and the horizontal line spacing to be increased; consequently, the content may be extended over the current screen. Dealing with the split text contents is most likely to increase cognitive demand and impede reading process. Indeed, these issues are dependent on the size of the digital devices (e.g., laptop, tablet or Kindle) and its software. Nevertheless, technology designers is imperative to take issues about how to optimise the spaces between words and lines (Reynolds & Walker, 2004) and how to control the spaces between the lines and length of lines (Reynolds, Walker & Duncan, 2006) into careful consideration to develop a less disruptive and more continuous screen visual experience for university students in reading academic materials.

12.4 Limitations

This thesis has always endeavoured to ensure a rigorous and systematic study, however, there are limitations relating to the data sample, data collection and data analysis. It should be emphasized that this study has been concerned with a quantitative examination with supplementary qualitative investigation. First, the statistical tests normally require a larger sample size to ensure a representative distribution of the population, however, the data sample of this comparative study on reading comprehension performance between screen and print is limited due to time frame and practical constraints. Therefore, the results should be treated cautiously when generalized or transferred. Second, this study aims to collect students' comprehension performances from a naturalistic and non-stressful reading environment, however, the presence of the researcher might influence participants in the process of data collection. Lastly, as an L1-Chinese-speaker who learns and uses English as a second language, this thesis inevitably has some language or cultural bias in translating the interview transcript from Chinese into English in the process of data analysis.

12.5 Future research

12.5.1 Text presentation and reading strategy

Previous research (Hou, Wu, & Harrell, 2017) indicates that the more the resemblance between digital and print medium, the less the difference between reading processing and comprehension performances. Viewing reading comprehension as an interaction between the text and the reader, there are two potential directions for future research to explore the gap between print and screen: text presentation the reading strategy. First, existing literature has shown that if the texts presented on screen and print have more similar characteristics, such as text display, text layout, one full page turning and restricted scrolling, the less disparity was shown between the comprehension performances across two medium. Still, one should note that these findings were on a basis of narrative texts and thus future research should also examine the expository texts in short or at length and enable the text presentation on two medium as identical as possible.

Second, future research should focus on the patterns of strategies and types of annotations related to academic reading in digital context. Though there is an extensive research literature on reading strategies in print, the investigation of the strategies in digital reading is relatively limited. In one study examining the usefulness of note-taking strategy while reading print and digital text for scientific discipline texts, Fiorella and Mayer (2017) found participants used more spatial strategies such as creating maps or drawings on print medium but they tended to use more verbal strategies such as words on laptop; but the spatial strategies (maps or drawings) significantly predicted learning outcomes in reading from paper rather than from laptop. More research is needed to replicate this finding in other disciplines and further examine what kind of print reading strategies can be equally effective in digital medium as well as what kind of reading strategies can be successfully transferred from a printed context to screen context. For example, future research may examine the latest emerging reading tablet iPad with an e-pencil and APPs (e.g., Notability and Marginnote) which have embedded more annotation features, such as highlighting, note-taking, creating mind map and drawings, so as to enable readers to write spontaneously on tablet as much like as paper. When the gaps between text presentation and strategy usage being reduced with the technological developers, it would be interesting to see whether the reading comprehension performances between print and screen tend to be similar, and whether onscreen comprehension remains inferior or superior compared to paper.

12.5.2 Reading preferences and habits

The current study shows that the reliance on printed texts may decrease with the time spent in an Anglophone academic context but the preference for digital texts may grow for some specific reading purposes. As new generations gain more experience in learning from digital texts and perhaps sufficient training of new technologies for digital reading, the picture portrayed in this study regarding the inferior comprehension for digital texts might change. Longitudinal studies are required to examine this hypothesis as the use of digital texts becomes more convenient and commonplace in an English-speaking academic context.

12.5.3 Metacognition

Previous research (e.g., Ackerman and Goldsmith, 2011; Ben-Yehudah & Eshet-Alkalai, 2018; Young, 2014) agree on the lower metacognitive accuracy of onscreen reading. It is suggested that the readers on screen answer comprehension questions more slowly and less accurately than printed readers due to an uncertainty, unrealistic confidence and even “over-confidence” (Ben-Yehudah & Eshet-Alkalai, 2018) in their level of comprehension (e.g., Sidi et al, 2015). However, the findings in current study did not fully support such accounts of differences in metacognitive processing between digital and print readers. The qualitative data from this study did suggest some uncertainty during the screen reading process, particularly in the case of L2-English-speaking students. Yet, the participants in this study indeed experienced greater feeling of uncertainty and spent more time on screen but achieved equivalent performance when answering some literal questions (e.g., recognition and recalling of details, getting main ideas. Interestingly, the L2-English-speaking students appeared to lack confidence in their performance on both print and screen, and they even held an underconfident view of their screen performance, by contrast of overconfidence experienced in Ackerman and Goldsmith (2011) research. A possible explanation for this difference is that L2-English-speakers read in their second language and thus they need to overcome their linguistic deficiency first. Then they perceive themselves to face a more challenging situation when reading on screen; but in fact readers on screen located textual information for answering literal questions in the same way as the readers on print. Thus, there seems to be incongruity between the comprehension performance and metacognitive judgements that are made with respect to specific comprehension tasks. Additional studies are required to examine the metacognitive monitoring between print and screen when reading in a second language, and further investigate the relationship between the reading medium, metacognitive accuracy and level of confidence.

12.5.4 Single-text and multiple text, linear text and hyperlinked

The theoretical and empirical models that inform the current research deal with the nature of single and linear text comprehension. The text used in this experiment excluded any hyperlinks because of the purpose of this study. However, hyperlinked texts are commonly seen in the academic study. A typical example is reading an journal article with in-text hyperlinked references or a text-book with external websites. This current research did not take these types of texts into account. In addition, when reading to write an essay, students are most likely to read several texts at the same time, on print or on screen. There are also well-articulated models of reading and learning when multiple text are implicated (Bråten & Strømsø, 2011; Rouet & Britt, 2011). This is another area for future research as to the effects of the medium of text delivery on reading comprehension for multiple texts.

12.5.5 Working memory

Future research should consider the use of language in digit WM span tasks. This WMC span tasks (FDR and BDR) for L1-Chinese-speaking students were conducted in their L1 Chinese. This is because it is assumed that the differences between L1 and L2 span tasks are less significant in the case of proficient L2 users (Service, Simola, Metsaenheimo & Maury, 2002; Van den Noort, Bosch & Hugdahl, 2006; Walter, 2004), which is the case for university-level participants in this study. However, it should be noted that the digit span can vary from one language to another for participants who speak different language. For example, recent research by Baddeley (2018) has found an average digit span of 6.7 for English students compared with 8.3 for Chinese. The author argues that in Chinese “it appears to be possible to pack more than one digit into a single syllable” (Baddeley, 2018, p. 159). Therefore, future research should replicate the digit span tasks in this study in both L1 and L2 to see if there are any differences. Furthermore, the use of L1 or L2 in WMC span tasks should be given careful considerations and serve the research purpose when testing participants who speak a different language. Lastly, this study has found that students with different L1s may use different strategies for answering comprehension questions, which could in turn differentially draw on working memory. For this reason, it would be important to note that future studies on WM employ a sufficient sample size that allows cross-language analysis of WM or comprehension performance.

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive Regulation of Text Learning: On Screen Versus on Paper. *Journal of Experimental Psychology-Applied, 17*(1), 18-32.

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior, 28*(5), 1816-1828.

