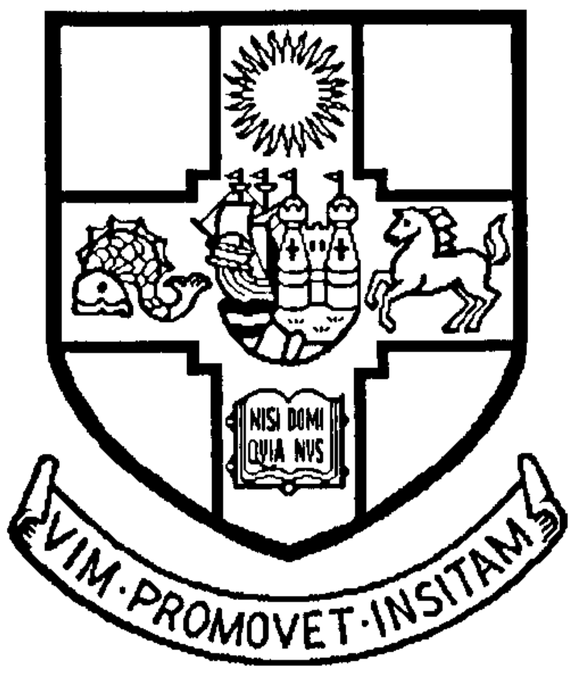
Towards a Model of Using Summarization Tasks as a Measure of Reading Comprehension

Graduate School of Education
University of Bristol



A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Social Sciences and Law

Abstract

From the perspectives of education, psychology and applied linguistics, this study reviewed the promises and problems of using summarization tasks as a measure of reading comprehension. A dynamic IFOE framework (*input*, *filter plant*, *output* and *evaluation*) of summarization tasks is proposed and empirically investigated through an ecological approach. 157 Chinese undergraduates were randomly assigned to one of the summarization tasks in a factorial design of 3 text types x 2 text presentation modes (*computer* vs. *print*) x 2 language orders (*English then Chinese* vs. *Chinese then English*). Post-summarization questionnaires and interviews were administered to elicit students' perceptions of the summarization tasks. In addition, their computer familiarity, and their reading, writing and translation abilities were measured. The quantitative data were therefore analysed to statistically model (i) the relationships between students' summarization performances and their other language abilities and computer familiarity, (ii) the impacts on the students' summarization performances of text type, presentation mode and use of a different language and (iii) the differential effects of two empirically developed assessment criteria. The qualitative perception data were analysed through a grounded theory approach, in relation to the three focal points above.

The results indicate that summarization is a dynamic process affected by various facets in the proposed IFOE framework. The only significant language ability predictor of summarization – English reading – explained a very small amount of the variance in summarization performances. Type of source text was as influential as English reading ability in affecting summarization performances. Furthermore, the use of the first language had advantage over the target language. However, the qualitative interview data revealed that students distinctly prefer English native speakers' performance as the unchallenged model for developing assessment criteria. Effects of text presentation mode seemed to have been mitigated by students' high computer familiarity whose effects on summarization were more psychologically anticipated than actually experienced by the students.

These results provide evidence that summarization tasks are both promising and problematic as a measure of reading comprehension. Implications for the use of summarization tasks are suggested, essentially arguing for an ecological and organic approach to task design and engagement of test-takers in developing assessment criteria. A series of studies is suggested to further explore and develop the potential of the IFOE framework in the use of summarization tasks for the assessment of reading comprehension.

In memory of Mother

Your mother is always with you.

She's the whisper of the leaves

as you walk down the street.

She's the smell of bleach

in your freshly laundered socks.

She's the cool hand on your brow

when you're not well.

Your mother lives inside your laughter.

She's crystallized in every teardrop.

She's the place you came from,

your first home.

She's the map you follow

with every step that you take.

She's your first love

and your first heartbreak...

and nothing on earth can separate you.

Declaration

I declare that the work in this dissertation was carried out in accordance with the Regulations of the University of Bristol. The work is original, except where indicated by special reference in the text, and no part of the dissertation has been submitted for any other academic award. Any views expressed in the dissertation are those of the author.

SIGNED _____ DATE 01.09.2005

Acknowledgements

It would not have been possible for this dissertation without help and advice from numerous people to whom I am very much indebted and grateful.

Professor Pauline Rea-Dickins, my academic supervisor, for her expert advice and academic inspiration to my dissertation research. It has been a pleasurable and productive time working with her since 2000 when I first started my study attachment at the Graduate School of Education.

I am also very grateful to the students who participated at various stages of this research: over 180 in pilot studies and 167 in main studies; without their voluntary participation, this would never be a research at all.

The three raters – Zhang Jing, Zhang Yan and Zhou Wei deserve absolutely special thanks here. Without them, this research would never have been completed.

I am also very grateful to other experts within and beyond the Graduate School of Education for various reasons.

Dr Katie Scott, Dr Judi Kidger, Dr Martin Crossley Evan, Mr Lawrence Cattermole, Mr Gordon Landreth, Mrs Judi Spenser and the Goulds helped to write the summaries at various stages.

Comments from members of the Centre for Research on Language and Education (CREOLE) at various stages of the dissertation research are greatly appreciated. Special thanks are due to Dr Katie Scott for her friendship and comments on an earlier version of this dissertation.

Dr Sally Barnes (Director of PhD Programme at GSOE) spared her time to advise me on the design of the post-summarization questionnaire. Professor Harvey Goldstein and Mr Anthony Hughes helped with upgrading my abilities in quantitative data analysis.

Dr Ngoni Chipere of the University of West Indies and Dr Jeanine Treffers-Daller of University of West England helped me with the CLAN programme at the initial stage of the vocabulary density analyses.

Professor Cyril Weir of University of Surrey Roehampton and Dr Craig Deville of University of Iowa commented insightfully on my presentations at Language Testing Forum 2002 and Language Testing Research Colloquium 2003 respectively.

Dr Lynda Taylor of UCLES ESOL Examinations kindly presented her 1996 PhD research, which was very helpful in setting up my own research, at CREOLE in 2002. Dr Robert Couchên of University of Ottawa (Canada) posted me his article on summary cloze, which would otherwise never be available to me. Dr Yasuyo Sawaki of Educational Testing Service (formerly University of California Los Angeles) sent

me her whole PhD dissertation that has been enormously helpful. Dr Marsha Bensoussan of University of Haifa (Israel) had very informative email communications, sharing with me her findings on using summary cloze test in research and practice.

Dr Richard Kiely served as the MPhil/PhD upgrade examiner and gave much useful advice on the design of the research and helped with the translation of a French rating scale and the pilot studies at Graduate School of Education. Professor Hong Shen of Peking University helped to check some of the Chinese versions of the research instruments.

Dr Guoping Wang of University of Michigan School of Medical Science, my cousin, has been, as always, very supportive and encouraging throughout my PhD study.

I am also very grateful to UCLES, Educational Testing Service and Reeds Information Limited who granted special permissions to use their copyrighted materials in this research.

Without the financial support from the Universities UK Overseas Research Studentship Awards Scheme (United Kingdom Scholarships for International Research Students), and the University of Bristol Postgraduate Research Scholarship, I may never have had the chance to carry on my research at the University of Bristol. The Graduate School of Education Research Student Conference Fund also made it possible to disseminate the research project and establish research networks. I am also very grateful to the Oz Osborn Memorial Fund which helped me to buy essential books and computer software for the project. Educational Testing Service granted not only its special permission to use a TOEFL past paper, but also helped enormously with my data collection process, with its TOEFL Small Grant for Doctoral Research in Second/Foreign Language Assessment.

Finally, but definitely not least, I owe too much to my wife and our daughter who had to be often reminded by her mother that Daddy was not available to play with her in the Clifton Suspension Bridge Park, one of her, and also my favourite, parks in Bristol.

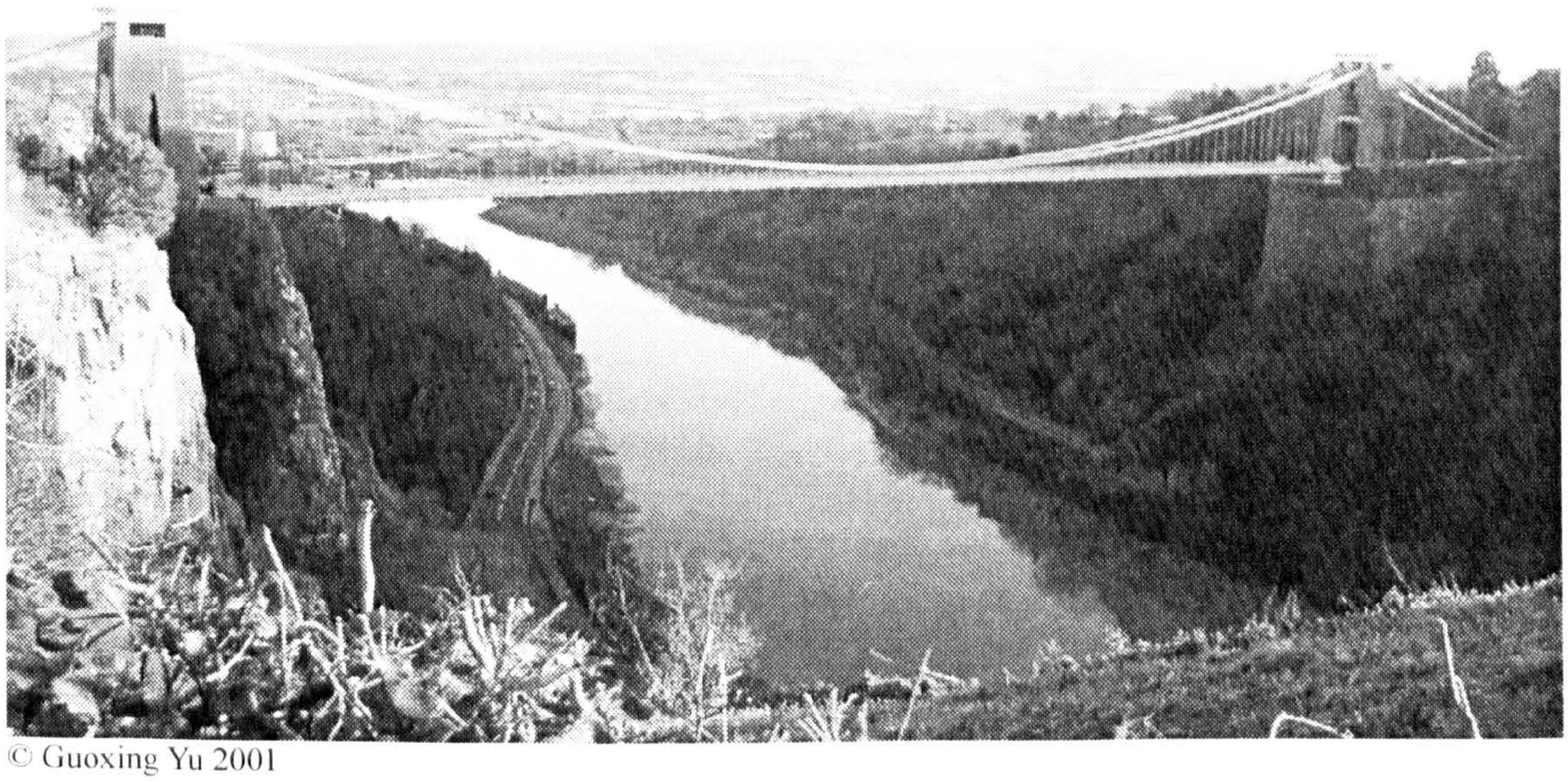


Table of Contents

List of Abbreviations and Acronyms

5%S	5% Score
CE	Chinese expert scoring template
C/E	Chinese then English
CEHS	HS score of Chinese summary according to expert template
CERSC	RSC score of Chinese summary according to expert template
CFQ	Computer Familiarity Questionnaire
CLAN	Child Language Analysis
CP	Chinese popular scoring template
CPHS	HS score of Chinese summary according to popular template
CPRSC	RSC score of Chinese summary according to popular template
C.S.L	Length of Chinese Summary
D-SS	Vocabulary density of student summary
D-ES	Vocabulary density of expert summary
E/C	English then Chinese
EE	English expert scoring template
EEHS	HS score of English summary according to expert template
EERSC	RSC score of English summary according to expert template
EFL	English as a Foreign Language
EP	English popular scoring template
EPHS	HS score of English summary according to popular template
EPRSC	RSC score of English summary according to popular template
ESL	English as a Second Language
E.S.L	Length of English Summary
ETS	Educational Testing Service
FCE	First Certificate in English
FCE-R	Reading paper of First Certificate in English
GLM	General Linear Model
GSOE	Graduate School of Education, University of Bristol
HS	Holistic score
IFOE	Input, Filter plant, Output, Evaluation framework
LANG	Language
LANGORD	Language Order
PRESMODE	Presentation Mode
PSI	Post-Summarization Interviews
PSQ	Post-Summarization Questionnaire
RSC	Right Statement Credit
SSS	Summary and Source text relationship Score
SPSS	Statistical Package for Social Sciences
TEFL	Teaching English as a Foreign Language
TEM	Test for English Majors (China)
TOEFL	Test of English as a Foreign Language
TOEFL-R	Reading section of Test of English as a Foreign Language
TST	Traditional Summarization Tasks
TWE	Test of Written English
TXT	Text type
UCLES	University of Cambridge Local Examinations Syndicate
VOCD	Vocabulary Density (a computer programme in CLAN)
WSP	Wrong Statement Penalty

List of Figures

List of Tables

List of Appendices

INTRODUCTION

Discontented with current practices of *purifying* reading comprehension test methods (see **1.1**) and *disempowering* test takers in test construction and interpretation (see **2.4.1** and **2.5.1**), I explored a rather old fashioned but scarcely researched reading comprehension test method – traditional summarization tasks (TST) – in the new era of communicative language testing and the information age. At first glance, it all sounds very unfashionable; however, it is also interesting to note that integrated reading-writing tasks such as summary writing are now being revived in language testing practice (e.g. in the next generation TOEFL to be introduced in September 2005). The previously neglected role of summary writing is probably due to claims that summary writing is a *muddied* measurement of reading comprehension (i.e. confounding reading and writing skills) (Alderson *et al.* 1995; Alderson 1996; Weir 1993, 2005), albeit without much empirical evidence to support this. On the other hand, the recent revival of summary writing as a test method also calls for empirical studies to gain better understanding of its previously claimed muddiedness, but simultaneously its intuitive appeal in communicative language testing. The investigation of summary writing has all the topicality of language testing research and practice.

In order to answer such calls and also to arrive at a better understanding of my own use of TST as a measure of reading comprehension, I reviewed the literature on summarization studies in psychology, education and linguistics and proposed a summarization framework for language testing purposes – IFOE (Input-Filter Plant-Output-Evaluation). All the four components of the framework (*input*, *filter* *plant*, *output* and *evaluation*) were investigated in an organic approach in this research (Figure 2.1), so as not to break the assessment ecology.

In particular, this research studied the summarization performances of 157 Chinese undergraduates. They were asked to summarize, in both their first language (Chinese) and foreign language (English), an extended English text, either computer mediated or paper presented, as a measure of their EFL (English as a foreign language)

reading comprehension abilities. The study explored IFOE factors such as (i) text types, (ii) text presentation modes (computer vs. paper), (iii) summarizers' EFL reading, writing (English and Chinese) and translation (from English to Chinese) abilities, (iv) languages used for writing the summaries (Chinese and English), and (v) assessment criteria (expert vs. popular template). The summarization tasks were organized in a factorial design of 3 text types x 2 text presentation modes (*computer* vs. *print*) x 2 language orders (*English then Chinese* vs. *Chinese then English*) and the written summaries were subjected to both expert and popular assessment criteria (see **2.5.1**). In addition, the students' computer familiarity, reading, writing and translation abilities were also measured through other instruments. The impact of these factors on students' summarization performances were analysed from two parallel datasets: (i) students' *actual* performances in the summarization tasks and other measurements such as computer familiarity, reading, writing and translation abilities, and (ii) their *perceptions* of the TST through post-summarization questionnaire and interviews.

This project is methodologically innovative in its investigation of the issues of test takers' values and their involvement in developing assessment criteria. This kind of involvement and empowerment of test takers and the comparisons of actual use of "expert" and "popular" scoring templates are innovative attempts to develop empirically the notion of "indigenous assessment criteria" (Jacoby & McNamara 1999). The empowerment of test takers is also ethical research conduct in the sense that the voices of the participants are not silenced; on the contrary, they are highly valued, as proposed by Bachman and Palmer (1996: 32).

The investigation of the under-researched construct – summary writing as a measure of reading comprehension – will contribute to our greater understanding of summarization tasks in the contexts of not only language testing but also more general psychological and educational assessments of discourse comprehension. The findings are also relevant to research and practice in the teaching of summarization skills to facilitate students' academic success. As such, the findings of this research will be of interest to four specific user groups, namely (i) examination bodies and councils engaged in researching appropriate ways to assess reading comprehension skills, (ii) theorists in discourse comprehension, (iii) academics researching language, psychological and educational assessments and (iv) language support staff and

university subject tutors providing their students with better scaffolding for the development of summarization skills vital for students' academic success.

This dissertation consists of five parts, organised in twelve chapters. Part I (Chapter 1) describes briefly the motivation and purposes of this research by reflecting on my discontent with and critique of current reading comprehension test methods and my personal philosophy in terms of organic foods and the ecology of educational assessment. Also presented at the end of this chapter is a brief introduction to the five research questions that have guided this research.

Part II (Chapter 2) reviews the literature on summarization studies in education, psychology and linguistics and proposes the IFOE framework. The literature review itself is however organized in the order of *evaluation*, *input*, *output* and *filter plant*.