Aebersold, J. A., & Field, M. L. (1998). *From reader to reading teacher: Issues and strategies for second language classrooms*: Ernst Klett Sprachen.

Al-Issa, A. (2006). Schema theory and L2 reading comprehension: Implications for teaching. *Journal of College Teaching & Learning (TLC), 3*(7).

Albashtawi, A. H., Jaganathan, P., & Singh, M. (2016). Linguistic Knowledge Aspects in Academic Reading: Challenges and Deployed Strategies by English-Major Undergraduates at a Jordanian Institution of Higher Education. *Higher Education Studies, 6*(3), 61.

Alderson, J. C. (2000). Assessing reading. Cambridge language assessment series. In: Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*: Ernst Klett Sprachen.

Ali, A. Z. M., Wahid, R., Samsudin, K., & Idris, M. Z. (2013). Reading on the Computer Screen: Does Font Type has Effects on Web Text Readability? *International Education Studies, 6*(3), 26.

Allen, L., Cipielewski, J., & Stanovich, K. E. (1992). Multiple indicators of children's reading habits and attitudes: Construct validity and cognitive correlates. *Journal of Educational Psychology, 84*(4), 489.

Allen, R. J., Havelka, J., Falcon, T., Evans, S., & Darling, S. (2015). Modality specificity and integration in working memory: Insights from visuospatial bootstrapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 820.

Alloway, T. P., & Gathercole, S. E. (2006). *Working memory and neurodevelopmental disorders*: Psychology Press.

Alptekin, C. (2006). Cultural familiarity in inferential and literal comprehension in L2 reading. *System, 34*(4), 494-508.

Alptekin, C., & Ercetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *JOURNAL OF RESEARCH IN READING, 33*(2), 206-219.

Alshan, L. M. (1964). *A factor analytic study of items used in the measurement of some fundamental factors of reading comprehension.* Teachers College, Columbia University,

Anderson, N. J. (1999). *Exploring second language reading : issues and strategies*. Boston, Mass. ;: Heinle & Heinle.

Anderson, N. J., & Cheng, X. (2004). *Exploring second language reading: Issues and strategies*: Foreign Language Teaching and Research Press.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. *Handbook of reading research, 1*, 255-291.

Applegate, M. D., Quinn, K. B., & Applegate, A. J. (2002). Levels of thinking required by comprehension questions in informal reading inventories. *The Reading Teacher, 56*(2), 174-180.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89-195): Elsevier.

Axelrod, R. B., & Cooper, C. R. (1996). The concise guide to writing. New York: St. In: Martins Press.

Bachman, L., & Palmer, A. (2010). Language assessment in practice. In: Oxford: Oxford University Press.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556-559.

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

Baddeley, A. (2007). *Working memory, thought, and action* (Vol. 45): OUP Oxford.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual review of psychology, 63*, 1-29.

Baddeley, A. (2018). *Working memories: Postmen, divers and the cognitive revolution*: Routledge.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A. D. (2015). Working memory in second language learning. *Working memory in second language acquisition and processing*, 17-28.

Bando, R., Gallego, F., Gertler, P., & Romero, D. (2016). *Books or Laptops? The Cost-Effectiveness of Shifting from Printed to Digital Delivery of Educational Content* (0898-2937). Retrieved from

Barrett, T. C. (1976). Taxonomy of reading comprehension. *Reading, 360*.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26*(3), 349-379.

Bayliss, D. M., Jarrold, C., Baddeley, A. D., & Gunn, D. M. (2005). The relationship between short-term memory and working memory: Complex span made simple? *Memory, 13*(3-4), 414-421.

Bazeley, P., & Jackson, K. (2013). *Qualitative data analysis with NVivo* (2nd / by Pat Bazeley, Kristi Jackson. ed.). London: SAGE.

Ben-Yehudah, G., & Eshet-Alkalai, Y. (2018). The contribution of text-highlighting to comprehension: A comparison of print and digital reading. *Journal of Educational Multimedia and Hypermedia, 27*(2), 153-178.

Best, R. M., Floyd, R. G., & Mcnamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology, 29*(2), 137-164.

Betzner, A. E. (2008). Pragmatic and dialectic mixed method approaches: An empirical comparison.

Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes, 45*(2), 122-159.

Bowey, J. A., & Rutherford, J. (2007). Imbalanced word-reading profiles in eighth-graders. *Journal of Experimental Child Psychology, 96*(3), 169-196.

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British journal of educational psychology, 75*(2), 189-201.

Bråten, I., & Strømsø, H. (2011). Measuring strategic processing when students read multiple texts. *Metacognition and Learning, 6*(2), 111-130.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Britt, M. A., Rouet, J.-F., & Durik, A. M. (2017). *Literacy beyond text comprehension: A theory of purposeful reading*: Taylor & Francis.

Broadbent, D. E. (1958). *Perception and communication*: Pergamon Press.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*(5-6), 489-503.

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*(6), 850-859.

Carnine, D., Silbert, J., Kame'enui, E., & Tarver, S. (2010). Direct instruction reading Boston: Merrill.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*(3), 386-404.

Cem, A., & Gülcan, E. (2015). Eye movements in reading span tasks to working memory functions and second language reading. *Eurasian Journal of Applied Linguistics, 1*(2), 35-56.

Ceran, D. (2015). Investigation of High School Students' Reading Compherension Levels According to Various Variables. *Educational Research and Reviews, 10*(11), 1524.

Chen, D.-W., & Catrambone, R. (2015). *Paper vs. screen: Effects on reading comprehension, metacognition, and reader behavior.* Paper presented at the Proceedings of the human factors and ergonomics society annual meeting.

Chen, G., Cheng, W., Chang, T.-W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education, 1*(2), 213-225.

Cho, B. Y., & Afflerbach, P. (2015). Reading on the Internet: Realizing and constructing potential texts. *Journal of Adolescent & Adult Literacy, 58*(6), 504-517.

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *JOURNAL OF RESEARCH IN READING, 42*(2), 288-325.

Clymer, T. (1968). Current Conceptions of Reading. *Innovations and Change in Reading Instruction, ed. Helen M. Robinson. The Sixty-Seventh Yearbook of the National Society for the Study of Education, Part,* 2, 7-29.

Connell, C., Bayliss, L., & Farmer, W. (2012). Effects of eBook readers and tablet computers on reading comprehension. *International Journal of Instructional Media, 39*(2).

Conway, A. R., 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.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science, 19*(1), 51-57.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Craik, F. I., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 599-607.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: qualitative, quantitative, and mixed methods approaches* (5th, International student ed.). Thousand Oaks, Calif;London;: Sage Publications.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health, 2013*, 541-545.

Crowther, D., & Lancaster, G. (2008). Research methods in education. In: Oxford: Elsevier Limited.

Crystal, D. (2007). *How language works*: Penguin UK.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466.

Daniel, D. B., & Woody, W. D. (2013). E-textbooks at what cost? Performance and use of electronic v. print texts. *Computers & Education, 62*, 18-23.

Davis, D. S., & Neitzel, C. (2012). Collaborative sense-making in print and digital text environments. *Reading and Writing, 25*(4), 831-856.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, 499-545.

Davis, L. C., Rane, S., & Hiscock, M. (2013). Serial recall of visuospatial and verbal information with and without material-specific interference: Implications for contemporary models of working memory. *Memory, 21*(7), 778-797.

Deacon, S. H., Wade-Woolley, L., & Kirby, J. (2007). Crossover: The Role of Morphological Awareness in French Immersion Children's Reading. *Developmental Psychology, 43*(3), 732-746.

Dehn, M. J. (2008). *Working memory and academic learning: assessment and intervention*. Hoboken, N.J: Wiley.

Dillon, A. (1992). Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics, 35*(10), 1297-1326.

Dole, J. A., Duffy, G. G., Roehler, L. R., & Pearson, P. D. (1991). Moving From the Old to the New: Research on Reading Comprehension Instruction. *Review of Educational Research, 61*(2), 239-264.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (Vol. 326): John Wiley & Sons.

Duff, S. C., & Logie, R. H. (2001). Processing and storage in working memory span. *The Quarterly Journal of Experimental Psychology Section A, 54*(1), 31-48.

Dundar, H., & Akcayır, M. (2012). Tablet vs. paper: The effect on learners' reading performance. *International Electronic Journal of Elementary Education, 4*(3), 441-450.

Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods, 39*(2), 212-223.

Eden, S., & Eshet-Alkalai, Y. (2013). The effect of format on performance: Editing text in print versus digital formats. *British Journal of Educational Technology, 44*(5), 846-856.

Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: a test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 972.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—principles and practices. *Health services research, 48*(6pt2), 2134-2156.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*: Sage.

Fiorella, L., & Mayer, R. E. (2017). Spontaneous spatial strategy use in learning from scientific text. *Contemporary Educational Psychology, 49*, 66-79.

Fisher, D., Lapp, D., & Wood, K. (2011). Reading for details in online and printed text: A prerequisite for deep reading. *Middle School Journal, 42*(3), 58-63.

Flood, D. L. J. (1978). Teaching Reading to Every Child *Journal of Teacher Education, 31*(5).

Gagné, E., Yekovich, C., & Yekovich, F. (1993). The cognitive psychology of school subjects. In: New York: HarperCollins.

Gandy, S. E. (2013). Informal Reading Inventories and ELL Students. *Reading & Writing Quarterly, 29*(3), 271-287.

Garland, K. J., & Noyes, J. M. (2004). Computer experience: a poor predictor of computer attitudes. *Computers in Human Behavior, 20*(6), 823-840.

Garland, K. J., & Noyes, J. M. J. C. i. H. B. (2004). Computer experience: a poor predictor of computer attitudes. *20*(6), 823-840.

Gathercole, S. E., Durling, E., Evans, M., Jeffcock, S., & Stone, S. (2008). Working memory abilities and children's performance in laboratory analogues of classroom activities. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 22*(8), 1019-1037.

Gernsbacher, M. A., Robertson, R. R., Palladino, P., & Werner, N. K. (2004). Managing mental representations during narrative comprehension. *Discourse Processes, 37*(2), 145-164.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71*(2), 279-320.

Goddard, N. (2012). Chapter 5 - Psychology. In P. Wright, J. Stern, & M. Phelan (Eds.), *Core Psychiatry (Third Edition)* (pp. 63-82). Oxford: W.B. Saunders.

Goldman, S. R. (2015). Reading and the Web: Broadening the need for complex comprehension. In *Reading at a crossroads? Disjunctures and continuities in current conceptions and practices* (pp. 89-103).

Goodwin, A. P., Cho, S.-J., Reynolds, D., Brady, K., & Salas, J. (2019). Digital Versus Paper Reading Processes and Links to Comprehension for Middle School Students. *American Educational Research Journal*, 0002831219890300.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and special education, 7*(1), 6-10.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*: Ernst Klett Sprachen.

Grabe, W. P., & Stoller, F. L. (2013). *Teaching and researching: Reading*: Routledge.

Graesser, A., Léon, J., & Otero, J. (2002). Introduction to the psychology of scientific text. *The psychology of scence texte comprehension. Mahwah: New Jersey: LEA*.

Graesser, A. C. (2007a). An introduction to strategic reading comprehension. In *Reading comprehension strategies: Theories, interventions, and technologies* (Vol. 2579, pp. 3-26).

Graesser, A. C. (2007b). An introduction to strategic reading comprehension. *Reading comprehension strategies: Theories, interventions, and technologies, 2579*, 3-26.

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., Singer, M., & Trabasso, T. (1994). Constructing Inferences During Narrative Text Comprehension. *Psychological Review, 101*(3), 371-395.

Green, T. D., Perera, R. A., Dance, L. A., & Myers, E. A. (2010). Impact of Presentation Mode on Recall of Written Text and Numerical Information: Hard Copy Versus Electronic. *North American Journal of Psychology, 12*(2).

Gygax, P., Garnham, A., & Oakhill, J. (2004). Inferring characters' emotional states: Can readers infer specific emotions? *Language and Cognitive Processes, 19*(5), 613-639.

Harvey, H., & Walker, R. (2018). Reading comprehension and its relationship with working memory capacity when reading horizontally scrolling text. *Quarterly Journal of Experimental Psychology, 71*(9), 1887-1897.

Henry, L. (2011). *The development of working memory in children*: Sage.

Herber, H. L. (1970). *Teaching reading in content areas*. Englewood Cliffs, N.J: Prentice-Hall.

Heriyawati, D. F., Saukah, A., & Widiati, U. (2018). Working memory capacity, content familiarity and university EFL students' reading comprehension. *Indonesian Journal of Applied Linguistics, 8*(1).

Hermena, E. W., Sheen, M., AlJassmi, M., AlFalasi, K., AlMatroushi, M., & Jordan, T. R. (2017). Reading Rate and Comprehension for Text Presented on Tablet and Paper: Evidence from Arabic. *FRONTIERS IN PSYCHOLOGY, 8*(257).

Hervey, S., & Higgins, I. (2002). *Thinking translation: A course in translation method: French to English*: Routledge.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160.

Hoover, W. A., & Tunmer, W. E. (1993). The components of reading.

Hou, J., Rashid, J., & Lee, K. M. (2017). Cognitive map or medium materiality? Reading on paper and screen. *Computers in Human Behavior, 67*, 84-94.

Hou, J., Wu, Y., & Harrell, E. (2017). Reading on Paper and Screen among Senior Adults: Cognitive Map and Technophobia. *FRONTIERS IN PSYCHOLOGY, 8*.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105*(2), 302.

Hulme, C., & Mackenzie, S. (1992). Working Memory and Severe Learning Difficulties. In: Lawrence Erlbaum Associates.

Jude, W. I., & Ajayi, O. (2012). Literal level of student's comprehension in Nigeria: A means for growing a new generation African scholars. *Journal of Education and Practice, 3*(7), 120-129.

Jung, J. (2018). Effects of task complexity and working memory capacity on L2 reading comprehension. *System, 74*, 21-37.

Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a Research Paradigm and Its Implications for Social Work Research. *Social Sciences, 8*(9), 255.

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.

Kelleher, J. D., & Dobnik, S. (2019). Referring to the recently seen: reference and perceptual memory in situated dialog. *arXiv preprint arXiv:1903.09866*.

Kerr, M. A., & Symons, S. E. (2006). Computerized Presentation of Text: Effects on Children's Reading of Informational Material. *Reading and Writing, 19*(1), 1-19.

Kerr, M. M., & Frese, K. M. (2017). Reading to Learn or Learning to Read? Engaging College Students in Course Readings. *College Teaching, 65*(1), 28-31.

Kim, H., & Kim, J. (2013). Reading from an LCD monitor versus paper: Teenagers' reading performance. *International Journal of Research Studies in Educational Technology, 2*(1), 1-10.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy, 65*(23), 2276-2284.

Kintsch, E., & Kintsch, W. (2005). Comprehension. In *Children's reading comprehension and assessment* (pp. 71-92): Routledge.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review, 95*(2), 163.