Part III first discusses the research approach of this project (Chapter 3), followed by the details of research questions and hypotheses, and data collection procedures (Chapter 4). Part IV consists of six chapters (5-10), of which the first component (Chapter 5) focuses on micro-level analyses on the data from each individual research instrument separately. The second component of Part IV (Chapters 6 to 10) presents the findings from macro-level analyses in the order of the five research questions. In Part V (Chapters 11 and 12), the findings to the five research questions and their implications are discussed within the IFOE framework. Directions for future research studies into the IFOE framework are suggested.

PART I

CHAPTER ONE

Motivation and Purposes of This Research

This chapter describes briefly my discontent with and critiques of current reading comprehension tests, in particular, (i) the gaps between theories of reading comprehension and practices of reading assessment, (ii) the problematic effects of test methods in distorting test takers' mental representations of text comprehension. To better measure students' reading comprehension abilities through a less distorting method, I propose a rather old-fashioned but scarcely researched reading comprehension test method – the traditional summarization tasks (TST), neglected largely due to its claimed confounding effects of writing abilities on the measurement of reading comprehension, albeit without much empirical evidence for this rejection. On the other hand, TST is currently undergoing a revival because of its "natural" appeal and close approximation to target language use. For both the rejection and the revival, there is an urgent need for empirical evidence to gain a better understanding of TST. Following my personal philosophy and practice of using TST to measure reading comprehension, I present briefly the motivation and purposes of my studies and research focus.

1.1 Problems with common practices of reading comprehension tests

The problems of current reading comprehension test methods are discussed briefly to highlight the gaps between theories of reading comprehension and practices of reading assessment, distortions of test takers mental representations of text comprehension, and deprivations of their rights in constructing meanings of the text.

1.1.1 Gaps between theories and practices of reading assessment

The theoretical understanding of reading comprehension has been compelling language testers to re-think their test construction. The acknowledged gaps between

theories of reading comprehension and practices of reading assessment are due to various factors. No party is to blame for such a gap, though it does appear that reading assessment

> has been, and commonly still is, driven either by language learning notions of communicative language performance or by assessment theory more generally, including the reasonably strong psychometric qualities of traditional reading comprehension tests,

rather than by theories of reading comprehension (Grabe 2000: 246). These gaps were also convincingly argued earlier by Farr and Carey (1986), Valencia and Pearson (1987), Just and Carpenter (1987), and Anderson *et al.* (1991). It is frequently the case that reading assessment tools have to pass through psychometric evaluations in terms of the traditional concepts of reliability and validity (Grabe 2000: 247), before they can be used with "confidence" for language proficiency tests.

1.1.2 Problematic effects of test methods in practice: selective critiques

One of the criteria for evaluating the practicality of a reading assessment tool is whether it is easy to score test takers' responses, with the ultimate aim of improving reliability, or scoring validity in Weir's (2005) terms. Taking multiple-choice questions (MCQ) as an example, these have their merits (for example, high scoring reliability), but are also widely criticized from various research perspectives. Students may comprehend the text but do not understand the questions, therefore, we may not be sure whether the exact point of misunderstanding lies in the text or in the questions. Neither are other test methods, such as short answer questions (SAQ), immune from this problem. Some research studies (e.g. Katz *et al.* 1990; Pyrczak 1974; Royer 1990) have found it is possible for students to answer questions without even reading the source texts; on the other hand, the presence of questions about a text may aid and stimulate comprehension by turning students' attention to specific parts of the text (Bensoussan 1982; Bensoussan *et al.* 1984), and therefore the on-going construction of students' mental representation of the text is affected (Gordon & Hanauer 1995).

Multiple choice also influences students' ability to demonstrate their comprehension by delimiting their possible responses (Bernhardt 1991). Choosing the right answer from four or five alternatives after reading a passage is not a normal reading activity except in test environments (Urquhart & Weir 1998), calling into question the construct validity of MCQ in testing reading comprehension. Talking

about the validity of MCQ, Weir (1988:57) expresses the similar concern that:

> the format is artificial and is increasingly perceived as an invalid measure for assessing comprehension by teachers, materials designers and language testers.

Test takers also stand a very good chance of getting the right answer by simply guessing, without understanding either the text or the question. As Heaton (1990) suggests, it is desirable to improve the reliability of a valid measure rather than to improve the validity of a reliable but less valid measure. However, validity is impossible to achieve without reliability (Bachman 1990; Ghiselli *et al.* 1981; Henning 1987), since "reliability is a necessary condition for construct validity, ..., However, reliability is not a sufficient condition for either construct validity or usefulness" (Bachman & Palmer 1996: 23).

In the same trend of achieving high scoring reliability, several "innovative" summarization tasks[1] as a measure of reading comprehension have also been researched, such as *summary-cloze*, *summary completion*, *best summary choice*. These tasks are problematic for various reasons (see also **2.4.1**). In summary cloze tasks, it is possible for test takers to gain a score without reading the source text, which raises the question of whether such tasks measure understanding of the source text or only of the summary (e.g. Courchêne & Bayliss 1995; Taylor 1996). Similar problems arise in best summary choice tasks, coupled with the extreme difficulty in designing such tasks (Huhta & Randell 1996). As demonstrated by Pyrczak (1974), it is possible for test takers to get the right answer to main idea comprehension questions without reading source texts. As with MCQ, Cloze and SAQ, these "innovative" summarization tasks also distort test takers' mental representations of text comprehension (see **2.4.1** for further discussion on the "innovative" summarization tasks).

1.2 Proposal of an alternative: traditional summarization tasks

It seems imperative that we design reliable and valid measures of reading comprehension abilities that are less (hopefully not) distorted by test methods *per se* while drawing confidently on current reading comprehension theories. The less the

[1] I use "innovative" to distinguish these forms of summarization tasks such as summary cloze, summary completion, and best summary choice from the traditional summarization tasks researched in this project. However, I do not endorse that these "innovative" forms are innovative.

influence from test methods, the less distorted mental representations of text comprehension, and therefore the better it is to tune to the "organic" reading process and the measurement of reading comprehension abilities. Research studies about test method effects on students reading performance (e.g. Kobayashi 1995; Riley & Lee 1996; Shohamy 1984; Wolf 1993) have found different test methods may involve different reading process, and therefore measure differentially reading abilities.

Out of my personal, philosophical, practical and professional interests, my preference has been to explore a less frequently used and under-researched reading comprehension test method –TST - whereby test method effects might be minimised and the approximation of mental representation of reading comprehension correspondingly maximised. In TST, students' reconstruction of a text is free from the potential influence or contamination of multiple choice or essay prompts (Bernhardt & Deville 1991). Test takers' mental representations of text comprehension may be less distorted than by MCQ, Cloze and SAQ. The definition of the term TST is discussed in detail in Chapter Two (see **2.1**).

1.2.1 Summarization as a personal interest in practice

As part of my own professional practice as a university lecturer, I have singled out the use of TST to measure my students' reading comprehension abilities at end-of-term examinations, but have been seriously challenged by colleagues. Their arguments are based on subjectivity in marking summary protocols and the confounding effects of writing ability (i.e. muddiedness), as well as other factors, on students' summarization performance. My usual defence is that summary writing is not the same as other composing activities, such as independent essay writing, and that our realistic and responsible aim is to get the best possible picture of students' reading comprehension abilities from the potentially "muddied measurement" because no "pure" measurement exists. Very often, "natural" or "organic" measurement is a "muddied measurement". As one of the key research questions, this study focused on whether and to what extent summarizers' linguistic abilities affected their summarization performances.

1.2.2 Personal philosophy of organic foods and ecological assessment

It is my personal philosophy that organic foods very often mean that they are muddied and that they are not purified. They may look ugly and dirty, but that is what they should look like. Similarly, to measure reading comprehension ability purely as many testers try to achieve is not an organic or ecological approach in itself. When a test method tries to purify something, it may well lose something that is essentially part of that construct being assessed, some "environment" that the ability relies on to exist. When we purify our measurement tools, we are also distorting and probably destroying the natural "environment" that reading ability relies on. In my understanding, rather than purify measurement tools, testers should purify, if possible, or extract reasonably well the reading ability measured by an organic method – though it may be muddied. Starting with something organic, and then trying to find the "real food" in the mud is one of the two guiding principles of this research.

The other guiding principle is that test takers have several important roles to play in test construction and interpretation, not only from the perspective of current understandings of test takers' human rights, but also from postmodernist interpretations of texts. Meanings of a text only reside in the reader who can interact with and interpret the text in various legitimate ways. Test constructors' understandings of a text are only one of the legitimate manifestations of the meanings of a text. Their understandings may not be the only understandings, nor those of the test takers who may have different but legitimate understandings. However, in most reading comprehension tests, test takers have to accept test constructors' understandings to earn a score. Thus, the challenge is if we can integrate the understandings not only of test constructors but also test takers to form a scoring template that is accepted by most of the stake-holders in testing endeavours. In this way, we might have an opportunity to release test constructors from unrealistic burdens, to empower and motivate test takers to be involved in the whole process of testing as they would not have to accept the imposed "authoritative" answers from test constructors.

1.2.3 Practices of summarization task in the profession of language testing: urgent need for empirical evidence

TST is a little researched area in the field of language testing and was for many years neglected as a reading comprehension test method largely due to its claimed muddiedness (e.g. IELTS, see Charge & Taylor 1997), albeit without much empirical evidence for this rejection (see further explanation on p.1). However, such integrated reading/writing tasks have undergone a revival recently in very high-stakes and large scale tests such as the next generation TOEFL because of the tasks' natural appeal and close approximation to target language use. For both the rejection and the revival of such tasks, systematic empirical evidence is urgently needed. On the personal front, it is also imperative for me to gain better understanding of the use of such tasks in my own teaching and testing practices and to be better able to face the challenges of colleagues (see **1.2.1**).

1.3 Brief introduction to the research focus

This research investigated students' *actual* performances and *perceptions* of traditional summarization tasks, with references to the key components of the proposed IFOE framework as explained in the Introduction (pp.1-2). Five research questions were examined to explore the *input* (text type and presentation mode), *filter plant* (language abilities), *output* (English and Chinese summaries) and *evaluation system* (expert and popular scoring templates) of the framework (see Figure 2.1).

Research Question One (RQ1)

What are the differences in score variances and students' attitudes between using expert and popular templates to evaluate their written summaries?

Research Question Two (RQ2)

Are students' summarization performances affected by their other linguistic abilities and if so, to what extent?

Research Question Three (RQ3)

What impact does the use of a different language and language order have on summarization performances and measurement of reading comprehension abilities?

Research Question Four (RQ4)

What are the effects of text presentation mode and students' computer familiarity on their summarization performances?

Research Question Five (RQ5)

What are the effects of text type on students' summarization performances?

1.4 Summary

This chapter briefly discussed the problems of some commonly used reading comprehension test formats and proposed the use of traditional summarization tasks (TST) out of my personal, philosophical, practical and professional interests. In the next chapter, studies on summarization in education, linguistics and psychology are reviewed, within a proposed framework of using summarization as a measure of reading comprehension.

PART II

CHAPTER TWO

Summarization as a Measure of Reading Comprehension

This chapter will first of all define "summarization" as a generic term used in this project (**2.1**), and then provide a detailed account of the premises (**2.2**), the promises (**2.3**), the practices (**2.4**) and the problems (**2.5**) of using TST as a measure of reading comprehension, drawing on literature from education, linguistics and psychology. In particular, the proposed IFOE (*input*, *filter plant*, *output*, and *evaluation*) framework will not only act as the reference point for the literature review, but also form the basis for this project's five research questions.

2.1 Defining summarization: starting points

What is a summary, and what is summarization? At first glance, the notions of summary and summarization are commonsensical and easy to understand. The COBUILD English Dictionary (Lingea Lexicon 2002, electronic version 4.11) gives the following entries for *summary* and its related synonyms:

> A ***summary*** of something is a short account of it, which gives the main points but not the details of it.
>
> An ***abstract*** of an article, document, or speech is a short piece of writing that gives the main points of it.
>
> A ***précis*** is a short written or spoken account of something, which gives the important points but not the details.
>
> A ***résumé*** is a short account, either spoken or written, of something that has happened or that someone has said or written.
>
> A ***synopsis*** is a summary of a longer piece of writing or work.

Merriam-Webster's Collegiate Dictionary (electronic version 2.5) defines summary as: "*abstract*, *abridgment*, or *compendium* especially of a preceding discourse":

> ***Abstract***: a summary of points (as of writing) usually presented in skeletal form.
> ***Abridgment***: a shortened form of a work retaining the general sense and unity of the original;
> ***Compendium***: a brief summary of a larger work or of a field of knowledge.

On closer scrutiny, however, these definitions are not as straightforward as they seem. How short is "short"? What constitutes "main points" and "important points"? Similarly, what are considered "the details"? And by whom? On what criteria? Is there any difference between "main points" and "important points" and "general sense"? Does a summary have to be written, but not spoken? It is understandable that dictionaries have to give concise definitions. However, even in the research literature, the definition of "summary" is not as straightforward as we might assume (Seidlhofer 1991, 1995). In the most extensive literature review on "summarization" so far, Hidi and Anderson's definition, based on N. Johnson (1983), seems as ambiguous as those in dictionaries:

> Traditionally, a summary is a brief statement that represents the condensation of information accessible to a subject and reflects the gist (central ideas or essence) of the discourse. (Hidi & Anderson 1986: 473)

So is McAnulty's more detailed definition (1981: 50):

> A summary is a condensed version, in your own words, of the writing of someone else, a condensation that reproduces the thought, emphasis, and tone of the original. It abstracts all the significant facts of the original – overall thesis, main points, important supporting details – but, unlike a paraphrase, it omits and/or condenses amplifications such as descriptive details...

What are considered "central ideas or essence"? Is it only by the "subject" or by the writer(s) of the source text(s) or by both? How brief is "brief"? Is there any difference between "supporting" and "descriptive" details? Are personal and evaluative comments allowed at all in a summary? Is a summary written for oneself or for other readers? Many questions remain to be clarified before reaching a generic definition of summary. As Ratteray (1985) points out there are at least seven types of summary[1] that have emerged in common usage, "a serious problem in much of this literature [i.e. the formal practice of summarization], however, is the assumption that only one kind of summary exists" (*ibid.*: 457).

Coupled with the ambiguity of the definitions of "summary" in practice, summarization is also closely linked with other terms such as text recall, main idea

[1] They are sequential summaries that retain the original order in which information was presented (including *abstract*, *précis*, *secretarial minutes*, *abridged digest*) and synthesizing summaries that alter this sequence to achieve specific objectives (including *locational digest*, *restructuring digest* and *review*).

comprehension, paraphrasing, and abstracting in the research literature (see also above the synonyms of *summary* given in the two dictionaries). In terms of psychological processes involved in the five reading purposes, they may be quite different; however, all five are closely related and all involve reduction and reconstruction of source texts. They are also very often considered to involve similar cognitive processes. Summary is a less-detailed recall; recall is "summary-plus-details" (van Dijk & Kintsch 1977)[2]; the term abstract "signifies an abbreviated, accurate representation of the contents of a document, without added interpretation or criticism" (ISO-214-1976). A paraphrase, in a looser sense, "may be close to a summary as soon as more detailed information of the paraphrased text is paraphrased with 'fewer words'" (van Dijk 1980: 102). Summarization, thus, comes under the umbrella term of "main idea comprehension".

The terminological chaos in "main idea comprehension" rubs salt into the wound of the already ambiguous definitions of summary. Pearson (1981) says that "the term main idea is but a main idea for a polyglot of tasks and relations among ideas", and that "the concept of main idea has not been defined consistently either in instruction or in research". Cunningham and Moore (1986) review main idea research and instruction and find the main idea world[3] is "confused". *A Dictionary of Reading and Related Terms* (Harris & Hodges 1981) further attests to this confusion with the note: "there is little agreement on what a main idea is" (p.188). Williams (1988) also laments this confusion in the definitions of main idea.

Because of the confusions in defining summary and its synonyms in the literature, I searched, at the initial stage of this project, *all* of the five related areas: main idea comprehension, summarization, recall, paraphrasing and abstracting, and then gradually focused on the two most relevant areas: summarization and recall for

---

[2] This theoretical claim is supported in research such as Goldman *et al.* (1995), which found the major difference between the two was that the likelihood of including elaborations was greater in recall than summarization.

[3] They found nine types of main idea comprehension: *gist* (a summary of the explicit contents of a passage), *interpretation* (a summary of what might be intended), *selective summary or selective diagram* (a summary of the literal words or phrases in the text obtained by selecting and combining some of them), *theme* (point about life or world made by passage), *title* (a name for the passage), *topic* (a label for the subject of the passage), *topic issue* (a word or phrase giving context for the passage), and *topic or thesis sentence* (a sentence in the text which best summarizes it).

literature review (see Yu 2005).

In this project, I used summarization as a more generic term. It was defined as a process of reducing and reconstructing a written text in a systematic way into a faithful and generalised *re*-presentation of the source text according to the needs of the summarizer. The product of summarization was therefore a written summary. My research focused on literal rather than critical, written rather than oral summarization performance to measure reading comprehension abilities.

2.2 Premises of summarization as a measure of reading comprehension: a theoretical perspective

Reading comprehension is a *sine qua non* for summarization. It is commonly held that comprehension is one of the prerequisites of summarization[4]. The reading process naturally involves summarization (van Dijk & Kintsch 1983), though not necessarily automatically (N. Johnson 1983). "Summarization requires the comprehension, evaluation, condensation, and frequent transformation of ideas that have been presented" (Hidi & Anderson 1986: 473-474). Since Bartlett (1932), there has been a host of publications on how the gist or main idea of a text is processed. For example, Gomulicki (1956: 90) states that "full understanding of a passage also demands an appreciation of the relative importance of its parts", a process involved in summarization (Hare & Borchardt 1984). Thorndyke (1975) finds high-level statements are more readily recalled and summarized than low-level statements in stories. E. Kintsch (1990) also finds that weak readers tend to read for microstructures and good readers for macrostructures. The amount of higher-level macroprocessing is commensurate with age. Other researchers conclude, similarly, that skill at comprehending important information in a text discriminates good from poor readers (Eamon 1978/1979; Smiley *et al.* 1977; Winograd 1984). For example, Winograd (1984) concludes that good and poor readers (36 poor and 39 good eighth graders and 37 adults) differ in what they considered important in a text, in what they included in

---

[4] See Nancy Johnson's (1983) categorization of the prerequisites for summarization of stories: (1) comprehending individual propositions; (2) establishing connections between propositions; (3) identifying the constituent structure of a story; (4) remembering the information in a story; (5) selecting the information to be represented in a summary; (6) formulating a concise but coherent representation of that information. She argues that summarization may not be an automatic entailment of reading comprehension for her young participants, contrary to what proposed by van Dijk and Kintsch in their various studies on adults' summarization.

their summaries of text, and in how they transformed original text. These are just a very few examples from the range of publications to demonstrate how the main idea of a text is processed and why summarization performance can be considered an important and inherent indicator of reading comprehension abilities.

2.3 Promises of summarization tasks in communicative language testing: a pragmatic perspective

The previous section briefly discussed that reading comprehension is a *sine qua non* for summarization so as to establish the theoretical premises of using summarization tasks as a measure of comprehension. In this section, I present some promising roles that summarization tasks can play in the practice of communicative language testing, in terms of their close approximation to target language use and the necessity of developing and measuring summarization skills.

2.3.1 Close approximation to target language use

From a language testing perspective, Bachman (1990; 1991; Bachman & Palmer 1996) mentions two fundamental requirements for ensuring the validity of effective language testing procedures. Firstly, the language abilities measured by the test must correspond to those abilities needed to carry out tasks in the target-language use situation, and secondly, features of test tasks, or test method characteristics, must correspond to critical features of target language use tasks. In this sense, summarization tasks have a natural appeal because they "simulate real-world tasks in which non-native readers have to read and write a summary of the main ideas of a text" (Cohen 1994: 174). Even if we are only interested in readers' organizational competence,

> a test requiring test takers to...summarize the propositional content in a reading passage, will involve the *full* [emphasis added] range of organizational characteristics. (Bachman 1990: 139)

In practice, TST, in which the test taker is instructed to summarize a text s/he is to read (or has already read) in his/her own words, has a long history of being used to measure reading comprehension abilities, and was considered a "very valuable" (Nuttall 1996: 206) reading task:

> It [summarization] demands *full* [emphasis added] understanding of the text, including the ability to distinguish between main points and examples, to perceive the relationships between the various parts of the argument, and so on (*ibid.*).

Getting the gist of a passage, usually in the form of a written or oral summary, is an essential communicative activity (Brown & Smiley 1978). A well-structured summary test may also "promote a richer, more interactive approach to reading than do comprehension tests that focus more on details" (*ibid.*: 203). Integration of reading with writing (e.g. summarization in this research), as Smith (1988) comments, is "one way of promoting engagement with a text which leads to better comprehension"[5].

The close approximation of the use of summarization tasks for testing purposes to real-world tasks is further attested in university students' experiences in language learning as well as their general academic study and future professional development. Summarization is critically related to their self-study as well as for presentations, assignments and other academic development tasks (Allison *et al.* 1994; 1995a; 1995b; Friend 2001; Maclellan 1997). Summarization tasks are also quite often the post-reading activity in many reading textbooks. In Weir *et al.*'s (2000) report on Advanced English Reading Test Project in Chinese universities, they identified that 6 out of 14 EAP (English for academic purpose) and 3 out of 6 EGP (English for general purpose) textbooks they analyzed had summary writing tasks. Summarization skills are also very often considered a "must" for university students (Davies & Whitney 1984; Holmes & Ramos 1993), and as an integral part of reading instructions (e.g. Aebersold & Field 1997; Urquhart & Weir 1998). The ability to summarize or extract main ideas has also been labelled the "hub in the wheel of reading comprehension" (Axelrod 1975: 383). Summarization skills are important in professional activities such as report writing and production of abstracts (see Ratteray 1985). Because of the unlimited source of information in print and in electronic forms that have become the norm (Brandow *et al.* 1995), readers (native AND non-native) have to read selectively and summarize what they have read. Truly, as Johns (1988: 79) comments, "whatever a person's interest in studying a foreign language, there seems to be no escape from the acquisition and development of summarising skills".

2.3.2 Necessity of developing and measuring summarization skills

Summarization then, not only has a *natural* appeal in the era of communicative language testing because of its close approximation to target language use, it is also a

---

[5] This should be interpreted cautiously, because, in a sense, the summarization task promotes better comprehension. It might also be interpreted as a well-motivated distortion of reading process.

necessary developmental skill that we all should acquire especially due to information inflation. Ever since the computer was invented, human beings have been attempting to simulate summarization processes in order to automatically produce summaries by computer programmes (e.g. Alterman & Bookman 1990; Edmundson 1964; Lehnert 1981, 1984; Luhn 1958; Paulson 1972). Recent developments in automatic text summarization tools[6] such as Microsoft Word's auto-summarize tool, the Automatic News Extraction System (Brandow *et al.* 1995), SimSum (Endres-Niggemeyer 2000), and the AutoExtract Summarizer (Byler 2001) reflect the fact that there are urgent commercial needs and vast markets for a means of summarizing unlimited amount of information on the one hand, set against the fact that satisfactory autosummarization programmes still have a long way to go before they can be relied on with confidence. On the other hand, the development and measurement of summarization skills by human beings thus seem irreplaceable by computer programmes, at least at the current stage of technology.

In summary, summarization not only has natural appeal in communicative language testing, it is also a much needed skill to efficiently use the vast amount of information available. Hence, the measurement of summarization performances as part of reading comprehension (see **2.2**) requires attention from the field of language testing (see also **1.2.3**).

2.4 Practices of using summarization as a measure of reading comprehension

In this section, I review current practices and related problems in the use of "innovative" summarization tasks as a measure of reading comprehension, followed by brief comments on the dearth of empirical studies on TST in the field of language testing.

2.4.1 Current practices and their problems

In language testing, TST as a measure of reading comprehension has historically been rejected, partly because of the potential problems in achieving high scoring

---

[6] See the special issues of *Information Processing & Management*, Vol.31, No.5, 1995; and *Computational Linguistics*, Vol. 28, No. 4, 2002; and also the edited book *Advances in Automatic Text Summarization* by Mani and Maybury (1999).

reliability (see **1.1.2**) and partly because of the confounding effects of writing abilities on the measurement of reading comprehension (see **1.2.3**). TST was therefore considered a "muddied measurement" in Urquhart and Weir's terms (1998: 121, see also Weir 2005) and by many researchers in the field.

Several so-called "innovative" approaches[7] have been used to modify TST on the basis that (a) summarization tasks ***can*** measure reading comprehension and (b) reliability of the measurement can be improved by using the more objective scoring methods. Such approaches include *summary-completion* (Bensoussan 1993; Mossenson *et al.* 1987; Pollitt & Hutchinson 1987; Pollitt 1993; L. Taylor 1996), *summary cloze* (Courchêne & Bayliss 1995; Hughes 1989: 122; Weir 1993: 89-90)[8], *gapped summary*, and *best summary choice* (Huhta & Randell 1996). These different forms of summarization tasks meet some of the psychometric requirements and have been used in large-scale tests such as IELTS (gapped summary or summary-completion), Cambridge Examination in English for Language Teachers (best summary choice), O Grade Examinations in English and in French, and Advanced Level Examinations in French in Scotland, and Cambridge First Certificate in English, Certificate in Advanced English, and Certificate of Proficiency in English.

However, in the different "innovative" forms of summarization tasks, there is a potentially serious problem that summarizers' mental representations of text comprehension may be distorted. In other words, the reading and summarizing process can be affected by the test prompts and methods. For example, although extreme care was taken by the researchers in designing summary-cloze tasks (e.g. Courchêne & Bayliss 1995; Taylor 1996), it was still possible to gain some scores without even reading the source texts[9]. This raises the question of whether

---

[7] Unfortunately, there is also a kind of terminology inflation. Summary completion as defined by one author could be very much the same as summary cloze defined by another author. I simply use the terms as used by the authors.

[8] Though Weir (1993) lists summary cloze as a reading test format, he is cautious about this as a measure of reading comprehension (pp.89-90), and he appears to favour summarization tasks as a measure of writing rather than reading ability (see Chapter 5 of his methodological book on language testing methods for details).

[9] Without too much effort, I got several answers correct in the summary-cloze tasks designed by the researchers (Courchêne and Bayliss 1995; L. Taylor 1996). In the context of the Hong Kong Examination Authority's (HKEA) Use of English Examination, Coniam (1993) found a substantial number of the summary cloze blanks in the examinations could be completed in without reference to the source texts.

summarization tasks are measuring the understanding of the source texts or only the researcher-imposed summaries *per se* with several blanks to be filled in. The presentation of a gapped summary can to some extent exert similar test method problematic effects as MCQ and Cloze (see **1.1.2**). In the case of best summary choice, Huhta and Randell (1996) suggest it is very time-consuming and much more difficult to prepare high-quality multiple-choice summaries than ordinary MCQ. Very often, language constructors fall into the pitfall lamented by Urquhart and Weir (1998)[10]:

> All too often test constructors take considerable periods of time reading and rereading texts and they peel off deeper and deeper levels of meaning. They then give candidates 20 minutes or so to reach the same depth of understanding under exam conditions. This is obviously a nonsense.

Not only is this unfair to test takers, but it may also be considered unethical conduct both in research and instruction. This unequal status between test constructors and test takers *disempowers* test takers, and also *depresses* test constructors. The question is whether a measurement tool can be found to examine test takers' comprehension abilities, while at the same time not distorting too greatly their mental representations of text comprehension. Meanwhile, this tool should not require too great an amount of time from test constructors. The traditional summarization tasks may present a promising alternative (see **2.3**). Although there has been a revival of the use of such integrated reading/writing tasks in large scale and very high stakes language tests like next generation TOEFL, before proceeding to a discussion of their potential, I should note that I am fully aware that TST is not perfect (see **2.5**), giving rise to one of the motives for examining TST in this research.

2.4.2 Scarcity of empirical research on summarization tasks in the field of second language testing

Summarization has been widely used as a research tool to (in)validate discourse processing models of normal and special-needs readers of different age groups, by looking at the differences in the process of summary writing between expert and novice summarizers (e.g. Brown *et al.* 1981; Brown & Day 1983; Brown *et al.* 1983; Garner 1982; Hare & Borchardt 1984; Winograd 1984; Yang & Shi 2003), as well as focusing on the informational (e.g. Johns 1985; Johns 1988; Johns & Mayes 1990;

---

[10] The following quote is cited from Weir *et al.* (2000: 64), originally from Urquhart & Weir (1998). However, no page number was given in Weir *et al.* (2000); I couldn't find the sentences in the original Urquhart & Weir (1998), either.

Kim 2001; Winograd 1984), topographical (e.g. Sherrard 1986) and linguistic analyses (e.g. Basham & Rounds 1984; 1986; Seidlhofer 1991, 1995) of the summaries produced by readers of a range of abilities and native languages. It has also been extensively used as a teaching or study skill to improve students' reading comprehension abilities (e.g. Bensoussan & Kreindler 1990; Cordero-Ponce 2000; Dermody & Speaker 1999; Jitendra *et al.* 1998; O'Mallan *et al.* 1993; Pressley *et al.* 1989; Taylor 1982) and in turn their content acquisition and subject learning (e.g. Arnold 1942; Bretzing & Kulhavy 1979; Friend 2002; Radmacher & Latosi-Sawin 1995; Selinger 1995; Stordahl & Christensen 1956).

In sharp contrast to the wide use of summarization tasks as a scaffold to improve students' reading comprehension abilities and their subject learning, summarization as a language testing method is not well documented, the exception being seventeen studies (Bensoussan 1993; Bueckendorf 1992; Cohen 1993, 1994; Courchêne & Bayliss 1995; Head *et al.* 1989; Huhta & Randell 1996; Kobayashi 1995; Pollitt 1993; Riley & Lee 1996; Sawaki 2003; Scott *et al.* 1996; Shohamy 1984; Stansfield *et al.* 1990; Stansfield *et al.* 1997; Stansfield *et al.* 2000; L. Taylor 1996). Four of these seventeen studies conducted by Stansfield, Scott and their colleagues focused on listening summarization tasks – Listening Summary Translation Exam (LSTE)[11]. Of the other thirteen studies focusing on the measurement of reading comprehension abilities (including those using "innovative" forms), only ***six*** (Bueckendorf 1982; Cohen 1993, 1994; Head *et al.* 1989; Riley and Lee 1996; Sawaki 2003) examined TST as the main focus of their studies. ***Four*** of the six (those by Cohen, Riley and Lee, Sawaki) focused particularly on performances of TST by L2 learners of English for Cohen's studies, French for Riley and Lee's, and Japanese for Sawaki's. However, none of these studies directly address the relationship between summarization performance and L2 reading comprehension abilities, a key focus of this research (see **4.1**).

It is also interesting to note that all the summarization research mentioned above in language testing has adults for participants[12] (except L. Taylor 1996, which looked

---

[11] The research on LTSE was quite similar to literal recall studies because the test takers were asked to write down as many details as possible from the short listening materials they had just heard.

[12] Even though some studies involved adult participants, it was mainly to set an adult standard to assess the children's performances.

at summary completion tasks with Key Stage 3 students in England), while the summarization research most frequently cited in the literature chiefly relate to children's summarization behaviours (e.g. Brown *et al.* 1981; Brown & Day 1983; Brown *et al.* 1983; N. Johnson 1983).

Taking into consideration the promises (see **2.3**), past rejection and recent revival of summarization tasks in large scale language assessment, more empirical evidence is badly needed to gain greater understanding of such tasks to inform a decision on rejection or revival. As Kim (2001: 570) laments, "we currently do not have a clear understanding of what our students do when they summarize an L2 text".

2.5 Further problematizing summarization tasks: a four-component framework

Although there is a dearth of research on TST in language testing, the host of publications in the related areas of education, linguistics and psychology provide sufficient material to *re*-view summarization tasks, taking a bird's eye view. Drawing from a number of such studies (Baumann 1986; Brown & Day 1983; Brown *et al.* 1983; Hare & Borchardt 1984; Hidi & Anderson 1986; Kintsch & Kozminsky 1977; Kintsch & van Dijk 1978), I selectively identify and further problematize some key issues which form the bases of the four-component IFOE (*input*, *filter plant*, *output*, *evaluation*) framework for using TST as a measure of reading comprehension (see Figure 2.1 at the end of this chapter). This section forms the major part of literature review on TST in the order of *evaluation*, *input*, *output* and *filter plant*. In particular, the key factors such as what and how to assess the quality of written summaries (**2.5.1** *evaluation*), source text type and text presentation mode (**2.5.2** *input*), language and language order used to produce the written summaries (**2.5.3** *output*), and summarizers' computer familiarity, language abilities and summarization strategies (**2.5.4** *filter plant*) are discussed taking an ecological and organic approach. Each factor is treated with equal weight, discussed in terms of its own merit and also its indispensable role and relation to other factors in the ecology of the assessment of summarization performances.

2.5.1 How to evaluate summarization performances

The rating of summary protocols is the most thorny issue and needs to be tackled first (Cohen 1993). Weir (1993) justifiably expresses concern regarding the marking of summary protocols to measure students' writing ability, drawing the attention of test constructors using this format as a measure of reading comprehension to the issue:

> To assess students responses reliably one needs to formulate the main points contained in the extract, construct an adequate scheme and standardize markers to it using explicit criteria and a script library. Some subjectivity inevitably remains and it is easy to underestimate the difficulty of marking reliably (Weir 1993: 154).

Alderson (1996: 225) holds a similar view on the subjectivity of marking written summaries. The potential for subjectivity and low scoring reliability are probably the main reasons why summarization tasks became unfashionable in the psychometric era (see **1.1.1**). However, taking into consideration the natural appeal of summarization tasks in communicative language testing, these and other difficulties "should not be taken as a reason for rejecting this form of test" (Cohen 1994: 203).

This section will review the methods of developing and implementing assessment criteria to evaluate summarization performances discussed in the literature (see EVALUATION SYSTEM of Figure 2.1 at the end of this chapter). In particular, it will focus on (1) the development of "ideal" summaries using existing models, experts and test takers, (2) key indicators of the quality of summaries, and how to evaluate them.

1) Developing "ideal" summaries

"Ideal" summaries generated from summarization models and/or written by experts are frequently used to evaluate test takers' summarization performances. This section will discuss the pitfalls of these taken-for-granted approaches and introduce the involvement of test takers themselves in developing assessment criteria.

a) Using summarization models to generate "ideal" summaries

i) The summarization models

A number of reading comprehension models which claim explicitly or implicitly

to be able to provide some explanation of the dynamics of summarization have been used extensively in summarization studies. The most prominent are:

- Kintsch and van Dijk's representational situation model (Kintsch 1974; Kintsch & van Dijk 1978; van Dijk 1980; van Dijk & Kintsch 1977, 1983);
- Rumelhhart's story schema (Golden & Rumelhart 1993; Rumelhart 1975, 1977);
- Meyer's (1975) prose structural content hierarchy system;
- R. Johnson's pausal unit analysis (Johnson 1970);
- Trabasso and his colleague's causal model (Trabasso *et al.* 1984; Trabasso & Sperry 1985; Trabasso & van den Broek 1985; van den Broek 1988; van den Broek & Trabasso 1986);
- Lehnert's plot units model (Lehnert 1981; Lehnert & Loiselle 1989).