Kintsch, W., & Rawson, K. (2005). Comprehension. The Science of Reading. *A handbook*.

Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? *Literacy, 42*(2), 75-82.

Kispal, A. (2008). *Effective Teaching of Inference Skills for Reading. Literature Review. Research Report DCSF-RR031*: ERIC.

Kretzschmar, F., Pleimling, D., Hosemann, J., Füssel, S., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Subjective impressions do not mirror online reading effort: Concurrent EEG-eyetracking evidence from the reading of books and digital media. *PloS one, 8*(2), e56178.

Kuzborska, I. (2015). Perspective taking in second language academic reading: A longitudinal study of international students' reading practices. *Journal of English for Academic Purposes, 20*, 149-161.

Lee, D.-S., Ko, Y.-H., Shen, I.-H., & Chao, C.-Y. J. D. (2011). Effect of light source, ambient illumination, character size and interline spacing on visual performance and visual fatigue with electronic paper displays. *32*(1), 1-7.

Lenhard, W., Schroeders, U., & Lenhard, A. (2017). Equivalence of Screen Versus Print Reading Comprehension Depends on Task Complexity and Proficiency. *Discourse Processes, 54*(5-6), 427-445.

Letton, M. C. (1958). *Individual differences in interpretive responses in reading poetry at the ninth-grade level.* University of Chicago, Department of Education,
Li, X.-h., Wu, J., & Wang, W.-h. (2007). Analysis of schema theory and its influence on reading. *US-China foreign language, 5*(11), 18-21.
Logie, R. H., & Logie, R. (1995). *Visuo-spatial working memory*: Psychology Press.
Luke, S. G., Henderson, J. M., & Ferreira, F. (2015). Children's eye-movements during reading reflect the quality of lexical representations: An individual differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1675.
Mangen, A. (2006). *New narrative pleasures? A cognitive-phenomenological study of the experience of reading digital narrative fictions*: Det historisk-filosofiske fakultet.
Mangen, A., & Kuiken, D. (2014). Lost in an iPad: Narrative engagement on paper and tablet. *Scientific Study of Literature, 4*(2), 150-177.
Mangen, A., & Schilhab, T. (2012). An embodied view of reading: Theoretical considerations, empirical findings, and educational implications. *Skriv*, 285-300.
Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research, 58*, 61-68.
Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, Computer Screens, or Paper: Does Reading Comprehension Change Across Media Platforms? *Applied Cognitive Psychology, 27*(4), 512-519.
Mayes, D. K., Sims, V. K., & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics, 28*(6), 367-378.
McCormick, S. (1992). Disabled readers' erroneous responses to inferential comprehension questions: Description and analysis. *Reading Research Quarterly*, 55-77.
McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*(3), 440.
McNamara, D. S. (2012). *Reading comprehension strategies: Theories, interventions, and technologies*: Psychology Press.
McNamara, T., Miller, D. L., & Bransford, J. D. (1991). Mental models and reading comprehension.
Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.
Mraz, M., Rickelman, R. J., & Vacca, R. T. (2009). Content-area reading. *Literacy instruction for adolescents: Research-based practice, 77*.
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*(1), 85-101.
Neijens, P. C., & Voorveld, H. A. (2018). Digital replica editions versus printed newspapers: Different reading styles? Different recall? *New media & society, 20*(2), 760-776.

Norman, E., & Furnes, B. (2016). The relationship between metacognitive experiences and learning: Is there a difference between digital and non-digital study media? *Computers in Human Behavior, 54*, 301-309.

Ortlieb, E., Sargent, S., & Moreland, M. (2014). Evaluating the efficacy of using a digital reading environment to improve reading comprehension within a reading clinic. *Reading Psychology, 35*(5), 397-421.

Pearson, P. D. (2009). The roots of reading comprehension instruction. *Comprehension Across the Curriculum: Perspectives and Practices in K-12*, 279-314.

Pearson, P. D., & Johnson, D. D. (1979). Teaching Reading Comprehension. In (Vol. 68, pp. 88).

Perfetti, C., & Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading, 18*(1), 22-37.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The Acquisition of Reading Comprehension Skill.

Pettit, N. T., & Cockriel, I. W. (1974). A factor study of the literal reading comprehension test and the inferential reading comprehension test. *Journal of Reading Behavior, 6*(1), 63-75.

Plaut, D. C. (2005). Connectionist approaches to reading. *The science of reading: A handbook*, 24-38.

Pokharel, P. K. (2018). Learning to Read and Reading to Learn in English. *Journal of NELTA Surkhet, 5*, 75-81.

Porion, A., Aparicio, X., Megalakaki, O., Robert, A., & Baccino, T. (2016). The impact of paper-based versus computerized presentation on text comprehension and memorization. *Computers in Human Behavior, 54*, 569-576.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*: Routledge.

Reynolds, L., & Walker, S. (2004). 'You can't see what the words say': word spacing and letter spacing in children's reading books. *JOURNAL OF RESEARCH IN READING, 27*(1), 87-98.

Reynolds, L., Walker, S., & Duncan, A. (2006). Children's Responses to Line Spacing in Early Reading Books or'Holes to tell which line you're on'. *Visible Language, 40*(3), 246.

Rockinson-Szapkiw, A. J., Courduff, J., Carter, K., & Bennett, D. (2013). Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education, 63*, 259-266.

Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. *Text relevance and learning from text*, 19-52.

Rupley, W. H., & Blair, T. R. (1983). *Reading diagnosis and direct instruction: a guide for the classroom*: Houghton Mifflin School.

Sally, W. (2019). Chinese students' applications to UK universities up by 30%. *The Guardian*. Retrieved from https://www.theguardian.com/education/2019/jul/11/chinese-students-applications-to-uk-universities-up-by-30

Samuels, S. J. (1994). Toward a theory of automatic information processing in reading, revisited.

Sanchez, C. A., & Wiley, J. (2009). To scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors, 51*(5), 730-738.

Schunk, D. H., & Zimmerman, B. J. (1994). *Self-regulation of learning and performance: Issues and educational applications*: Lawrence Erlbaum Associates, Inc.

Sellen, A. (1997). A comparison of reading paper and on-line documents.

Service, E., Simola, M., Metsänheimo, O., & Maury, S. (2002). Bilingual working memory span is affected by language skill. *European Journal of Cognitive Psychology, 14*(3), 383-408.

Shahriza Abdul Karim, N., & Hasan, A. (2007). Reading habits and attitude in the digital age: Analysis of gender and academic program differences in Malaysia. *The Electronic Library, 25*(3), 285-298.

Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology, 22*(2), 261-273.

Shenoy, P., & Aithal, P. (2016). A Study on History of Paper and possible Paper Free World. *International Journal of Management, IT and Engineering, 6*(1), 337-355.

Sidi, Y., Ophir, Y., & Ackerman, R. (2016). Generalizing screen inferiority - does the medium, screen versus paper, affect performance even with brief tasks? *Metacognition and Learning, 11*(1), 15-33.

Siegenthaler, E., Wurtz, P., Bergamin, P., & Groner, R. (2011). Comparing reading processes on e-ink displays and print. *Displays, 32*(5), 268-273.

Singer, L. M., & Alexander, P. A. (2017a). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education, 85*(1), 155-172.

Singer, L. M., & Alexander, P. A. (2017b). Reading on Paper and Digitally: What the Past Decades of Empirical Research Reveal. *Review of Educational Research, 87*(6), 1007-1041.