The summarization process in the "representational situation model" involves the construct of four macro-rules: deletion, generalization, selection and construction, expanded to six and then five by Brown and Day (Brown & Day 1983; Brown *et al.* 1983). Their six rules of summarization are (a) delete trivial material, (b) delete redundant material, (c) substitute a superordinate term for a list of terms, (d) substitute a superordinate action for a list of subcomponents of that action, (e) select a topic sentence, and (f) invent a topic sentence if one does not already exist. Later (c) and (d) were conflated into a single superordination rule. Hare and Borchardt (1984), borrowing from Brown and Day, propose five rules: (a) collapse lists, (b) use topic sentences, (c) get rid of unnecessary detail, (d) collapse paragraph, and (e) polish the summary, in a direct instruction programme for low-income minority high school students to improve their summarization skills.

Rumelhart's schema, a story grammar approach to text analysis, provides a root node for a hierarchical tree structure that expands to arbitrary depth, as the schemata on each level are instantiated and expanded in a recursive manner. According to Rumelhart (1977), a summary is determined by the amount of detail required and by the amount of information dominated by the "nodes" at the bottom of the schemata "tree".

Meyer's system (1975) is based on "Fillmore's (1968) case grammar and Grimes's (1975) semantic grammar of propositions" (Golden *et al.* 1988: 140). The structural content hierarchy of a text is identified on the basis of logical relations

among propositions. A superordinate structure (the top-level) determines the overall structure of a text, and therefore subsumes all the content and relationships in the text.

R. Johnson's (1970) pausal unit analysis involves a group of fluent readers who divide the passage into pausal units, i.e. those points where a speaker would pause. Next, another group of fluent readers rate the importance of each pausal unit to the theme of the passage on a four-point scale. The readers eliminate the first quarter of the units they consider least important, repeating the procedure twice more until only one quarter of the pausal units are left. Though Johnson's pausal unit analysis lacks strong theoretical foundation, it has proved quite practical in research (e.g. Cavalcanti 1987).

In the causal model, the importance of text units is determined by two variables: the number of causal relationships a text unit has with others, and whether or not the unit falls on the causal chains, through three tests of idea units: counterfactual reasoning test, temporal precedence test, temporal co-existence test (Trabasso and Sperry 1985). The more causal relationships a given unit has with other units, the more important it is and the better chance it will be summarized. If the unit lies on the causal chain that connects the opening event and the final outcome, it will be more important and better summarized than if it does not lie on this chain.

Plot units analysis starts with the configurations of primary affect states of the narratives; the configurations consist of primitive and complex plot units "whose overlapping structures allow us to measure the connectivity and symmetry of character interactions", and thus the framework for text summarization is provided (Lehnert 1981).

ii) Capability and practicality of using these models for generating summaries from extended texts

Besides the use of these models (see **i** above) by their "creators", there is a considerable body of publications on the application of the models by other researchers to study summarization performances. There are two broad trends in analyzing source texts to generate "ideal" summaries which in turn are used to evaluate students' summarization performances.

- readers' ratings of perceived importance (e.g. Brown & Smiley 1977; Johnson 1970; Swoope & Johnson 1988; Thomas & Bridge 1980; Winograd 1982, 1984),
- more formal analyses of the text structure or content to formulate the connectivity or hierarchy of the propositions (e.g. Kintsch 1974; Kobayashi 1995; Meyer 1975; Rumelhart 1977; Sawaki 2003; Upton 1993).

The methods proposed by Kintsch and van Dijk (1978), Meyer (1975) and Johnson (1970) seem to be most widely accepted and extensively used to generate "ideal" summaries[13]. However these are not without problems from both theoretical and empirical viewpoints. In this section, I briefly comment on these issues in terms of the capability and practicality of using these models for the current project to conduct rigid[14] text analyses to generate ideal summaries.

Different predictability of the models

Comparative studies on the predictability of the models have shown that the ideal summaries or recalls generated by the models will be quite different from each other if the texts used are structurally different from those on the basis of which the models were developed (e.g. Mills *et al.* 1993).

Time commitments and requirement on strong expertise

As Bernhardt (1991: 202-203) points out, developing a scoring template from a text hierarchy such as Meyer's protocol system is very time-consuming (see also Urquhart and Weir's (1998) comments on the construction of reading comprehension tests, p.19) . It has "traditionally taken between 25 and 50 hours per 250-word text", in addition to half to one hour scoring time:

> Such time commitments as well as expertise needed in instrument development make it difficult to justify the use of such a procedure even in research, let alone in classroom environments (Bernhardt 1991: 202-203).

The requirement for a high level of expertise also calls into question inter- and intra-expert agreement in analysing texts. Mills *et al.* (1993) asked judges

---

[13] Please see pp.83-88 (Seidlhofer 1995) for those studies using van Dijk and Kintsch (1978, 1983) and p.129 for those using Meyer's, mainly during the period from 1980 to 1990.

[14] Schnotz (1983) criticizes the seemingly rigid analysis of text's propositions as a subjective manner in which the propositions are identified and defined.

(presumably experts, although no description of who the judges were is provided: experts or laymen) to re-analyze their eight texts within a span of 10 days, using Trabasso and Sperry's guidelines (1985) and found "in all cases, all judges made a substantial number of changes in the second set of analyses" (Mills *et al.* 1993: 292).

Incapability of handling extended texts

These models have all been developed primarily from short and narrative texts. When being used to analyze lengthy and complex texts, they "appear to be generally no more capable of describing text structures in an adequate manner" (Schnotz 1983: 178). Extended texts, say over 2200 words, make propositional and structural analysis monumentally difficult (besides time commitments and expertise). Because a summary is a discourse in "its own right" (van Dijk 1977, 1980; van Dijk & Kintsch 1977), it is possible to use these models to investigate a *novice's* writing too (Golden *et al.* 1988). However, in the current project I confront the task of analyzing many summaries one by one, requiring an enormous amount of time.

In terms of discourse types, Kintsch (1982) himself has criticized the overwhelming use of his own and other models on narrative texts, although Kintsch and van Dijk's and Meyer's models have also been used in expository texts.

Exclusion of individual reader factors

The majority of the participants involved in the validation of these models were first language summarizers of English, very often psychology students at American colleges, through laboratory-like experiments. Whether these models are still applicable to other participants of different background is questionable.

Since these models were first of all developed to analyze the texts, they are more or less text-driven and fail to incorporate reader factors such as their background knowledge and topic interest. Kintsch and van Dijk admit the importance of individual summarizer factors[15] in several publications (Kintsch 1988; Kintsch & Greene 1978; Kintsch & van Dijk 1978; van Dijk & Kintsch 1977, 1983). They call

[15] When Kintsch (1974) proposed his model, in the early stages he failed to include reader factors; in the late 1970s and early 1980s, he, van Dijk and others revised the model to include reader factors theoretically.

their 1978 model, far more frequently cited in the literature than their 1983 model (N=1391 for Kintsch & van Dijk 1978; N=28 for van Dijk & Kintsch 1983 as on 4th July 2005, according to ISI Web of Knowledge citation statistics), an:

> abstract, structural description of macrounderstanding [which] could hardly provide a sound explication of individual differences and differences in tasks, goals, or interests in the formation of macrostructures. (van Dijk & Kintsch 1983: 192).

However dynamic their revised model (1983) appears, in practice, reader factors are not valued as much as the so-called inherent macrostructures of a text to generate an ideal summary.

Seidlhofer (1995) first conducts a conceptual evaluation of these models (in particular, Kintsch and van Dijks' and Meyer's in Chapter 4 and Chapter 5 respectively). She also uses summaries (of an expository text) produced by Austrian students of English as a foreign language to empirically evaluate whether these students' largely intuitive use of summarization strategies correlate with what the models suggest. A general finding is that there are mismatches between theory and practice and that the models may not be capable of accounting for the summarization performances of foreign language learners.

A number of empirical studies further reveal the importance of incorporating reader factors into analyzing summarization process and product, such as:

- age (e.g., Adams *et al.* 1990; Byrd 1985; Craik & McDowd 1987; Jackson & Kemper 1993),
- cognitive styles of field-dependence/independence (e.g. Crowley 1987; Mast 1988; Rickards *et al.* 1997; Wilson 1984),
- literate expertise (e.g. Cumming *et al.* 1989),
- prior knowledge of the content and structure of the text (e.g. Afflerbach 1990; Balajthy & Weisberg 1990; Barry & Lazarte 1995, 1998; Gauntt 1989; Hadwin *et al.* 1999; Head 1986; Kiewit 1997; Lambiotte & Dansereau 1992; Loyd & Steele 1986; Steele 1985; Swoope & Johnson 1988; Wilson 1984),
- summarization purposes in relation to judgments of priority or importance/interest of information in a source text (e.g., Wade *et al.* 1999; Zuck & Zuck 1984)

Thus, the central role of the characteristics of individual summarizers should be recognized, rather than relying solely on "rigid" analyses of source texts using the models.

b) Using individual summarizers to develop assessment criteria

Another trend in developing assessment criteria is to invite experts to produce "ideal" summaries. However, taking into consideration the important impact of individual characteristics on summarization performances (see above), exclusive use of a limited number of native speaker experts to generate assessment criteria is also questionable. Meanwhile, the involvement of test takers themselves is increasingly valued in developing assessment criteria (see also Chapter 3). Below, I discuss the use of both native speaker experts and test takers to develop assessment criteria. For further discussion on the effects of individuals' factors (e.g. language proficiency, cultural variations) on summarization performances, please see **2.5.4**.

i) Routine use of native speakers in developing assessment criteria

English native speaker experts have widely been used by default to write and validate language test items (e.g., Lado 1961; 1986), to create scoring templates, and to evaluate test takers' performances as judges (e.g., Barnwell 1989; Brown 1991; Shi 2001). However, this is despite arguments concerning "the futility of the definition of native speaker" (Savignon 2003) and considerable ambiguity or vagueness, both conceptual and practical, on such issues as the identity of native speakers of English, the degree of native-speakerness, and the degree of native speakers' superiority over non-native speakers in terms of, for example, their English reading comprehension abilities.

Native speakers have also enjoyed supremacy in the literature on summarization studies. Native speakers have been widely used to evaluate test takers' summarization performances, either as judges or to set standards. For example, Johns (1985) uses ten experts (academic skills reading and writing instructors) to perform the summarizing task "to set a standard" (p.499) for her three experimental groups. Sarig (1989) suggests involving as many interpretations as possible from diverse professional backgrounds and levels of expertise to reach the Meaning Consensus Criterion

Answer. Cohen (1993) uses scoring templates developed from the summaries of nine Hebrew-speaking and nine English-speaking experts to judge Hebrew summaries of Hebrew texts and English summaries of English texts respectively. Similarly, Corbeil (2000) invites five native speakers of English to produce summaries of the English source texts and five native speakers of French for the French source texts to "determine the total number of main ideas in the texts" (p.41). Kobayashi (2002) also uses English native experts to write up summaries of relatively short texts (350 words on average) as her "baseline" to evaluate English summaries written by Japanese learning English.

However, the degree of agreement between and within experts on which main ideas and connecting ideas should be kept in an ideal summary is questionable, not to mention the ambiguity of the difference between main and connecting ideas (see **2.1**). For example, Cohen (1993) finds the summaries of his Hebrew-speaking experts reflect an "80% average agreement", and the summaries of the EFL experts reflect an "85% average agreement". Cohen (1993: 137) however finds "even the experts did not fully agree on which ideas were essential to the construction of a meaningful summary", a finding reflected in Mills *et al.* (1993). The other studies mentioned in the previous paragraph seem to have accepted native speakers' summaries without any concern.

Apart from the potential differences in summarization performances between native speakers experts themselves,

> there is growing evidence that native speakers perform *neither* [original emphasis] uniformlv (sic) well on tests of all aspects of language ability, *nor* [original emphasis] uniformlv (sic) better than do non-natives (Bachman 1990: 248-249).

Bachman further comments that:

> the very concept of 'native speaker' as actual individuals has been rejected by many scholars, and the problems of identifying the characteristics that might be used to define even the prototypical native speaker are virtually impossible to resolve. Furthermore, ... 'native speakers' vary considerably in their control of different aspects of language proficiency, it is unreasonable to consider them as any more than a norm group... (Bachman 1990: 343)

In fact, it is futile to attempt to define what a native speaker is (Savignon 2003). From

a theoretical viewpoint, Davies (2003: 197) convincingly argues that the native speaker in applied linguistics is both a fine myth and a reality:

> We need it as a model, a goal, almost an inspiration. But it is useless as a measure; it will not help us define our goals.

He thinks that the fundamental opposition between native and non-native speakers is a matter of power relations: native speaker membership is determined by the non-native speaker's willingness (or lack of willingness) to assume confidence and identity. In the next section I address this power relationship by discussing the use of non-native test takers to develop assessment criteria.

ii) Hidden values of test takers in developing assessment criteria

The wide use of English native speakers in language testing is encountering increasing epistemological and ethical challenges triggered by postmodernist perspectives (see also **3.1** and **3.2**). Language testing as a social practice (McNamara 2001; Shohamy 2001a, 2001b) would embrace the equal status of native speaker test developers (or the experts) and the normally non-native speaker test takers (or the novice) in the business of an EFL test. One way to address this power relationship is through involving test takers in test development. Student involvement in developing assessment criteria, as proposed by Wolf *et al.* (1991) and Birenbaum (1996), can promote not only ethicality but also positive impacts of tests on language learners, as Bachman and Palmer argue (1996: 32):

> We would suggest that one way to promote the potential for positive impact is through involving test takers in the design and development of the test, as well as collecting information from them about their perceptions of the test and the test tasks. If test takers are involved in this way, we would hypothesize that the test tasks are likely to be perceived as more authentic and interactive, and that test takers will have a more positive perception of the test, be more highly motivated, and probably perform better.

In a study on the effectiveness of summary writing and short-answer questions to improve advanced reading comprehension in a foreign language, Bensoussan and Kreindler (1990) find classroom discussions in which students negotiate the scoring keys of summaries prove to be "extremely valuable". The negotiation motivates their students (179 freshmen in the advanced reading course of EFL at Haifa University, Israel) to become intensely involved with the text and more critical of their responses. Similarly, Cohen (1993: 144) observes that one of his raters "felt that using a key

based on the judgment of 'experts' skewed the assessment away from the level of students being assessed", and therefore he suggests a compromise for future research, namely building a rating key from both test takers and examiners, as in Bensoussan and Kreindler (1990).

The use of test takers themselves to develop assessment criteria is theoretically possible and desirable, and empirically achievable. Kintsch and Kozminsky (1977: 497) find that their 48 college student participants, who summarized a text of about 2000 words into 60-80 words immediately after processing the text, "agreed quite well on what to include in their summaries whether they read or listened to the stories", and that "if one takes the propositions most frequently used by the subjects in their summaries and puts them into the right order, one can construct a popular summary for each story" (p. 495).

The extensive body of literature on research into the use of English native speaker experts, and indeed non-native speaker "experts" (e.g., Hill 1997; Shi 2001), in EFL test development and performance judgments overshadows the dearth of empirical studies into the use of (non-native) test takers *per se* in contributing to the development of assessment criteria. Although such potential has already been taken seriously in language teaching (e.g., Bensoussan & Kreindler 1990) and specific purpose language testing (e.g., Douglas & Myers 2000) as part of attempts to develop "indigenous assessment" criteria (Jacoby & McNamara 1999), there is little evidence of research that compares the *actual* use of scoring templates generated from experts and test takers respectively to evaluate the performance of the *same* group of test takers (but see Turner & Upshur 2002). The current project endeavours to fill this gap (see **4.1.1**).

2) Defining key quality indicators and the methods to evaluate them

After deciding who is to develop assessment criteria (native speaker experts and/or test takers themselves), it is necessary to define the key quality indicators of a written summary and the methods to evaluate them. This section critiques various *independent* and *integrated* quality indicators and how to quantify them for language testing purposes. Independent quality indicators refer to the discrete-like

characteristics of a summary, while integrated indicators refer to the summary's overall and holistic-like characteristics[16].

a) Defining independent quality indicators and evaluation methods

The quality of a summary is multi-dimensional. The content coverage, structure and succinctness form its major independent quality indicators. Both weighted and unweighted scoring systems have been used to quantify these indicators. Below, I discuss the defining of these indicators and the two evaluation methods.

i) Defining independent quality indicators

As Rost (1990) argues, evaluating summaries poses serious problems because a summarizer faces a myriad of choices regarding which information to reproduce and a range of strategies to represent the reproduction in a summary. Researchers have focused on analyzing content and structure as the two main quality indicators of a summary. Winograd (1984) evaluates the quality of a summary by examining how the sentences have been transformed from the source texts (*reproduction*, *run-on*, *combination*, *low-level invention*, *high-level invention*). Johns (Johns 1985; Johns & Mayes 1990) uses a similar method. However, she focuses more on the content than structure and develops a scale to measure whether a particular idea unit is a replication or distortion of the source text. Golden *et al.* (1988) develop a rating scale which reflects both the structure and content of summaries. Garner and McCaleb (1985) are more concerned with the percentages of the inclusion and exclusion of the important ideas of different levels. Kirby (Hadwin *et al.* 1999; Kirby & Pedwell 1991; Stein & Kirby 1992) develops a four-level scale to assign importance to the content covered (*theme*, *main ideas*, *important ideas* and *less important ideas*). Other approaches such as topographical (e.g. Sherrard 1986) and linguistic (e.g. Basham & Rounds 1984; 1986; Seidlhofer 1991) analyses are also popular among researchers in understanding the qualitative difference in summarization performances between novices and experts. Some studies combine various aspects of these indicators (e.g. Cumming *et al.* 2005; Kim 2001). Taking Cumming *et al.* (2005) as an example, many discourse features are analysed including lexical (e.g. word length, ratio of different words to total words written) and syntactic complexity (e.g. number of

[16] I deliberately try to avoid using terms such as *analytical* and *holistic* scoring methods, although the independent quality indicators may be quite similar to *analytical*, and integrated to *holistic*.

words per T-unit, number of clauses per T-unit), rhetoric (e.g. quality of propositions, claims, warrants, and oppositions in argument structure), and the pragmatics (e.g. orientation to source evidence) of the written discourse produced in integrated reading/writing tasks for the field test of next generation TOEFL.