Smith, R. J., & Barrett, T. C. (1979). *Teaching reading in the middle grades*: Addison Wesley Publishing Company.

Smyth, M. M., & Pendleton, L. R. (1989). Working memory for movements. *The quarterly Journal of experimental Psychology, 41*(2), 235-250.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*: Rand Corporation.

Snowling, M. J., & Hulme, C. (2008). *The science of reading: A handbook* (Vol. 9): John Wiley & Sons.

Son, L. K. (2007). Introduction: A metacognition bridge. *European Journal of Cognitive Psychology, 19*(4-5), 481-493.

Stakhnevich, J. (2002). Reading on the Web: Implications for ESL professionals. *The Reading Matrix, 2*(2).

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 402-433.

Støle, H., Mangen, A., Frønes, T. S., & Thomson, J. (2018). Digitisation of reading assessment. *Learning to Read in a Digital World, 17*, 205.

Stoltzfus, E. R., Hasher, L., & Zacks, R. T. (1996). Working memory and aging: Current status of the inhibitory view. *Working memory and human cognition*, 66-88.

Stoop, J., Kreutzer, P., & G. Kircz, J. (2013). Reading and learning from screens versus print: a study in changing habits: Part 2–comparing different text structures on paper and on screen. *New Library World, 114*(9/10), 371-383.

Sweet, A. P. (2005). Assessment of reading comprehension: The RAND reading study group vision. In *Children's reading comprehension and assessment* (pp. 21-30): Routledge.

Tashakkori, A., & Teddlie, C. (2008). Quality of inferences in mixed methods research: Calling for an integrative framework. *Advances in mixed methods research*, 101-119.

Taylor, A. K. (2011). Students learn equally well from digital as from paperbound texts. *Teaching of psychology, 38*(4), 278-281.

Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of memory and language, 39*(2), 195-217.

Trakhman, L. M. S., Alexander, P. A., & Silverman, A. B. (2018). Profiling reading in print and digital mediums. *Learning and Instruction, 57*, 5-17.

Tun, P. A. (1989). Age differences in processing expository and narrative text. *Journal of Gerontology, 44*(1), P9-P15.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language, 28*(2), 127-154.

Urquhart, A. H., & Weir, C. J. (2014). *Reading in a second language: Process, product and practice*: Routledge.

Usó, E., & Ruiz-Madrid, M. N. (2009). Reading printed versus online texts. A study of EFL learners strategic reading behavior. *International Journal of English Studies, 9*(2).

Vacca, R. (1988). Reading and Learning to Read. *Reading Psychology, 9*(1), 80-84.

Van den Broek, P., Bohn-Gettler, C. M., Kendeou, P., Carlson, S., & White, M. J. (2011). When a reader meets a text: The role of standards of coherence in reading comprehension.

Van den Noort, M. W., Bosch, P., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist, 11*(4), 289-296.

Vicari, S., Bellucci, S., & Carlesimo, G. A. (2006). Evidence from two genetic syndromes for the independence of spatial and visual working memory. *Developmental medicine and child neurology, 48*(2), 126-131.

Walter, C. (2004). Transfer of reading comprehension skills to L2 is linked to mental representations of text and to L2 working memory. *Applied Linguistics, 25*(3), 315-339.

Wang, S., & Bai, X. (2016). University Students Awareness, Usage and Attitude Towards E-books: Experience from China. *The Journal of Academic Librarianship, 42*(3), 247-258.

Wästlund, E. (2007). *Experimental studies of human-computer interaction: Working memory and mental workload in complex cognition*: Department of Psychology.

Wästlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior, 21*(2), 377-394.

Waterman, A. H., Atkinson, A. L., Aslam, S. S., Holmes, J., Jaroslawska, A., & Allen, R. J. (2017). Do actions speak louder than words? Examining children's ability to follow instructions. *Memory & cognition, 45*(6), 877.

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly journal* *of experimental psychology. A, Human experimental psychology, 49*(1), 51-79.

Weaver, C. A., & Kintsch, W. (1991). Expository text. *Handbook of reading research, 2*, 230-245.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528): John Wiley & Sons.

Wen, Z., Mota, M. B., & McNeill, A. (2015). *Working memory in second language acquisition and processing* (Vol. 87;87.;). Bristol;Buffalo;: Multilingual Matters.

Young, J. (2014). A study of print and computer-based reading to measure and compare rates of comprehension and retention. *New Library World, 115*(7/8), 376-393.

Yu, G. (2005). *Towards a model of using summarization tasks as a measure of reading comprehension.* University of Bristol,

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review, 27*(1), 69-78.

Zeelenberg, R., Pecher, D., Shiffrin, R. M., & Raaijmakers, J. G. (2003). Semantic context effects and priming in word association. *Psychonomic Bulletin & Review, 10*(3), 653-660.

Zhou, Y. L., McBride-Chang, C., & Wong, N. (2014). What is the role of visual skills in learning to read? *FRONTIERS IN PSYCHOLOGY, 5*, 776.

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E., & Curiel, J. M. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading, 2*(3), 199-220.

Appendix A Ethical form

The Secretariat
Level 11, Worsley Building
University of Leeds
Leeds, LS2 9JT
Tel: 0113 3434873
Email: ResearchEthics@leeds.ac.uk

UNIVERSITY OF LEEDS

Handan Lu

School of Education

University of Leeds

Leeds, LS2 9JT

ESSL, Environment and LUBS (AREA) Faculty Research Ethics Committee

University of Leeds

31 December 2020

Dear Handan

Title of study:	**An investigation into differences between screen reading and print reading comprehension in the academic study**
Ethics reference:	**AREA 17-020**

The above project was reviewed by the AREA Faculty Research Ethics Committee at its virtual meeting of 28th September 2017. The following documentation was considered:

Document	Version	Date
AREA 17-020 1. Ethical Review Form.docx	2	12/09/17
AREA 17-020 2. Information sheet.docx	2	12/09/17
AREA 17-020 3. Consent-form.docx	2	12/09/17

On the basis of the information provided, the Committee requested further information/ clarification of the following matters before approval can be granted:

General comments		
The submission addresses the substantive ethical issues of the research proposal, but there are some entries that need clarification. These are explained below.		
Application section	**Comment**	**Response required/ amended application required/ for consideration**
3.1 and throughout	It is recommended that you avoid the term ‘native speaker’, which is highly politically charged and contested. Please consider alternative descriptions (e.g. L1-English speaker).	For consideration (Yes. All “English-native-speaker” has been changed to “L1-English speaker”.)

3.1	What is meant by the term 'merits'? Presumably this refers to 'advantages' as opposed to disadvantages.	Yes. It means "advantages". "Merits" has been changed to "advantages". (p3)
4.6	The assertion that the research will enhance student learning needs to be substantiated by explaining how this will be achieved, e.g. will teachers and students be provided with any feedback that will have potential impact on the quality of teaching and learning?	Amended application See 4.6 (marked as red)
7.5	It would be advisable to suggest a point up to PhD submission and/or publication when it would be too late for the data to be withdrawn.	Amended application and information sheet See 7.5 and information sheet (marked as red)
8.3	The statement 'replacement terms or vague descriptions' is far too vague in itself to explain precisely how the data will be anonymised. It therefore needs rephrasing. The reviewers suggest you replace this phrase with a more precise statement on how anonymity will be ensured, e.g. through the use of pseudonyms or by avoiding the publication of contextual	Amended application See 8.3 (marked as red)

	information that could lead to the identification of individuals.	
8.4	It would be better to answer 'No one' rather than N/A.	Amended application See 8.4 (marked as red)
Consent form	Point 6, final sentence – please change 'my records' to 'this data'. 'My records' could potentially mislead participants to think that their personal data will be used.	Amended consent form See consent form

A response should be sent to the Committee which addresses each of these points, and further consideration will be given to your response. Please highlight or use a different colour font to indicate the changes to your application form and supporting documents. Students are strongly advised to discuss their response with their supervisor before it is submitted.