Another line of research in evaluating the quality of written summaries has focused specifically on content and the structural relationship between source texts and summaries, for example, Winograd (1984), Johns (1985; Johns & Mayes 1990), and Sherrard (1986). Stein and Kirby (1992: 224) suggest the relationship between source and summary, for example the extent to which a summary contains a verbatim copy or integration of several ideas from different locations in the source text, reflects the depth of the comprehension process of summarizers.

A summary is a discourse "in its own right". Some summaries may be written succinctly, some may be bullet-pointed somewhat like a note, and not concise or coherent. The succinctness of a summary is not only related to the coverage of information when the length is held constant, but also reflects the summarizer's ability to use the language and its syntactical rules. It is an important indicator of the quality of a summary and the language ability of the summarizer. Bensoussan and Kreindler (1990: 59) use a "bonus" system to reward succinct summaries. Suppose one student who included 3 propositions out of 5 received a grade of 60%, "if the summary were written succinctly, the total score was 65%. The inclusion of additional information penalized students from receiving the five-percent bonus for writing to the point".

ii) Weighted or unweighted systems

In the literature on summarization studies, both weighted and unweighted systems are used to assign scores for the inclusion or exclusion of particular content such as "idea units". The weighted system credits points for units/propositions according to their position in the hierarchy of the source text or perceived importance, for example, one point for the least important and four for the most important (e.g. Hadwin *et al.* 1999; Johnson 1970 as the most cited one; Kirby & Pedwell 1991; Stein & Kirby 1992). The unweighted system, however, credits every unit/proposition equally, regardless of its level of importance. For example, in unweighted partial credit system, complete inclusion of a scoring unit is assigned two points, an

incomplete inclusion one, and exclusion zero. In unweighted dichotomous system, the scoring method is simplified to either inclusion or exclusion.

Bernhardt finds very high correlations between total test scores obtained from the weighted and unweighted dichotomous systems and concludes that "there is enough overlap in the scores to argue that both systems are tapping the same behavior" of recall performance (Bernhardt 1991: 216). Deville and Chalhoub-Deville's statistical analyses indicate "there is essentially no difference in the relative total scores whether... [they are] scored dichotomously or are weighted", so:

> researchers and classroom teachers can forgo the weighting system and simply score the protocols dichotomously. Dichotomous scoring will save... the time and effort currently being expended on the process of weighting propositions (Deville & Chalhoub-Deville 1993: 126).

Borderia-Garcia and Oskoz (2001) used the two scoring systems and found the correlations between the two scoring systems were .959.

b) Defining integrated quality indicators and evaluation methods

Independent indicators help to illuminate various aspects of the quality of a summary and also may serve well the specific research focus of particular studies. However, the overall quality of a summary does not easily emerge. It is also very difficult to quantify the quality of summaries for the purpose of measuring the summarizers' reading comprehension abilities. In this section, I describe two holistic assessment criteria and augmentation scoring methods to promote rater reliability – a key issue in holistic assessment (see Huot 1990).

i) Defining integrated quality indicators of written summaries

In the literature, there are two detailed scales for evaluating overall quality of summaries[17], one developed by an individual researcher in Canada, the other by

---

[17] Valette (1977: 252) suggested three scoring criteria: (1) accuracy of summary [5 points = all major elements included; 3 points = most major elements included; 1 point = less than half the major elements included]; (2) intelligibility of summary [5 points = a native speaker could easily understand the summary; 3 points = a native speaker could understand the summary only with effort; 1 point = a native speaker would have serious difficulty understanding the summary]; (3) use of tenses [5 points = all verbs in appropriate tense; 3 points = one to three errors in verb tense; 1 point = four or more errors in verb tense]. In a strict sense, Valette's scoring criteria still focused on the individual characteristics of a summary separately; they did not generate a holistic score as did the other two scales.

Educational Testing Service.

Rivard's (2001: 186) four-point holistic scale[18] for evaluating the language in pupils' written summaries focuses on the *brevity, coherence*, and *effectiveness of conveying meaning*:

> **Features of 4 point summaries**
> - The organization of the summary reveals a concern for clarity and coherence.
> - The pupil expresses himself with ease.
> - The limited number of language errors enhances the quality of the communication.
> - The summary is a reformulation, reflects the original text well without commentary.
> - The pupil provides evidence of a concern for brevity.

> **Features of 3 point summaries**
> - The organization of the summary is clear and generally coherent.
> - The pupil expresses himself clearly.
> - Some minor errors do not damage the transmission of the message.
> - The summary is a reformulation, and reflects adequately the original text without commentary.
> - The pupil provides, overall, evidence of a concern for brevity.

> **Features of 2 point summaries**
> - The organization of the summary sometimes lacks clarity and coherence.
> - The pupil expresses himself in an awkward way.
> - Errors, which are sufficiently numerous to be sources of distraction in the transmission of the message.
> - The summary contains some borrowing from the original text and some commentary.
> - The pupil sometimes shows a concern for brevity.

> **Features of 1 point summaries**
> - The summary lacks clarity and coherence.
> - The pupil expresses himself in an obscure or confusing way.
> - Frequent errors damage the transmission of the message.
> - The summary contains several borrowings from the original text and several commentaries.
> - The pupil rarely shows a concern for brevity.

LanguEdge™ Courseware Handbook for Scoring Speaking and Writing (Educational Testing Service 2002) also provides some useful guidelines for establishing the overall quality of a written summary. For the integrated reading/writing task in the courseware, which is however not exactly the same as the summarization tasks defined in this project (see **2.1**), the scoring guidelines in the handbook emphasise the features of a written product such as *accuracy*, *effectiveness*, and *logicality* of conveying the meaning of the principal ideas of a source text, as well as *appropriateness* of the reader/writer's own language:

---

[18] The original scale was in French. My thanks are due to Dr Richard Kiely who helped with the translation of the scale. However, any errors are mine.

- the accuracy of the convey of "principal ideas" of the source text,
- the accuracy and appropriateness of sentence and word formations,
- the appropriateness of the use of the reader/writer's own language and language from the source text,
- the connections among or logicality of ideas.

It becomes evident from the two holistic scales above that the salient quality indicators of a summary embrace the source-summary relationship, in particular, the accuracy, effectiveness and logicality of conveying the source text's principal ideas in the summarizer's own words.

However, neither independent nor integrated quality indicators *alone* can present the full picture of the characteristics of a summary. A new evaluation scheme is needed to incorporate both independent and integrated quality indicators in order to represent a fuller picture of the features of a summary (see also **4.2.4**).

ii) Augmentation method

An augmentation scoring method involves two stages of rating. At the first stage, the rater assigns an integer-level rating (e.g. A or B) that best describes the level of proficiency represented by a written product. However, it is very likely that the written product contains some elements of an adjacent integer-level (either higher or lower), but not in sufficient abundance to warrant the adjacent rating. Therefore, at the second stage, the rater indicates whether the written product leans towards a higher or lower distribution within that integer level, by using + for higher and – for lower than the benchmark scale (Penny *et al.* 2000). In the case that the rater thinks an integer-level rating reflects the quality of the written product, no augmentation is given.

In theory, increasing the length of scale, within a certain range, is desirable and useful for "boosting inter-rater reliability", as evidenced in some comparative studies using more- and less-point scales (for a review see Penny *et al.* 2000). Cronbach *et al.* (1995) also propose using a kind of augmentation method by allowing raters to assign a decimal for an integer-level to reduce errors. It is not an easy task for a researcher (or test constructor) to define succinctly and accurately the characteristics of each level. Similarly, it is difficult for raters to distinguish a fine difference between the levels. By using the augmentation method, the existing number of ratings (levels)

could be expanded threefold. However, "this expansion is substantively different from simply expanding a scale by a factor of 3" (Penny *et al.* 2000: 150). The expansion of scales by the augmentation method produces the same desirable effects of increasing inter-rater reliability, but imposes a substantially lower memory load on raters than simply increasing scales in the scoring guide. What is more, an augmentation method also reflects the practice of many teachers who give such scores, for example, A^-, C^+, or B, in marking their students' essays.

2.5.2 Effects of text input

Hidi and Anderson (1986: 473) argue that the cognitive demands of summarization are dependent upon "qualities of the text to be summarized, the whereabouts of the text during summarization, and the type of summary to be produced". In this section, I review how the qualities of the source text (see INPUT of Figure 2.1 at the end of this chapter), such as its discourse type, length, readability, presentation mode, whereabouts and organisational features, could affect summarization performances.

1) Text type

The effects of discourse type on recall of first language readers and second language readers are reported in the seminal studies of Meyer and Freedle (1984) and Carrell (1992) respectively. It is evidenced that some types are relatively easier to summarize than others, for example narratives are arguably easier than expository or argumentative texts because people are more familiar with this type of discourse in daily life. Even different degrees of "narrativity" (defined as "type of narrative organization of events") also affects students' use of summarization strategies and their performances, as found in Giora and Shen's (1994) research on senior high school students in Tel-Aviv.

Narrative and expository texts have been extensively used in studies on summarization in the fields of education, linguistics and psychology (see **2.5.1**), as is also the case in language testing research and the majority of the empirical studies on summary writing as a measure of reading comprehension (e.g. Cohen 1993; 1994; Kobayashi 1995; L. Taylor 1996) used expository and narrative texts. Many of them

were re-written for particular research purposes and they were usually short and simply structured (see also the above examples), although a few exceptions used authentic non-abridged and lengthy texts to look into the summarization processes (as opposed to measuring reading abilities) (e.g. Cumming *et al.* 1989). All the texts used in these studies were presented on paper (see 4 below).

2) Text length

Text length has been regarded as an important factor in summarization. It influences partly the density of the summary if its length is held constant. Therefore text length has a strong impact on the cognitive load of the task (Kirkland & Saunders 1991). Text length also influences readers' allocation of their time for reading and for producing summaries. As mentioned above, the texts used in summarization studies are very often relatively short; most of the texts are around 400 words. This raises the question of the motivation and purpose of summarizing *already* short texts on the one hand and the extent to which the findings from such studies can be compared with those in which the summarization process involved extended texts on the other.

3) Readability or summarizability

Readability has traditionally been considered to be one of the factors influencing text comprehension and therefore is also important to consider as one of the factors to affect summarization performance, because comprehension is one of the premises of summarization (see **2.2**). In the vast literature on readability (for a review, see Klare 1984; Weaver & Kintsch 1991), a wide range of factors has been identified as contributing to the readability index, such as "topic progression", "vocabulary load", and "syntactic complexity". However, readability indexing has been criticized because alone it may not capture the whole picture of the difficulty of a text. Whilst acknowledging this, the polemic is beyond the scope of this dissertation. Along with measuring readability of a text, other methods such as expert judgments and test takers' introspections and retrospections have been employed to determine text difficulty in language testing research. However, these methods are also problematic. For example, the debate between Weir (Urquhart & Weir 1998) and Alderson (Alderson 2000) on Alderson's (1990; 1991) research into the dimensionality of reading skills/strategies raises the issue of the use of expert judgments (see also **2.5.1**

for a general discussion on the use of native speaker experts, and Barati 2005)

It should be pointed out that readability is "an interaction between a text and the reader's prose-processing capabilities, rather than … some innate property of a text" (Miller & Kintsch 1980: 335). Furthermore, the readability of a text may not be the same as its summarizability – a key issue in the current project.

The readability of a text may be further complicated when it is presented on a computer screen. The features of a text such as font, colour, and the resolution of the computer can all impact on the text's readability and consequently on summarization performance. The following section discusses the potential effects of text presentation mode on summarization performance.

4) Text presentation modes

In reading tests, the explosion of electronic resources along with the already wide range of printed information used, calls for investigations into effects of the text presentation mode on reading comprehension. The nature and degree of mode effects to be observed will determine the extent to which results from computer-based reading tests are generalisable to target language use domain tasks, which may involve both paper- and computer-based reading (Bachman 2000:9). Alderson holds the same view:

> It is important to know whether processing text on screen is different from processing from print - not only because of the potential fatigue effect due to screen glare, but also because generalisations from screen-based reading to print-based reading may not be justified (and vice versa) (Alderson 2000).

Comparative studies of paper- and computer-based reading comprehension in second language tests are scarce, as Chalhoub-Deville and Deville (1999) and Sawaki (2001) point out in their systematic reviews of the effects of text presentation mode on second language reading comprehension and its measurement. There is no research, to the best of my knowledge, which has looked into the differences between reading to summarize using paper- and computer-based extended texts in the field of language assessment. This research aims to examine the potential effects of text presentation modes on students' summarization performances (see **4.1.1**).

Research in other areas may provide insights to contribute to our understanding of reading-to-summarize computer-mediated and long texts. From the perspective of ergonomics, Dillon (1992) reviews empirical studies on individuals' screen reading (mostly proofreading rather than reading comprehension), and finds that reading from the screen is 20% to 30% slower than reading from paper. Dillon also finds that factors such as the quality of visual image, and the availability and quality of text manipulation facilities are important in determining the effects of text presentation mode, for example if the texts do not fit into one screen and therefore require scrolling or paging. When speed is a requirement of screen-reading, Dyson and Haselgrove (2000) find that there is a greater trade-off in speed-accuracy in understanding of details than in "higher order" questions such as main idea comprehension.

However whether and to what extent these factors of computer-mediated texts have a significant effect on reading comprehension also depends on the readers' familiarity with reading on a computer or the potential for anxiety – a key issue related to text presentation mode (see further discussion in **2.5.4**).

5) Text availability

A further inherent element of the quality of a text is whether and how long the text is available during the summarization processes. Whether a text is presented line by line as in some psychological research, whether a text is removed right after reading, how many times a text is exposed to the readers, and how long a text can be available to the readers, could all have effects on the participants' summarization processes, for example, on allocation of time and attention to details, and on managing the cognitive load of summarization tasks (Kirby & Pedwell 1991; Stein & Kirby 1992). In other words, text availability, or exposure to text, has the potential to affect participants' summarization process and product.

From the learning enhancement perspective, Kirby and Pedwell (1991) and Hidi and Anderson (1986) discuss in detail the benefits of text-absent summarization to facilitate deeper learning process, while text-present summarization encourages a copy-delete strategy and surface processing (e.g. Kirby & Pedwell 1991). However, the results may well be attributable to the fact that Kirby and Pedwell use short texts

(see **4.2.2**). The unavailability of a lengthy text could dramatically increase the cognitive demands of the summarization tasks (Hidi & Anderson 1986). What is more, this unavailability could also impact on summarization performance in terms of the confounding effects of memory. Participants with good memories may well be advantaged. However, it should be pointed out that the availability of a text could also make it possible for the summarizers to simply copy some parts of the text – a key strategy to which novice summarizers frequently turn.

6) Organizational features

Besides text type, presentation mode (print or screen), and length and whereabouts, the organizational features[19] of a text have been found to be particularly prominent indicators of a text's summarizability. In studies on summarization, recall and main idea comprehension (e.g. Carrell 1992; Fletcher 1990; Garner & McCaleb 1985; Gauntt 1989; Hare *et al.* 1989; Lorch & Lorch 1985, 1986, 1995, 1996; Lorch *et al.* 1993; Lorch *et al.* 2001; Powell & Isaacson 1984; Sanchez *et al.* 2001; Schwarz & Flammer 1981), there is general agreement that the macro-organizational features of a text such as headings and subheadings influence readers' summarization process and product (Brooks *et al.* 1983; Lorch & Lorch 1996; Lorch *et al.* 2001; Sanchez *et al.* 2001). It is especially true for those readers who tend to pay close attention to headings and consequently produce the most accurate summaries (Hyönä *et al.* 2002). The presence or absence of text titles also has impact on learners summarization/recall performance (Kim 1989; Schwarz & Flammer 1981). The micro-organizational features such as signal words ("therefore", "because", and "in sum" etc.) and italics also signify the structural and content importance of a proposition in the text (see Garner and McCaleb 1985).

2.5.3 Type of summary to be produced

The cognitive demands of summarization are dependent upon not only the qualities of the source text to be summarized, but also "the type of summary to be produced" (Hidi & Anderson 1986: 473). In this section, I focus on the languages to

---

[19] Organizational features or signals, according to Lorch (1989), are writing devices that emphasize the topics of a text and their organization without communicating new semantic content. These are also termed the structural features of a text in some research. They include headings, titles, illustrations with/without captions, and so on.

be used to produce the summaries and how the summaries are to be presented to raters (see OUTPUT of Figure 2.1 at the end of this chapter).

1) Languages and language users

As identified by Bachman's (1990) facets of test methods, the language of the expected response plays a very important role in summarization tasks. There are apparently two combinations of choices – oral or written, and first or second language – to be made in the summarization of source texts.

The perceived confounding function of writing and reading abilities on summarization performance, especially in a second language, make Alderson raise "the question of whether the first language responses would be more suitable in this form of test [summarization]" (Alderson 1996: 225). In empirical studies, it seems that the use of the first language for summarization tasks is more favoured than a second language (e.g. Cohen 1994; Kim 2001; Scott *et al.* 1996; Stansfield *et al.* 1990; Stansfield *et al.* 1997; Stansfield *et al.* 2000).

However, very few empirical studies have compared the use of first and second languages in summarization tasks (e.g. Lee 1986; Long & Harding-Esch 1978). Lee (1986) strongly favours the use of first language, concluding that a "native-language recall task yields more evidence of comprehension, which might be masked by a target-language recall task" (p.208). Long and Harding-Esch (1978) find that "second language deficit" (p.273) is evident in their participants' performances in both summary writing and recall after listening to two speeches lasting about ten minutes. Half of their English native speakers (n=5) were tested on the English version of one speech and the French version of the other and vice versa for the remainder of English native speakers learning French (n=5). The French native speakers (n=10) were tested in a similarly balanced design. They find their participants who are "highly proficient in their second language" (p.283) nevertheless produce significantly less important information, significantly more false information, omissions and substitutions, and fewer words in the summarization tasks in their second language than in their first language. However, I consider Long and Harding-Esch's (1978) findings questionable because of the very small number of cases involved.

Taking a broader view of research, the benefits of using first and second languages for construction tasks such as open-ended questions (not necessarily summary writing) are also empirically supported in Shohamy (1984) (Hebrew speakers learning English), Lee (1987) (English speakers learning Spanish), and Godev *et al.* (2002) (college level learners of Spanish as an L2). Welling-Slootmaekers (1999) also argues that use of the first language (Dutch) instead of the target foreign language improves the assessment of their pupils' foreign language ability. However, van Elmpt and Loonen's (1998) study reports no significant difference between using the first language (Dutch) and the target language (English) to answer comprehension questions[20]. Similarly, Bensoussan and Kreindler (1990) find that the language in which students respond (Hebrew and English) does not affect the scores.

In research into summarization for other purposes, such as improving reading comprehension and (in)validating summarization models, language effects seem less pronounced than in research into summarization as a measure of reading comprehension ability. This may be due to the fact that the majority of those studies focus on first language summarizers of English (very often in American contexts).

The issue of which language(s) to use for the summarization tasks is apparently linked with the language(s) in which the summarizers are proficient. The summarization process and product of various groups including Arabs (Alhaidari 1992; Ayari 1998; Bensoussan & Kreindler 1990), Brazilians (Cohen 1994; Holmes 1996; Holmes & Ramos 1993), Hebrew-speakers (Bensoussan & Kreindler 1990; Cohen 1993), Japanese (Kobayashi 1995), Koreans (Kim 1995, 2001), and Taiwanese and Minnanese[21] (Mahoney *et al.* 1997; Stansfield *et al.* 1997; Wu & Stansfield 2001) have been studied. However, summarization performances of Chinese students are less well-documented. Furthermore, as Johns (1985) notes, university students' summarizing skills have not been as well documented as those of elementary and secondary students (e.g. Brown *et al.* 1983; Day 1980; Winograd 1984). This is still true about 20 years on, though there is increasing interest in research into university students' summarization processes, for example Yang and Shi's (2003) research

[20] These two research done in the Netherlands are cited from Alderson and Banerjee's (2001) state-of-the-art review in language testing and assessment.

[21] These are just two dialects of Chinese.

involving six MBA students at a Canadian university.

Research shows that summarizers, either young learners or adults, frequently employ a "zero strategy" – verbatim or nearly verbatim repetition of text propositions - and a "selection" strategy – selection of parts of the text in verbatim form (e.g. Brown & Day 1983; Fløttum 1985). There is no reason to assume that the participants in this project do not use these summarization strategies. When it happens in Chinese summarization tasks, the summarizers may translate literally, repeating verbatim the original text but in Chinese. I am well aware of the possible confounding effects of students' translation and Chinese writing abilities on their summary writing in Chinese (see also **2.5.4**). The side-effects of translation and first language writing ability on L2 summarization performance appear to be common-sense. However, no research to date has been conducted to look into these effects.

Closely related to the use of two languages for summarization tasks, summarizers' proficiency levels and other literacy expertise in both languages may also affect their summarization. This issue is further reviewed in **2.5.4**.

2) Handwritten vs. Word-processed

How a summary is to be produced, handwritten or typed, is also an essential consideration nowadays because of the availability of computers, although it might not have been an issue in the studies reviewed by Hidi and Anderson (1986). This section briefly reviews the debates on whether and to what extent the quality of a piece of writing, typed or handwritten, may affect scores.

In the literature on educational measurement, it has been found that not only the order in which an essay-type paper is marked, i.e. rating context variables (Coffman & Kurfman 1968; Daly & Dickerson-Markman 1982; Hales & Tokar 1975; D. Hughes *et al.* 1980), but also handwriting and quality of presentation (Briggs 1970; Chase 1968, 1979, 1983, 1986; Graham *et al.* 1989; D. Hughes *et al.* 1983; MacCann *et al.* 2002; Markham 1976; Powers *et al.* 1994), influence the scores a handwritten essay-type paper receives. *Ceteris paribus*, raters tend to favour well-handwritten papers which receive higher scores than their poorly-handwritten counterparts.

In second language testing, the quality of handwriting and presentation, very often defined as its legibility, in students' written scripts has also been claimed to affect rating performances, which in turn result in differences in scores awarded (Hamp-Lyons & Kroll 1997). Raters comment on the legibility of handwriting (e.g. Milanovic *et al.* 1996), however, its effects on scores are yet to be empirically established (Alderson & Banerjee 2002).

Charney (1984) finds that quality of handwriting plays a more significant role when raters have restricted time to read and rate a handwritten essay. With time pressure, raters tend to depend on characteristics such as handwriting quality in the essays "which are easy to pick out but which are irrelevant to 'true writing ability'" (*ibid.*, cited in Shaw 2003: 7). Vaughan (1991) identifies handwriting and overall presentation of an essay as a significant factor, second only to the content of the essay, influencing a rater's decision-making process.

Brown (2003) compares 80 IELTS Task-Two essays (40 handwritten essays are re-typed, resulting in 80 essays) which are judged using IELTS band scales; the results show that, contrary to expectations, (*a*) the handwritten scripts are consistently marked higher than the typed versions, and (*b*) "the handwritten scripts with poor legibility [showed] the greatest score difference between versions", which means those test candidates with bad handwriting and poor presentation are advantaged rather than disadvantaged (Brown 2003: 138).

In a similar comparative study investigating the impact of legibility on ratings awarded to FCE handwritten and word-processed scripts, Shaw's (2003) findings are in line with Brown's in that his three experienced raters did not seem to favour the improved legibility in the word-processed scripts. These findings contradict the intuitive belief that well-presented essays tend to receive higher scores than poorly-presented essays, as evidenced in the literature of general educational assessment. However, Shaw justifiably points out that his three raters felt that it would be "difficult not to be influenced by bad handwriting" (Shaw 2003: 10), and that there are a number of advantages of word-processed scripts over their handwritten counterparts (p. 9):

- both strong and weak scripts are easier to read;
- poor handwriting is not penalised. Raters appear not to be unduly influenced by neatness of presentation which is exhibited through handwriting;
- errors of spelling and punctuation are accentuated when typed and are more easily identifiable;
- all scripts look similar in general appearance before reading thereby facilitate rater objectivity;
- typed texts facilitate paragraph identification.

Although mixed, the effects of handwriting and word-processing found in educational measurement literature in general and second language testing in specific clearly call for cautions in the interpretations of test results from either method.

2.5.4 Facets of filter plant

The central player in the IFOE framework is the *filter plant* – the summarizer – that can activate and coordinate interactively the other three components of the framework. Various characteristics of the *filter plant*, such as the purposes for which summaries are written, summarization strategies, language abilities and literacy expertise, topic familiarity and interest, and computer familiarity, all work together to affect summarization performances (see FILTER PLANT of Figure 2.1 at the end of this chapter).

1) Test instructions: audience and purpose

Research in foreign language writing illustrates that the intended audience of a piece of writing plays a very important role, affecting its quality and style. But unfortunately, in most summarization research, participants are simply asked to write a summary after reading a text, and they are not told explicitly for whom and for what purposes they are writing the summary (Hidi and Anderson 1986; L. Taylor 1996). Weir admits that the TEEP (Test in English for Educational Purposes) summarization task (to measure writing rather than reading) "might be enhanced if the candidates were given an explicit addressee for the task" (Weir 1993: 153). As an example of summarization task designed for IELTS, Alderson (2000) clearly states that:

> You are writing a brief account of the eruption of Mount St. Helens for an encyclopaedia. Summarise in less than 100 words the events leading up to the actual eruption on May 18.

It is true that in real-life situations summary writing serves various purposes (Ratteray 1985; Russell 1994). A major distinction is whether it is written for oneself or others. The main intended audience of a "reader-based" summary is *others*, therefore also termed a "*public*" summary by Swales and Feak (1994). A "writer-based" summary is for "*private*" use; the audience is the writer himself/herself, and therefore it is not constrained by any outside factors. It is very often incomparable with summaries of other writers (Hidi and Anderson 1986). In order to compare summarization performances for language testing purposes, "reader-based" summaries may be more able to provide comparable data than "writer-based" summaries[22].

In close relation to the intended purpose/audience of a summary is its length. If a summary is written for oneself, it can be as long as the summarizer thinks appropriate for his/her particular purposes. In a test or research context, however, usually it is the researcher or test constructor that arbitrarily determines the length. In Cohen's (1994) study on summarization, participants wrote their summaries without a word limit or time limit in their first language. In Cohen's other study (1993) on summarization performance in Hebrew, he used an American "common practice" approach (i.e.80-100 words) to determine the length of summary protocols (Cohen 1993: 145, note 3).

There are further challenges to determine the lengths of summaries when two languages are used to summarize the same source text (see also **2.5.3**). It may be particularly the case when the two languages are from different families. Should summarizers be required to use the same number of words, for example, for both English and Chinese summaries? Even though some research has found differences in density between English and Chinese, there is no general or systematic determination of how many words on average a given English/Chinese text will contain to convey the same meaning. It may be up to the individual writer. To illustrate the potential for

---

[22] In Cohen's two published studies on summarization (1993, 1994), there is a contradiction between his views of the frequency of occurrence of "writer-based" and "reader-based" summaries in real life. In 1993 he stated: "Real summaries are usually prepared for others who have not read the text and simply want to know what it is about" (p.132), that is to say, real summaries are usually reader-based. However, in 1994 he declared: "... respondents on a test are usually required to furnish a *reader-based* [original emphasis] summary rather than the *writer-based* [original emphasis] summary that they would most likely prepare in the real world - as when, for example, they make notes on a reading assignment" (p.175). That is to say, real summaries are usually writer based. I am inclined to concur with Hidi and Anderson's (1986) view that most real life summaries are usually writer-based.

difficulty, "we" is only one word in English, but modern Chinese would use two scripts "我們"; "early in the morning" (four English words) is equivalent to two Chinese scripts "清晨". Because of the word limit, some students may resort to using old Chinese, still legitimately in use in some cases. Old Chinese is much denser than modern Chinese, for example, "I get up early in the morning" is equivalent to two old Chinese scripts "晨起", while modern Chinese may need five scripts to convey exactly the same meaning.

2) Cognitive demands, strategy training and group work

Summarization is without doubt a cognitively demanding task (Kirkland & Saunders 1991). Many researchers have found that children have enormous difficulties in summarizing and very often children can only use deletion-copy strategies – a "zero strategy" (e.g. Brown & Day 1983; Day 1980; K. Taylor 1986). It is accepted that summarization is a late developmental skill (Hidi and Anderson 1986). However, even adults also find it challenging, whether in L1 (e.g. Winograd 1984) or L2 (e.g. Cumming *et al.* 1989; Kim 2001; Shih 1992). Direct instruction in summarization strategies helps both children and adults to improve their summarization performance (Brown & Day 1983; Brown *et al.* 1983; Cordero-Ponce 2000; Day 1986; Friend 1995, 2001, 2002; Gajria 1989; Gajria & Salvia 1992; Guido & Colwell 1987). Pressley *et al.*'s (1989) review finds positive effects for summarization training on elementary school children's reading comprehension. However, research also shows that even seventh graders have good knowledge of what summarization is and how to summarize by using basic strategies such as copying and selecting important elements (Brown & Day 1983), and that, although adults are well aware of how to write a good summary, they do have difficulty with the *actual* summarization process. In L2 contexts, this difficulty may be due to low language proficiency (Cohen 1994; Connor 1984; Connor & McCagg 1983), not necessarily to low awareness of summarization strategies. For example, Cohen (1994) attributes his participants' difficulty in distinguishing superordinate, nonredundant material from the rest largely to an insufficient grasp of foreign language vocabulary.

Day (1980) examines the effects of explicitness of summarization strategy training, and finds that easy summarization rules such as "deletion" need no

instruction (classroom teaching of summarization strategies), and that a brief explanation of how to work with difficult summarization rules, such as "superordination", produces immediate, efficient use. Cohen's (1993: 143) simple guided instructions (test directions) on how to produce a good summary had a "mixed effect on the summarizing of native-language [Hebrew] texts but somewhat positive effect on the summarizing of foreign-language [English] texts"; in item-by-item analyses of the summaries, he finds "the guided instructions appeared to be both helpful and detrimental" (*ibid.*). In listening translation summarization tasks designed by Stansfield and colleagues, test takers were given a chance to read brief instructions (test directions) on how to write good summaries, although, unlike Cohen (1993), they did not seem to be interested in investigating whether and to what extent these brief instructions were used by the test takers.

In order to enable students to write better summaries, some studies ask them to work in groups or pairs as a kind of mediation. However, contrary to expectations, the findings suggest that adult EFL university undergraduates and postgraduates (Allison *et al.* 1994, 1995a, 1995b) and first language university undergraduates (Hooper *et al.* 1994) working alone produce significantly more drafts and longer summaries than those working in pairs, when there is no limit on the length of summaries. The effectiveness of individual summarization is also supported by Rybczynski (1987) in a study of the effects of children's (first language sixth grade readers of average and above average reading ability) individual and cooperative summarization on learning outcomes relating to important ideas from a social studies text.

The findings of these studies seem to suggest that in this project (*a*) providing student participants with detailed summarization strategy training, although perhaps desirable for such cognitively demanding tasks, may not be essential since they would already have some knowledge on how to write a good summary in their first language and (*b*) individual rather than cooperative summarization tasks may be more appropriate for language testing purposes.

3) Language proficiency and literacy expertise

Earlier on in the discussion of the type of written summary, I reviewed the use of first and/or second language for summarization (see **2.5.3**). In this section, I focus on

the relationship between summarization performance and language proficiency and literacy expertise in both the first and the second language of a summarizer.

It is "a pity that summarizing has become unfashionable" (Nuttall 1996: 206). This is particularly true of traditional summarization tasks, rejected in large-scale tests as a "muddied measurement", in Urquhart and Weir's terms (1998:121). They have suggested avoiding:

> tasks such as selective summary based on prior reading of texts – where the extended writing involved in task completion might interfere with extrapolations we might wish to make concerning candidates' reading abilities alone (*ibid.*).

Weir reiterates concerns regarding "muddied measurement" throughout his most recent book on test validation.

> ... given that in many places in the world employers, admissions officers, teachers and other end-users of test information want to know only about a candidate's reading ability *per se*, then we must where appropriate address the problems in testing this and try to avoid other constructs, such as writing ability, interfering with its measurement. (Weir 2005: 88)

When Weir (*ibid.*) comments on integrated listening/writing tasks to measure listening comprehension, concerns of "muddied measurement" are again raised as a serious issue:

> In the latter case [testing understanding of a spoken passage through an integrated writing task such as a selective summary of the discourse] the danger of muddied measurement cannot be ignored, i.e., are we testing listening and/or writing? (Weir 2005: 101)

K. Taylor (1986) studies the summarization performances of young American children (4th and 5th graders) in two experiments. He finds students' performance on a standardized reading comprehension test does not predict accurately their ability to find and produce the main idea in their written summaries. He therefore concludes that this study:

> should cast some doubt about the significance of the role of reading in the process of writing a summary. Obviously, we must be able to comprehend what we read to summarize, but apparently summarizing requires certain written language skills which are apart from and may be more complex than mere reading skills. (K. Taylor 1986: 206).

To Cohen's (1994) surprise, he finds that one of his five Brazilian participants (annonymised as Ana) in a small-scale case study receives the highest score from one of the two raters on the three summary tasks although she is "considered the lowest in

proficiency based on her grades and teacher's appraisal of general performance in the EAP course" (p.194). However, his high-proficiency participants do consistently outperform the medium-proficiency ones. It should nevertheless be pointed out that "the test was untimed, and Ana took the longest time" which may have compensated for her weakness in proficiency (p.194). Other factors in terms of the research design may also have played a significant role in Ana's higher achievement than expected, namely that the participants were allowed to use a dictionary for the summarization tasks and/or:

> the teacher's rating of the students' proficiency may have been based more on reading fluency and accuracy in writing the target language than on summarising ability *per se*. (Cohen 1994: 194)

Admittedly, there may be other explanations for Ana's unexpected higher achievement, such as the unreliability of the summary writing test *per se* and the marked differences between the raters' performances (Cohen 1994: 201).

Other empirical research into correlations between summarization and other test methods have thus far produced rather different if not conflicting results. Thomas and Bridge (1980) find a high correlation (r=.80) between their eighth-grade students' cloze scores and the summarization scores. L. Taylor's (1996) summary completion tasks have a high level of correlations (over .73) with independent reading measures (teacher assessment and the national test of English). Head *et al.*'s (1989) study finds that (a) topic interest, writing ability, and summarization training had some degree of influence on their seventh-graders' (first language reader) ability to summarize a social studies text of about 570 words, and (b) that "multiple-choice and summarization measures shared very little overlap in the kinds of text comprehension". They therefore call for caution in the use of summarization as a measure of reading comprehension. However, their conclusion is not that convincing. Looking from another perspective, it may be that the multiple-choice questions have low correlations with other measures of reading comprehension such as summary writing, because "there is some evidence that in tests of reading MCQ tests only exhibit a low correlation with other measures of reading (see Weir 1983)" (Weir 1993: 97). Further questions arise from their research design which, for example, could have prompted the participants to focus specifically on main idea comprehension which in turn could have carry-on effects on their performances in those MCQ focusing on

main idea comprehension of the same texts. Summarization can "promote engagement with a text which leads to better comprehension" (Smith 1988) and therefore the participants may be better off in terms of the scores they receive from the MCQs answered immediately after the summarization tasks.