The Committee is not able to approve your application at this stage so you are unable to begin your research. Please do not hesitate to contact us if you have any questions. Advice can also be sought from the Research Ethics Senior Training & Development Officer: http://ris.leeds.ac.uk/EthicsTraining.

Yours sincerely

Jennifer Blaikie

Senior Research Ethics Administrator, the Secretariat

On behalf of Dr Kahryn Hughes, Chair, AREA Faculty Research Ethics Committee

Appendix B Interview question outlines

The interview questions have two parts. Part one is related to the reading process that the participants had undertaken. Part two is related to the reading habits in general in the academic study.

Part one has two separate sets of questions: one for the participants who read on screen and the other for the participants who read on paper.

The questions were based on the researcher's observation notes and reading test results. In Part one, participants were asked to briefly re-read the text and see their answers sheet. *These questions were aimed at understanding participants' strategies used to answer specific questions during the test. In Part two, participants were asked to talk about their general reading strategies and habits.*

Part One

Interview questions for **screen reading** participants

1. Did you find it easy and quick to locate the answers for question number X when reading on laptop screen?
2. Do you think it time-consuming to find answers for question X?
3. Can you answer question number X without turning back to the text on screen?
4. Why did you leave the question number X blank?
5. How did you answer question X if you cannot find the direct answers in the text?
6. If you read a long sentence or a difficult paragraph on screen, such as "…", what would you do? For example, did you annotate (highlight or make notes) on screen?
7. Did you think it's easy and useful to annotate on screen during the test?

8. Did you scroll often when you read on screen during the test?
9. Did you think anything that may distract you from understanding the text, for example, frequent scrolling while reading?
10. What were the most challenges faced when answering questions number X, X and X?

Interview questions for **<u>print reading</u>** participants

1. Did you find it easy and quick to locate the answers for question number X when reading on paper?
2. Do you think it time-consuming to find answers for question X?
3. Can you answer question number X without turning back to the text?
4. Why did you leave the question number X blank?
5. How did you answer question X if you cannot find the direct answers in the text?
6. If you read a long sentence or a difficult paragraph, such as "…", what would you do? For example, did you annotate (highlight or make notes) on paper?
7. If you used annotation, did you think that it could help you understand the text?
8. What were the most challenges faced when answering questions number X, X and X?

Part Two

1. During reading for academic purposes do you typically read electronic or hard copies of materials?
2. For what other purposes do you read on screen?
3. For what other purposes do you read on print?
4. When you read texts on print such as a journal article, do you generally read slower or quicker compared to reading on screen?
5. What might be the reasons that make you read slower (or quicker) on print?

6. If you read on screen, do you read on your laptop, desktop or on other devices? Why is that?
7. Do you think the familiarity of using software affects your decision as to whether to read on screen or print?
8. Do you think your English language proficiency affects your decisions to whether to read on screen or print?
9. If you read in L1, do you typically read on screen or print?
10. If you read in L2, do you typically read on screen or print?

Appendix C Reading Comprehension Assessment - text

CHAPTER 2

The marketing environment

CHAPTER CONTENTS

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- Explain the nature of the business environment, and the relationship between the firm and its environment.
- Understand the problems of dealing with the micro and macro environments.
- Describe the relationship between the elements of the business environment.
- Explain the effects of demographic change on marketing.
- Discuss the nature and sources of competition.
- Explain how technological change can transfer between industries.

Introduction

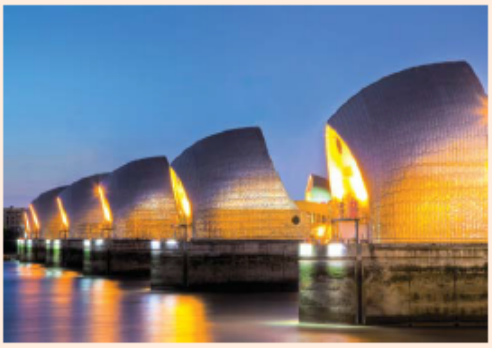
No business operates in a vacuum. Decisions are made within a context of competition, customer characteristics, behaviour of suppliers and distributors, and of course within a legislative and social framework. People working within organisations are contributing to the welfare of society and of each other, and obtaining satisfaction of their own needs in return: this complex network of exchanges results in a better standard of living for everybody.

From a marketing viewpoint, managing the exchange process between the firm and its customers comes highest on the list of priorities, but it would be impossible to carry out this function without considering the effects of customer-based decisions on other people and organisations. A stakeholder is any individual or organisation affected by the firm's activities – neighbours, suppliers, competitors, customers, even governments – and all of these will have some input into marketing decisions, either directly or indirectly.

Some environmental factors are easily controlled by managers within the firm, whereas others cannot be changed and must therefore be accommodated in decision-making. In general, the larger the firm, the greater the control over its environment: on the other hand, large firms often find it difficult to adapt to sudden environmental changes in the way that a small firm might.

In order to assess the impact of different environmental factors, managers first need to classify them.

Preview case study: Costain West Africa



Costain is a major international civil engineering company. Founded in Liverpool in 1865, the company became one the largest British civil engineering companies in the ensuing years. Costain was involved in building the Mulberry harbours used on D-Day, the Channel Tunnel, the Thames Barrier, Hong Kong's airport, and many other large-scale construction projects. At one time Costain operated in 25 countries, but during the 1990s the company contracted and began to concentrate on the UK market, due to the recession that heralded the decade.

Costain West Africa was originally founded in 1948, to take over the Holt construction business. The company became fully independent of Costain UK when the parent company began its partial withdrawal from overseas markets: Costain West Africa is quoted on the Nigeria Stock Exchange, and is the largest construction company in Nigeria, if not in the whole of sub-Saharan Africa.

Any company in the construction business is likely to be affected by recessions and financial crises: major capital projects are frequently put on hold when money runs short, because an organisation can always cope for another year or two without its new headquarters, and the new bridge can always wait – after all, it wasn't always there, was it? So many construction companies face hard times – as do their suppliers and subcontractors. Costain West Africa was affected by the financial crisis of 2008 as much as any other organisation, but managed to survive and even flourish.

Surviving a financial crisis is no mean feat – but Costain West Africa met this sudden shift in the marketing environment with a uniquely African approach.

Macro environment Factors that affect all the firms in an industry.

Micro environment Factors that affect one firm only.

Internal environment Factors that operate within the organisation.

External environment Factors that operate outside the organisation.

Classifying environmental factors

Factors within the environment can be classified in a number of ways. First, the environment can be considered in terms of those elements that affect all firms within the industry (the **macro environment**), as opposed to those elements that affect only the individual firm (the **micro environment**). In general, the macro environment is difficult to influence or control, whereas the micro environment is much more within the firm's control.