From a broader conceptualisation of language proficiency and literacy expertise, i.e. without specifying reading and writing abilities as such (see above), Corbeil (2000) and Johns (Johns 1985; 1990), and Cumming *et al.* (1989) study summarization product and process respectively, in order to examine the relationship between summarization performances and summarizers' first and second language abilities broadly defined.

In Corbeil's study (2000), 99 English speaking university students registered in first- to fifth-year courses of French as a second language were asked to summarize an English text and a French text each of around 600 words into 145-word summaries in randomised order. The summaries were then evaluated according to (a) the number of "main ideas" and the total number of "idea units" included and (b) an adapted scale of Johns and Mayes's (1990) which further taps into the quality of idea units, for example, whether they were correct replications (e.g. direct copying at sentence level, combining idea units within a paragraph or across paragraphs, correct invention of idea units) or some kind of distortion of the source (e.g. incorrect replacement of noun or verb phrase, deletion of essential information or addition of inaccurate information, combinations of distortion, inaccurate metastatements, inclusion of personal comments). The main aim is to examine the roles played by the participants' English (first language) summarization skills and by their French (second language) proficiency on their summarization performance in French. The results show that second language proficiency and first language summarization skills *both* contribute to some aspects of second language summarization performances, but to different degrees. It seems that some summarization skills in the second language are "differentially affected" by first language summarization skills and second language proficiency. Some principal findings relevant to this current research are:

- Students' ability to include main ideas in their first language "directly affects" their performance in including main ideas in their second language summarization task, whereas the effects of second language proficiency seem to be less

pronounced than their first language summarization skills.

- In the case of direct copying, both first language summarization skills and second language proficiency have significant effects on second language summarization performances. A "good lexical knowledge of the second language is nonetheless necessary to paraphrase instead of copying verbatim" (p.49).
- Students having a good command of macro-rules of summarization in their first language attempt to do the same in their second language summarization.

Similarly, as in Corbeil's study, Johns (1985; Johns & Mayes 1990) examines the quality of a written summary, using various criteria such as inclusions, replications, and distortions of idea units of source texts in written summaries. In Johns (1985), 54 "underprepared" English native speaker freshmen (defined as those who had a low grade point average in secondary schools and low scores on university entrance examinations), 53 "mainstreamed" or "adept" freshmen, and 21 "advanced" (senior or graduate) students in an upper-division linguistics class for prospective ESL teachers were asked to summarize a short selection from a freshman American history textbook into around 100 words (I counted the original number of words as around 700), with no time limit. It is found that the "underprepared" students include statistically significantly fewer main idea units (generated according to the summaries produced by 10 experts, see also **2.5.1**) than the adept and the advanced students. Substantially more idea-unit level reproductions are made by the "underprepared" students than combinations of macro-propositions in their summary samples, so are distortions at idea-unit level. With the same research questions and approach as Johns (1985), Johns and Mayes (1990) examine the summary protocols of 80 ESL students of two levels[23] of language proficiency (high and low, 40 each). However, in this research, the participants were asked to summarize a 588-word text from a textbook for low-intermediate English for Business students into 85-115 words, within an allotted time period (not stated). It is found that there are a few differences between the high and the low level groups, but not nearly as many as in Johns (1985) comparing the "underprepared" and the "adept" native speakers. For example, the low proficiency students include far more direct copying of idea units than the high group.

---

[23] The low proficiency participants were "registered in 'remedial' ESL reading and writing classes in the Academic Skills Center at San Diego State University, and the high group from "sophomore (advanced) composition classes, designed for non-native speakers" (Johns & Mayes 1990: 256).

However, they do not differ significantly in replications, combinations or distortions of idea units.

Although not explicitly stated, Corbeil (2000) and Johns (1985, Johns and Mayes 1990) seem to imply that the language proficiency of the summarizers *do have* differential effects on their summarization performances according to their scales for evaluating summary protocols.

In Cumming *et al.* (1989), the focus is on the cross-linguistic relationships in thinking processes of summarization between two languages. Fourteen Anglophone undergraduates of French at a Canadian university were asked to summarize two challenging newspaper articles, each 6 pages long, one in English and the other in French, at one week intervals[24]. The participants were at the beginning to intermediate level of proficiency in French. They wrote English summaries of the English text and French summaries of the French text, while thinking aloud their summarization processes. To these researchers, summarization is fundamentally a problem-solving activity. It is found that the thinking processes in summarizing a challenging text in one's second language seems to be fundamentally similar to those involved in summarization in one's mother tongue. The use of problem-solving strategies "relate[s] closely to the literate expertise people have developed" (p.213), correlating also with the qualities of the written summaries in both languages. However, the use of these strategies seems unrelated to their proficiency in French (second language), although the qualities of the written French summaries are also related to participants' levels of French proficiency. As Cumming *et al.* (1989) point out, their research is limited because of the small number of participants in only two tasks.

4) Cultural variations

I have argued that native speaker experts may produce very different summaries from non-native speaker test takers from various angles (see **2.5.1**). This is also true between different cultural groups. In this section I focus on cultural variations in summarization.

---

[24] The time the participants were allotted to finish the tasks was not reported.

Basham (1986; 1987 cited in Cohen 1994) views summary writing as a cultural artefact. Summarizers are required to demonstrate not only their composing strategies in both reading and writing but degree of familiarity with assumptions about the nature of "objectivity" and "display of knowledge" which are implicit in Western academic culture. For example, her Alaska Native participants are found to include a number of features termed "oral", "informal", or "involved", as opposed to those termed "written", "formal", or "content-centered" demonstrated by experienced summarizers. This may reflect in part a lack of experience with the expectations of summarization tasks. Furthermore, she adds that the tendency of Alaska Native students to personalize their summaries is also attributed to a number of cultural values. Similarly, Moore (1997) also finds summarization practices differ across cultural groups of EAP students.

Research findings in this field are much more complex than whether or not there are cultural variations in summarization practices. Shi (2004) finds that third-year Chinese university undergraduates learning English as a foreign language borrowed significantly more texts from the source without appropriate referencing than first-year native speakers of English studying in a Canadian university, when they were asked to produce a written summary. In the study of differences between first and second language readers recalling expository text, Connor (1984) finds recall of higher level ideas may not be affected by their first language background. In a related study on cross-cultural differences in written paraphrases of English expository prose, Connor and McCagg (1983) also find that non-native English speakers' attention to detail and support for generalizations is much weaker than English native speakers, but recall of main points does not vary greatly. It seems that main idea comprehension is less susceptible to indigenous cultural variations in their studies. They further find that non-native English speakers appear to be constrained by the original structure of the source text, but native English speakers feel much freer to rearrange the original propositional order. In their interpretations, these differences are attributable to language proficiency rather than cultural difference. However, cultural variations in summarization practices are supported by Ayari's (1998) study on how Middle East Arabs and other Asians (both having TOEFL scores of over

500)[25] perform in oral and written summarization tasks. He finds the non-Arab Asians' (Chinese, Koreans, and Japanese) preference for written summarization tasks is statistically significantly different from the Arab groups of all language proficiencies. In French-immersion and Francophone secondary schooling contexts (Senior 1 to Senior 4), Rivard (2001) studies 400 students' performance in writing a summary of a science text. It is found that summaries produced by Francophone students are "generally linguistically superior", "generally better organised, stylistically superior", and contain "fewer errors of language" than those of French-immersion students' (p.184). However, by the end of secondary[26], only style differentiates between the written summaries of these two groups (p.184). Although he did not attribute these differences explicitly to cultural variations between the two groups, there seemed to be an implicit claim that French-immersion and Francophone students wrote summaries differently, although "with few exceptions, all [French-immersion] students had been in the immersion programme since Kindergarten" (p.173).

5) Topic interest and familiarity

Readers' interest, either cognitive or personal, in the text topic, for example, was found to be interacting with whether they would perceive a particular element of a text should be included in the construction of a summary protocol and how (Alexander & Jetton 1996; Head 1986; Head *et al.* 1989; Schellings *et al.* 1996; Wade *et al.* 1999). Reader's familiarity with the topic of the text also influenced how s/he summarized or recalled it (Afflerbach 1990; Carrell 1983; Hahn & Smith 1986; Kiewit 1997); however, Swoope and Johnson (1988) found no significant effect of readers' prior knowledge on their written summarization performances of expository

---

[25] There are two interesting experiments looking into the effects of summarization on gains of TOEFL scores (Ward & Xu 1994). The first experiment compares the gains from summarization skills training and use of commercially prepared TOEFL materials, the results showing that there is no statistically significant difference between the TOEFL score gains of the two groups (Group One gaining 7 points on average, n=7 students receiving instruction in summarizing skills, and Group Two gaining 13 points on average, n=14 receiving TOEFL preparation from commercially prepared TOEFL materials). The second experiment compared TOEFL score gains with two additional groups of English-as-a-Second-Language (ESL) students who had been in ESL classes in the United States for 9 months. One group (n=61) reported using summarization skills in class; the other (n=25) reported never using them. Over the 9 months of ESL study, the group using summarization skills had an average TOEFL score gain of 61 points, and the other group's gain averaged 42 points.

[26] This seemingly longitudinal claim was questionable because it was based on the change of summarization performance of students from *different* grades rather than a true-sense longitudinal follow-up of the *same* group of students from Senior 1 to Senior 4.

texts. These two factors of readers' interest and topic familiarity with the text to be summarized were also found to be interacting with their reading abilities. Winograd (1984) found in the eighth graders of his study, poor readers chose sentences that were "interesting" and rich in detail when asked to select the most important sentences in a text. In contrast, fluent readers used text cues and background knowledge to identify important text elements in text. Carrell (1983) found intermediate ESL students were not affected by familiarity with the subject matter they read for recall.

6) Computer familiarity or anxiety

Computer familiarity or anxiety may play a significant role if the source texts for summarization tasks are computer-presented (see **2.5.2**). In relation to this, I review the literature on the possible relationship between test takers' computer familiarity/anxiety and test performance. Due to the lack of studies on the relationship between computer familiarity and summarization performance, the literature review in this section is necessarily set in a broader research context (i.e. beyond language testing and applied linguistics).

Not only does text presentation mode play an important role in screen reading, but readers' familiarity with computers is also an integral factor determining how readers interact with the screen-displayed texts. Carol Taylor and her colleagues (Eignor *et al.* 1998; Kirsch *et al.* 1998; Taylor *et al.* 1998; Taylor *et al.* 1999) investigate the relationship between computer familiarity and performance on computer-based TOEFL, and find no meaningful relationship between level of computer familiarity and level of performance on the computerized language tasks after controlling for English language ability. The research may mitigate language testers' concerns regarding the possible disadvantages to lower computer familiarity test takers of TOEFL, however, it does not actually probe the possible advantages towards higher computer familiarity test takers. For their particular research aims, they conducted a computer tutorial as an intervention scheme to equalize the computer familiarity levels between the "low-computer-familiar" and "high-computer-familiar" examinees, grouped according to their answers to a computer familiarity questionnaire (see also **4.2.2** and **5.1**). Without such intervention, the examinees' performance might be different.

In a study of the effects of computer-based placement test administrations on test anxiety and performance (math, reading, and written English essay) of 72 college undergraduates at an American university, Shermis and Lombard (1998) find age and computer anxiety are statistically significant predictors of reading performance (Nelson-Denny Reading Test Form E), but no predictors (age, gender, computer anxiety, test anxiety, and personality) are statistically significant in terms of performance on the free written English essay (a 500-word essay on a current social issue). In American school contexts, Russell and Haney (2000) find "written tests administered on paper underestimate the achievement of students accustomed to working on computers" in their two experiments (Russell 1999; Russell & Haney 1997).

O'Sullivan *et al.* (2004) investigated university students' (mainly Chinese university undergraduates) written performances in two delivery conditions (computer and paper-and-pencil). They found (*a*) there were no significant differences in the scores awarded for the candidates' written performances under the two conditions, (*b*) the effect of computer familiarity/anxiety on written performance was negligible, and that (*c*) "a similar cognitive process is most probably being employed in completing the writing tasks under different delivery conditions" (p.50).

It is a rather misty picture of the effects of computer familiarity/anxiety on test performance, especially when the incomparability of the so-called computer familiarity index in different research contexts is taken into consideration. Those defined as high-computer-familiar in one study may be in the low-computer-familiar group in research in a different context.

2.6 Summary

In summary, this chapter first of all defined the generic term – summarization – used in my research, and discussed the *premises*, *promises* and *practices* of using summarization tasks as a measure of reading comprehension. It then *problematized*, from the perspectives of language assessment, a wide range of variables that could impact on students' summarization performances. A four-component framework of summarization tasks was proposed and summarized in Figure 2.1.

INPUT

- Text type (***texts of three types***)
- Text length (***2100-2300 words***)
- Text presentation mode (***computer, print***, audio-recorded)
- Number of text(s): ***single***-/multiple-sourced
- Text exposure chances/availability/whereabouts (***present***, absent)
- Readability (***Flesch-Kincaid Grade Level 11-12***)
- Lexical density (by ***VOCD***)
- Syntactic complexity
- Rhetorical conventions
- Summarizability
- Pre-summarization tasks (***reading***, listening, reading+listening)
- Summarization strategy training (full pre-test, ***partial on-site***)
- …

FILTER PLANT

- Reading ability (***L2 measured by FCE & TOEFL***, L1)
- Writing ability (***L2, L1***)
- Translation ability (***L2→L1***, L1 →L2)
- L1 literacy/experience in summarization (*by* ***questionnaire + interviews***)
- Computer familiarity (*by* ***computer familiarity questionnaire***)
- Topic familiarity with the source text (by ***questionnaire + interviews***)
- Cognitive styles (field-dependent, field-independent…)
- Summarizer's age (***20s***)
- Summarization purposes (***reader***-/writer-based, ***public***/private…)
- L2 vocabulary knowledge
- Summarization strategies
- Task mediation (***individual***, pair, group work)
- …

OUTPUT

- Production or selection (summary-choice, summary-cloze, ***summary production***)
- Languages (***L1***, ***L2***, oral, ***written***)
- Presentation mode (Word-processed, ***handwritten*** …)
- Length (***constrained***/free)
- ***Literal*** or critical
- …

EVALUATION SYSTEM

- Content coverage (***RSC***, ***WSP***)
- Summarization strategies employed (copy, deletion, integration…)
- Succinctness (***5% Score***)
- Topographical relationships with the source text (***SSS***)
- Whose criteria (***expert***, ***popular***, auto-summarizer)
- Syntactic complexity
- Lexical density (***D for English summaries***)
- Summaries in 2 languages (differences between L1 and L2)
- Overall quality (***holistic score by augmentation method***)
- Other discourse features
- …

Figure 2.1 IFOE framework for summarization as a measure of reading ability

Note: Items italicized represent research focus of this project and are discussed in detail in Chapter 4.

Arising from my personal philosophy, research interests and professional practice, the key variables of interest for this project include text type and presentation mode (*input*), handwritten/word-processed language (*output*), intended audience, summarization strategy training, cultural variations and the summarizer's computer familiarity and various linguistic abilities such as reading and writing and literate expertise (*filter plant*), and key quality indicators and methods to evaluate them (*evaluation*). There are certainly many other factors that can affect, individually and interactively, students' summarization performances (Figure 2.1). However, in language testing, very often we have to "*underspecify* [original emphasis], both in designing language tests, and in interpreting test scores. That is, when we design a test, we cannot incorporate all the possible factors that affect performance" (Bachman 1990: 31). The aforementioned key variables are examined under various experimental conditions and elaborated in Chapter 4 of Part III, after the discussion of the paradigm and the epistemological bases of this research in the following Chapter 3.

PART III

Research Approach and Design

Part III consists of two chapters. Chapter 3 describes the paradigm and the epistemological bases of this project. Chapter 4 delineates the research questions and hypotheses, research protocols and data collection procedures. At the end of Chapter 4, the methods of evaluating students' summaries are also discussed briefly (see Figure 3.1).