The environment can also be classified as internal or external. The **internal environment** comprises those factors that operate within the firm (the corporate culture and history, staff behaviour and attitudes, the firm's capabilities) and the **external environment** comprises those elements that operate outside the firm (competition, government, customers). A problem for firms lies in deciding where

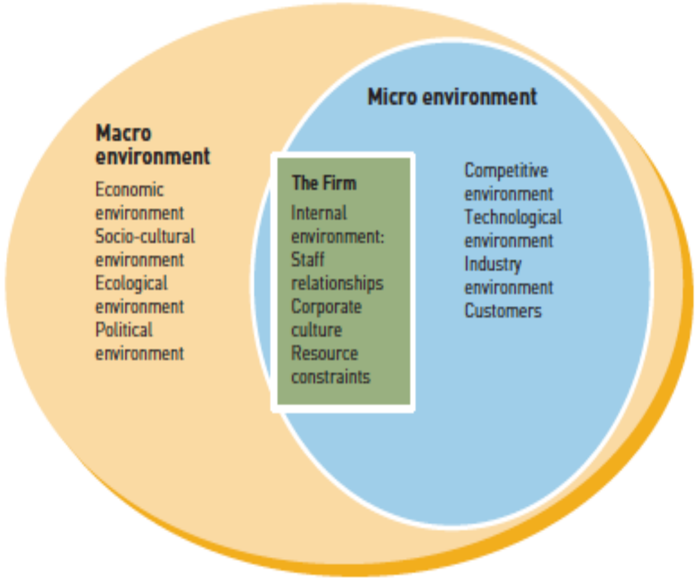


Figure 2.1 Environmental factors

the boundaries lie: for a truly customer-orientated company, customers might be considered as part of the internal environment, for example. Figure 2.1 shows how these factors relate. In effect, the firm operates within a series of layers of environmental factors, each of which has a greater or lesser impact on the firm's marketing policies. As a general rule, the further out the layer is, the more difficult it is for the firm to control what is happening: only the very largest firms have control, or even influence, on the macro environment.

The macro environment

The macro environment comprises those factors which are common to all firms in the industry. In many cases the same factors affect firms in other industries. Government policy, the economic climate and the culture within the countries in which the firms operate are common factors for all firms, but will affect firms differently according to the industries they are in.

In some cases there will be overlap between the micro environment and the macro environment. For example, a very large, global firm operating a subsidiary in a small country might regard the government of the country as part of the micro environment, since it is possible for the firm to control what the government does. This has certainly been the case with major fruit-importing companies operating in Central America. On the other hand, although competitors are usually regarded as part of the micro environment, a firm which is large enough to control an industry might be regarded as part of the macro environment by smaller firms in the same industry.

ECONOMIC ENVIRONMENT

The economic environment is basically about the level of demand in the economy. Most national economies follow the boom-and-bust economic cycle: every seven or eight years the economy goes into **recession**, which means that the production of

Recession A situation in which gross national production falls for three consecutive months.

goods and services shrinks and unemployment rises. A recession is a period of three consecutive months or more in which output shrinks, and the consequences may or may not be serious: during periods of recession, consumers are likely to postpone major purchases such as washing machines or new carpets due to uncertainty about employment security, and (by the same token) businesses will cut back on capital expenditure for such items as new factories or machinery. Borrowing is likely to reduce as consumers and firms become less confident about their ability to repay, and consequently demand drops still further.

In most cases recessions 'bottom out' within a few months or a year, but the financial collapse of 2008 created a worldwide recession in which many economies failed to recover for more than five years after the initial crisis. Governments tried many different measures to restart the world economy, but with little success.

Think outside this box!

If governments are so poor at controlling the economy, wouldn't it be better to leave things well alone and let Nature take its course? After all, there are so many factors to take into account in the way the economy works – people's confidence, the availability of manufacturing capacity, the activities of other countries and companies, and so on. Governments in the 19th century only concerned themselves with the defence of the realm and the internal security of its citizens – running the army and the police is a big enough task, surely!

On the other hand, the 19th century was marked by revolutions and rioting throughout Europe as starving people revolted against their governments. Maybe having a job and putting food on the table is a security issue after all.

Governments have a fine balancing act to perform in ensuring that the economy remains stable, and thus provides citizens with a good standard of living and a degree of confidence about the future. The problems caused by the financial crisis of 2008 have far exceeded government power to control: even when several governments act together, the situation can only be managed partially. In recent years, governments have controlled the economy largely by setting interest rates, and by controlling their own taxation and expenditure regimes. Both of these have a strong impact on marketers, because they affect people's willingness to spend on consumer goods and also (for firms that deal directly with the government) affect the size of the potential market. Non-profit organisations may feel the effects even more strongly, since many are funded from government grants and contracts, which may be cut back in times of austerity.

Within the European Union (EU) the common agricultural policy is an example of government intervention. The EU intervenes in agricultural markets, buying up and stockpiling food in order to maintain prices and smooth out supplies. However, this policy has resulted in the so-called 'wine lakes' and 'butter mountains' when continuing surplus production is bought and stockpiled, until eventually it has to be dumped on world markets or destroyed. On the other hand, the EU specifically prohibits governments from favouring their own national suppliers when ordering such items as computers or office equipment – all such tenders must be thrown open to suppliers in all member states.

Economic changes can be monitored in several ways. The business press typically provides informed analysis of economic changes, and national treasury officials in most countries also produce impact assessments. These are of variable quality according to the countries concerned. In some countries the assessments are as objective as it is possible to make them, since this allows companies and individuals to make informed judgements. In other countries the treasury produces distorted reports for reasons of political expediency, in order to support the party in power. Some universities and business schools also publish information and forecasts based on their own econometric models, and these may offer a different perspective from those forecasts produced by the government.

SOCIO-CULTURAL ENVIRONMENT

Socio-cultural forces fall into four categories, as follows:

1. **Demographic** forces. Demography refers to the structure of the population, in terms of factors such as age, income distribution, and ethnicity.
2. **Culture**. This refers to differences in beliefs, behaviours and customs between people from different countries.
3. Social responsibility and ethics. Derived in part from culture, ethical beliefs about how marketers should operate affect the ways in which people respond to marketing initiatives.
4. **Consumerism**. The shift of power away from companies and towards consumers.

Demographics The study of the structure of the population.

Culture The set of shared beliefs and behaviours common to an identifiable group of people.

Consumerism The set of organised activities intended to promote the needs of the consumer against those of the firm.

The relationship between these elements is shown in Figure 2.2. These relationships will be explained in more detail throughout this section.

Demographic forces are affected by variations in the birth-rate and death-rate, by immigration and emigration, and by shifts in wealth distribution, which may be caused by government policies. The demography of Western Europe has shifted dramatically over the past fifty years as the birth-rate has fallen and improvements in medical care have pushed the average age of the population sharply upwards. The birth-rate in Western Europe as a whole is now lower than the death-rate, so that the population would be shrinking were it not for immigration from Eastern Europe and the Third World. In some countries the situation is approaching

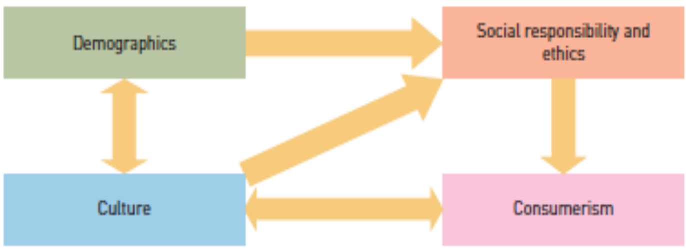


Figure 2.2 Socio-cultural environment

crisis point: for example, Spain has introduced a policy of contacting expatriate Spaniards in Latin America and encouraging them to return home. The Spanish government estimates that it needs 10,000 immigrants per annum to maintain the population.

The problem of depopulation and an ageing population is that many of the older people are retired, and therefore need to be supported by the productive members of society.

An influential report prepared for the EU in 2002 showed that the 15 member states (at that time – there are now 25 members) had experienced considerable immigration, virtually no outmigration, and dramatically reduced birth-rates. Coupled with the increased life expectancy (now around 75 for men, and 81 for women) the net result has been a reduction in the under-25 age group and increases in both the working population and the elderly population. These changes have happened over a thirty-year period from the mid-1970s (Cruijsen et al. 2002). The report goes on to say that entry by the new Eastern European member states will change this pattern in the short term, since these countries have lower life expectancies. During the 1990s (following the collapse of Communism in Europe) Eastern European countries have themselves experienced demographic shifts, notably a dramatically reduced birth-rate. These demographic shifts are thought to be the result of worsening health care, fear over job security, and less healthy lifestyles. The authors expect the following demographic shifts as a result of expansion:

1. Population decline will occur several years sooner.
2. Population ageing will be slightly suppressed.
3. Population dejuvenation (reduction in under-25s) will become stronger in future decades.
4. Expected decline of the working population will hardly change.

So far, experience has borne out these findings. There have been dramatic shifts in populations (several million Poles have emigrated to other EU states, for example), so there has been no decline in the working population of the original 15 member states.

From a marketing viewpoint, these changes offer both opportunities and threats. Clearly products aimed at a youth market are likely to decline, whereas products aimed at older people will be in greater demand. In practice, however, this may lead to surprises: for example, an assumption that almost all 70-year-olds have mobility problems may have been true thirty years ago, but improved health care and healthier lifestyles probably mean that most 70-year-olds in the 21st century are as fit as 50-year-olds were in the 1960s. The increase in the elderly population is not expected to peak out until the 2040s, and even this assumption depends on limited improvements in health care and the life expectancy of the very old – in other words, it assumes that people will not live much beyond 100 years old (Cruijsen et al. 2002).

A further demographic change (general to Europe) is the increase in single-person households. This has come about through an increase in the divorce rate and increasing affluence: young people no longer live with their parents until they marry, as was the case in the 1950s. At the other end of the age scale, large numbers of widowed elderly people continue to live in the former marital home. In several EU countries single-person households now represent the largest category of household: the UK's 2011 census revealed that single-person households had increased from 17% of all households in 1971 to 31% in 2011. Two-person households represented 32% of all households, making those two categories far and away the greatest proportion of households in the UK (Census 2011).

Real-life marketing: Think small

Most companies like the idea of being big. Retailers especially like to have big, well-stocked shops: a wider range of merchandise means more opportunities to sell something, after all. As in other aspects of marketing, though, you should be prepared to do something the others aren't doing – and thinking small has certainly been a success story for some firms.

The idea is to use small outlets, for example at railway stations and airports, to sell a limited range of goods that all fall into a particular category. This idea is used by firms such as Tie Rack and Sock Shop, who locate in high-footfall areas such as transport hubs. These companies pay a relatively low rent because the premises are very small, but they have a high turnover because people know they can get what they want quickly and easily – someone who has just spilled coffee down his tie can buy another one for that important meeting, for example.

For the idea to work, you should follow these rules:

- Think outside the box.
- Look for a resource that is currently unused or at least under-used.
- Specialise! This is essential for small firms – only very large firms can afford to be all things to all people.
- Don't try to compete head-on with the big companies.

The implications for marketing are widespread. For house builders, smaller homes and starter homes (e.g. flats) will show increased demand. This may mean that smaller models of domestic appliances will be more popular, that pack sizes of cereals and other foods will be smaller, that furniture will be smaller and perhaps more adaptable (for example futons, which convert from sofas to beds) and that security devices will be more popular as more people leave their homes unattended when they go to work. Such a rapid increase in single-person households represents a major challenge for many marketers since it implies a considerable shift in market demand for almost every consumer product.

Income distribution and wealth concentration are also part of the demographic structure. Income is a somewhat fluid concept: pre-tax income does not mean a great deal, since an individual's salary may be heavily or lightly-taxed according to the country concerned and the level of income of the individual. Disposable income is the income remaining after income tax and other deductions, but of course this is not the end of the story – basic household expenses need to be met such as mortgages, local authority property taxes, household bills and so forth. This leaves an amount which the individual can spend in any way he or she chooses: this is called discretionary income. There is, of course, a conceptual problem here in distinguishing between necessities and discretionary purchases. Housing is an example – a relatively wealthy person might choose to live in a small house, and thus have an extremely small mortgage and a correspondingly high discretionary income. Someone else might decide to live in a large house, and have very little discretionary income as a result. In either case, the choice of house was freely-made, so the house purchase might be considered in the same way as the purchase of a particular brand of bread or make of car. Clothing is even more problematical – wearing some kind of clothing is obviously essential, but the fashion industry is founded on the basis of attracting discretionary income, so the line between necessity and discretionary purchase is somewhat blurred.

Appendix D Reading Comprehension Assessment - questions

1. What affects making business decisions in an organization?

2. How does the network between people in an organization bring about a better living standard?

3. List some stakeholders that will have effect in making marketing decisions.

4. What is the relationship between the firm and its environment?

5. In Costain West Africa case, why did Costain Company survive despite the financial crisis in 2008?

6. In what ways can the marketing environment factors be classified?

7. How does a firm relate to these factors?

8. List one or two examples for each type of marketing factor.

9. why are the boundaries between the internal and external environment sometimes hard to decide?

10. How the quality of economic changes assessment is varied in different countries?

11.Why is the external environment impossible to control?

12. Given that some macro factors are common for all firms, why will they affect the firms differently?

13. What is the problem of classifying the macro and micro environment factors? Please give examples.

14. What is boom-and-bust economic cycle?

15. What would be the consequences of economic recession?

16. What changes would the increasing single-person households make to the market?

17. Can you infer what measurements the government took to recovery the economy in 2008 recession?

18. Is it true that the government in 19 century is too poor to control the economy due to the defence of realm?

19. What do governments do to control the economy and stabilize the markets?

20. How does the non-profit organization benefit from government control?

21. Can you list some supporters of government control?

22. Can you think of any opponents against government control?

23. How does the EU intervene in the agriculture markets?

24. What does the intervene result in?

25. How do you think about EU intervenes in markets?

26. Can you predict the results that EU makes computers or office equipment's supply open to all members bring about?

27. How is the quality of economic changes assessment varied in different countries ?

28. What factors will influence social-cultural environment?

29. In Figure 2.2, how are the factors within social-cultural environment affected with each other?

30. Why does Spanish government encourage Latin Spaniards to return home?

31. What would be the consequences of depopulation and an aging population?

32. What were the changes for the fifteen states members of the EU during the thirty years after the mid-1970?

33. What brought about the decrease of birth rate of Eastern European countries during1990?

34. In one report about EU in 2002, what changes already happened and what did the author predict in the future?

35. What do you expect the working population change in the original fifteen member states?

36. How did the age demographic changes affect the market in West Europe?

37. What evidences can you list to explain the increase in single-person households in several European countries?

38. In the "Real-life marketing"(P35), what information was used for the companies like Tie Rack and Sock Shop to make their marketing decisions?

39. Can you predict other changes that increasing single-person households would bring to consumer-product companies?

40. How do you illustrate the distinction between "necessities and discretionary purchase"?

41. How do you illustrate the distinction between "necessities and discretionary purchase"?

42. Based on your view, how do you explain the conclusion "line between necessity and discretionary purchase is somewhat blurred"?