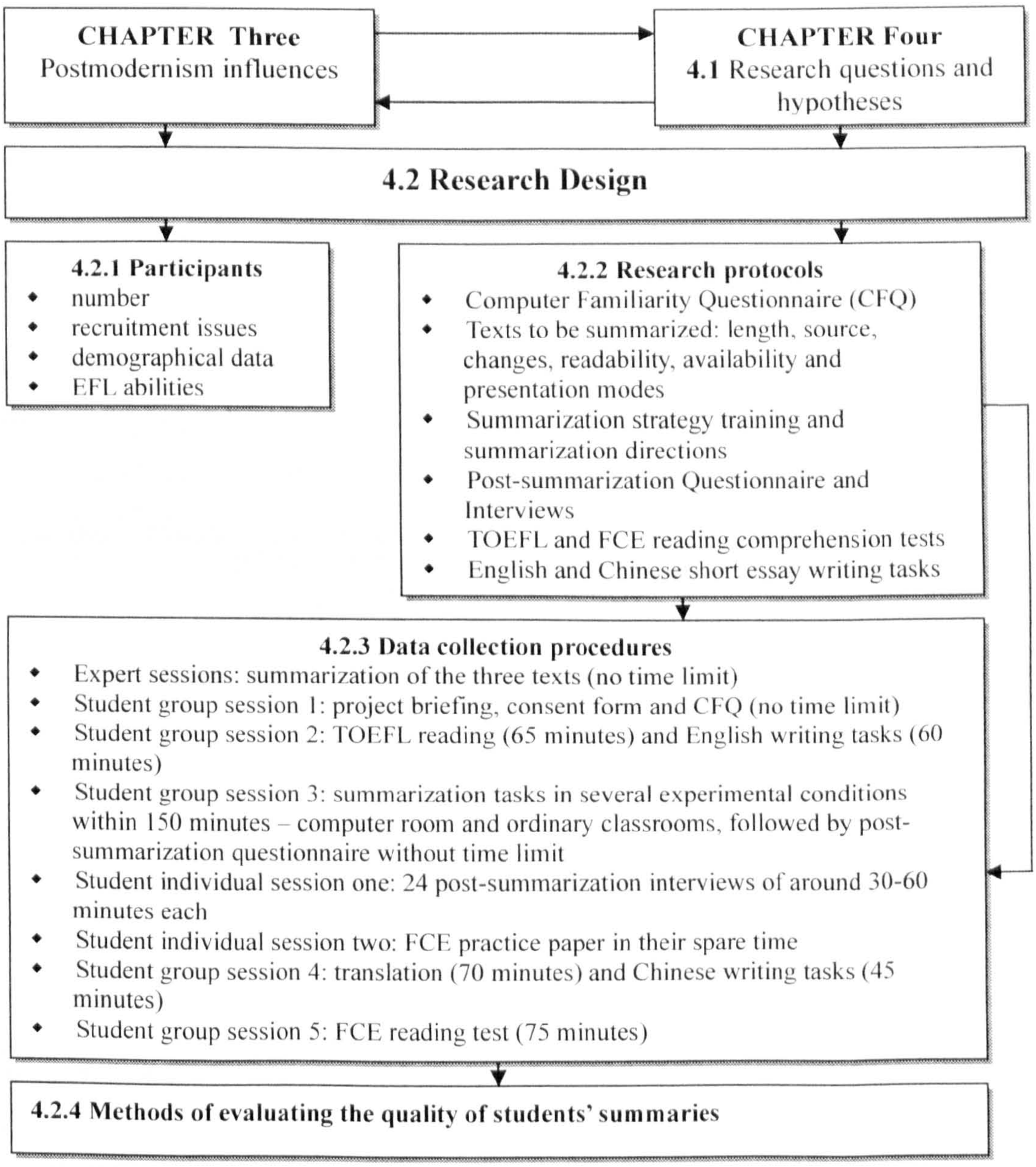


Figure 3.1 Conceptual organization of Part III

CHAPTER THREE

Postmodernist Influences

Though language testing research has been predominantly psychometric and positivistic, the field is constantly changing, influenced by postmodernism[1]. This chapter presents four major themes of postmodernism (*individuality, indeterminacy, interpretation, and immediateness*), reflecting on postmodernist influences on language testing researchers' understandings of (a) text interpretations in reading comprehension tests, (b) ethics of language testing research and (c) applications of integrated quantitative and qualitative research methodologies.

3.1 Four major themes of postmodernism

The debates on modernism and postmodernism as social philosophies are mainly around the themes of objectivity/subjectivity and certainty/uncertainty. When discussing these differences between modernism and postmodernism, dichotomous tables are often used to try to differentiate clearly between the two. Though sorting ideas into these dualisms may itself be contrary to a postmodern approach, a table such as that given below does provide a means of contrasting the two approaches:

modernism	**postmodernism**
certainty and predictability	uncertainty and provisionality
universality(across time and space)	locality and particularity (individualized self experience)
transparency and understandability	indeterminacy
order of nature and structures	ambivalence of human design

Modernism assumes that knowledge is certain, objective, and good and that, in principle, knowledge is accessible to the human mind (see Grenz 1996). The assumption of objectivity leads modernists to claim access to *the* dispassionate and universal knowledge. In addition to assuming that knowledge is certain, objective and

---

[1] Not necessarily adhering to strong postmodernism, I would rather use a mild term: 'postmodernish'. It is very difficult indeed to find a social theorist who admits to being a postmodernist. Even Jean Baudrillard and Jacques Derrida, said by commentators to be amongst those most central to the debate about postmodernism, have both denied being postmodernists! There are also some interesting examples of educational theorists who propose to apply postmodernism perspectives to research, while expressly, sometimes vehemently, declaring they are not postmodernists at all and that they are only suggesting we might use postmodernism perspectives in our research (Blake 1996, 1997; Blake *et al.* 1999).

universal, modernists also assume that it is inherently good for human beings and that there is an optimistic truth seeker as the representative of mankind and the bridge to the ultimate aim: knowing society and implementing beneficial social changes. Postmodernism rejects the foundational assumptions upon which modernism was built (Grenz 1996).

> Postmodernism is a style of thought which is suspicious of classical notions of truth, reason, identity and objectivity, of the idea of universal progress or emancipation, of single frameworks, grand narratives or ultimate grounds of explanation. Against these Enlightenment norms, it sees the world as contingent, ungrounded, diverse, unstable, indeterminate, a set of disunified cultures or interpretations which breed a degree of scepticism about the objectivity of truth, history and norms, the givenness of natures and the coherence of identities (Eagleton 1996: vii)

Though there are many such syntheses of postmodernism (Atkinson 2000, 2002; Bereiter 1994; Bereiter *et al.* 1997; Constas 1998; Greene 1993), the above is certainly not the complete definition of postmodernism. The term 'postmodernism' is open to numerous interpretations – one of the themes, indeed, of postmodernism. It has a multiplicity of referents and yet at the same time its proponents also resist definitions. Therefore, only the generally assumed understandings of postmodernism[2] in educational research, particularly in language testing research, are discussed here. This section explores four of the many interpretative and inter-related understandings of postmodernism: individuality, indeterminacy, interpretation, and immediateness, to characterize the postmodernist challenges to language testing.

Postmodernism treats the world (or actually the worlds) as a text (or texts) which can be read differently by each reader as an individual at different times; therefore, the world can be read differently by each knowing self of the world[3]. There will be as many interpretations as the single self reads the world at different times. There is no reality at all; there are only interpretations, whatever they are. Reality is one of the products of language and is therefore relative. In reality, there is something unexplainable, unfathomable, and "unpresentable" (Lyotard 1984). Postmodernist interpretation is an introspective and individualized understanding. For postmodernists there are infinite interpretations. There is no final meaning for any

---

[2] Postmodernists may well reject or say nothing about whatever definitions "outsiders" impose on them. It is very often one of the reasons why postmodernists are criticized for irresponsible nihilism.
[3] This also implies that a literary text itself can be read differently by readers. This will be discussed further in **3.2.1**.

particular sign, no notion of unitary sense of text; no interpretation can be regarded as superior to any other. Anything goes with anything. "All is difference".

Derrida argues that all that emerges in the knowing process is the perspective of the self who interprets the world. Foucault (1977; 1980) asserts that every interpretation of the world is also an act of power, knowledge is always the result of power.

> Power produces knowledge... Power and knowledge directly imply one another... There is no power relation without the correlative constitution of a field of knowledge, nor any knowledge that does not presuppose and constitute at the same time power relations (Foucault 1977: 27-28, trans. Alan Sheridan, cited in Grenz 1996).

Foucault views power as part of the nature of the social itself (including the knowing self), it is not only a manifestation of something imposed on people from outside, but also constructed from around the self – the individuality or the subjectivity. Power "operates ever more insidiously through disciplinary institutional forms which masks themselves as forms of truth and knowledge" (Olssen 1999:176). Foucault's point is that modern power is centerless, located neither in the State nor in any other single source. Since power co-exists with language, and language is everywhere, power also comes from everywhere. Moreover, wherever there is a power, there is also resistance.

Rorty proposes that we should simply give up the search for truth and be content with interpretations, and that "systematic philosophy" be replaced with "edifying philosophy" which "aims at continuing a conversation rather than at discovering truth" (cited in Grenz 1996; Rorty 1979:393). Similarly Rosenau (1992:8) describes postmodernists as those who simply:

> seek to "locate" meaning rather than "discover" it. They avoid judgment ... they offer "readings" not "observations," "interpretations" not "findings." ... They never test because testing requires "evidence," a meaningless concept within a post-modern frame of reference.

Postmodernism does not assume knowledge is objective, because there is no so-called "uni" of the "universe", and the universe is not mechanistic and dualistic but rather historical, relational, and personal; therefore reality is relative, indeterminate, and participatory. Interpretations of the world are relative to the local participatory community of the self (Grenz 1996). A feature that is common among postmodernists is to reject grand theoretical approaches or "metanarratives" entirely. Rather than

searching for a theoretical approach that explains all aspects of society, postmodernism is more concerned with examining the variety of experiences of individuals and groups and it emphasizes differences over similarities and common experiences. In the view of many postmodernists, the modern world is "fragmented, disrupted, disordered, interrupted" and unstable – and may not be understandable on a large scale (Rosenau 1992: 170). This requires the reader to interpret texts, but not impose on others the reader's interpretation of texts (*ibid.*). Everyone should have the right to interpret themselves and the worlds around them.

> What is needed here is a preservation of differences, a tolerance of ambiguities, and at the same time is a resistance to forced unity (Dallmayr 1987:107).

Postmodernism rests on this ontology of difference that celebrates individuality within a discourse of "fragmentation", "uniqueness" and "specificity". It is also an anti-foundationalism approach that celebrates cultural relativism and plurality: there is no single universal correct view of the interpretations of social existence or text.

The individuality, interpretation and indeterminacy of a postmodernist approach are based on understandings of the immediate. Postmodernism not only breaks with the past, it also denies a concern with the future. The only possible deconstruction of social explanation is limited to the immediate. Even the immediate is very difficult to explore, because of the increasing rapidity of change in the modern world.

Before proceeding to a discussion of postmodernist influences on or challenges to language testing research, this section concludes the philosophical debates between modernism and postmodernism by drawing from Best and Kellner's (1997) explanation of postmodernism as "an emerging paradigm"(p.253) in its Kuhnian sense. They attribute the following elements to this new paradigm:

- a rejection of universal and unifying schemes of thought, the need to emphasize the plurality, difference and fragmentation.
- a rejection of order and fixed meaning, the need to emphasize the play, uncertainty, and ambiguity.
- a rejection of objectivity and truth, the need to favour relativism and simulation

In language testing research and practice, these two sides of the coin (*rejection* and

need) are also becoming increasingly evident, particularly in text meaning constructions and ethics, as well as research methodologies.

3.2 Postmodernist influences on language testing

Language testing research may be considered the most conservative in social sciences as, generally speaking, positivism has been the dominant paradigm in language testing research, aiming at generalizability and the application of research findings in similar contexts. Many language testing researchers have been greatly influenced by psychometrics. However, there is a significant increase of publications recently urging for a reconsideration of the dominant psychometrics. The aim to achieve reliable language testing, that is, to make language testing as objective and predictive as possible, has now been mixed with a postmodernist inclination among researchers (Lewy 1996), with more and more beginning to realize there are some things which are unpresentable (in Lyotard's term). The following section describes the epistemological foundations of this project by reviewing postmodernist influences on language testing from three perspectives: (i) text interpretations in reading comprehension tests, (ii) ethicality and (iii) integration of quantitative and qualitative research methodologies.

3.2.1 Text interpretations in reading comprehension tests: a compromise between modernist and postmodernist approaches

The postmodern perspective on texts "implies that meaning originates not in the production of a text (with the author), but in its reception (by the reader)" (Rosenau 1992: 37), therefore, we can have different interpretations of texts. This understanding of texts presents researchers of reading comprehension with a considerable dilemma. Given the differences in understanding a particular text, the issue is: how are we to determine (if at all) which understanding is "correct", and which is "incorrect"? Postmodernists would say that all understandings are possible and equally right or equally wrong. The notion of correctness proper is inappropriate and theoretically misleading. But, then, what should language test constructors and curriculum and examination boards do?

Almost synchronous with the first postmodernist writings – such as Foucault (1977) and Lyotard (1984) – leading figures in applied linguistics and language

testing research began to realize the indeterminacy of knowledge and the illegitimacy of the so-called "correct" and universal interpretation of a text. Corder (1973), for example, when referring to the interpretation of a spoken text, emphasised that our interpretation of the text may be appropriate or inappropriate to our needs, but never "right" or "wrong". Sarig (1989), from an ethical point of view of language testing, commented on the relativity of meaning:

> the text may have a unique meaning for a learner, be it for developmental or cultural reasons. In such a case, it is not for test developers to pass judgment on the learner's reading of the text, let alone infer that it is necessarily a product of deficiencies in text processing.

Sarig, however, stressed that "Each text opens up a wide range of legitimate, potential meanings, but that potential is not limitless" (*ibid.* 81)

From a postmodernist viewpoint, the meaning of a text is not inherent in the text *per se*, it emerges only as the interpreter enters into dialogue with the text and because the meaning of a text is dependent on the perspective of the one who enters into dialogue with it, it has as many meanings as it has readers or readings. This notion of reader-generated meaning is also the quintessence of interactive and constructive reading theories.

Alderson (2000), while accepting this postmodern notion of text understandings, declares that:

> there must be some acceptance at a common-sense level that some interpretations of text are simply wrong: they do not represent any plausible interpretation of an author's possible intentions.

The problem, then, is how to decide which are acceptable interpretations and which are not. Test constructors will need to be able to answer that question, since it is surely not adequate to say that somebody understands a text only when there is agreement with the test constructor's interpretations[4]. However, Alderson proposes:

> Tests should be open to the possibility of multiple interpretations. Test designers should be as open as possible in the range of different interpretations and understandings they accept (*ibid.*, 29).

[4] Though this is very often the common practice of language testing, especially in reading comprehension multiple choice questions where the readers have to choose one of the best choices already there in order to pass the examination.

3.2.2 Ethics of language testing research: responsible rather than reliable

The notion of acceptable interpretations of test does not only include the interpretations of the test paper *per se,* it also includes the interpretations of test scores, test use, and many other related aspects of test. This leads many language test researchers to think very seriously about the power relations among various test stake-holders, because a single test score has the potential to be interpreted and sometimes manipulated differently by different stake-holders at different times (Rea-Dickins 1997). Language testing is no longer for the common good. It can be used for different purposes, intended or unintended, crude or disguised, when interpreted (Shohamy 2001b). It can be manipulated for various purposes, for example, immigration control to exclude political refugees ("undesirable aliens") on the basis of their inadequate language proficiency (Davies 1997) and can actually be detrimental to a person's life. Hamp-Lyons (1998) suggests that the current interest in ethics in language testing stems from the fact that language testers have had a "positivist" approach to our discipline: that "the object of our enquiry really exists". In a postmodern world, it does not. It is a fiction. She holds that "the growing interest in ethics reflects a post-modern concern with self-evaluation and self-reflection" (*ibid.*, 329). In a position paper discussing Codes of Ethics of the International Language Testing Association, Boyd and Davies (2002: 303) point out that the current proliferation of codes of ethics in different professions and fields "may be viewed as an attempt to provide a modernist response to the challenge of postmodernism", and that "the challenge of postmodernism cannot be easily dismissed".

"Ethics in a postmodern world is local, temporary, and without a logical base" (Fulcher). Similarly, Boyd and Davies (2002: 303) propose that the challenge of postmodernism "may be accommodated by recognizing the need for a variety of Codes of Practice" for different institutions working in very different cultural and political settings.

The awareness of language testing as a social practice has been increasing since Messick's (1989; 1992; 1996) work on "consequential validity" (McNamara 1998). His seminal notion of "consequential validity" has been widely accepted by language testers, particularly when examining the washback effects of language tests, and it is

also the theoretical foundation of Shohamy's critical language testing theory (2001a; 2001b). McNamara (2001: 333) attributes this increase to the "intellectual changes triggered by postmodernism, where models of individual consciousness have been reinterpreted in the light of socially motivated critiques".

The long-lasting lament on the variability of text interpretations in language testing and the increasing debate on ethicality among language testers reflect awareness of the limits of language tests and also language testers' growing "self-evaluation and self-reflection" (Hamp-Lyons 1998) about their responsibility as researchers as well as common individuals in the world. There are more things of which we are uncertain than certain. Even those things we are "certain" about may well be uncertain, and fragmented. More and more language testing researchers are giving up the traditional notion of predictive validity (e.g. Banerjee 2003) and calling for more concern and respect with the present, with the individuals. It is now gradually being acknowledged that test takers are not simply the researchable and researched objects, but that their human rights should be fully respected and their views incorporated as much as possible into test development and validation processes.

3.2.3 Integrated quantitative and qualitative research methodologies

Calls for ethical approaches are being felt in research methodologies. This paradigm shift from the vacuum of pure considerations of reliability and validity to critical language testing (Shohamy 2001a, 2001b) and postmodern sensitivity to power relations among various stake-holders is gradually gaining momentum among researchers and practitioners. A special issue of *Language Testing* (volume 18, No. 4, 2001) on language testing as social practice (e.g. McNamara 2001; Shohamy 2001a), in a sense, reflects this trend or at least indicates language testing researchers' concerns and dissatisfactions with the dominant positivist paradigm that very often employs very sophisticated statistical tools to make "applicable" conclusions. McNamara, May and Hill (2002) point out that:

> The challenge to the positivist epistemology of applied linguistics generally ... and of language testing in particular ... has favored non-quantitative research methodologies ... (*ibid.* 222).

Though quantitative approaches to language testing research are still dominant,

we increasingly see evidence of the integration of qualitative and quantitative research methodologies (see Banerjee & Luoma 1997).

3.3 The postmodernist inclination of this project

This project embraces the aforementioned postmodernist influences. The notion of multiple interpretations of text meanings has significant implications for the project in terms of its use of individuals, both experts and test takers, to generate assessment criteria (see **2.5.1**). Students' voices are greatly valued. They are not simply treated as something researchable and to be researched. Their views on the test tasks and their actual performances are equally cherished (see **2.5.1** and **4.2.2**). In terms of research methodologies, students' *perceptions* of the summarization tasks play as important a role as statistical modelling of their *actual* summarization performances.

However, whilst this project adopts some postmodernist ideas and departs from orthodox positivist approaches, it is nonetheless in the positivist camp in aiming to gain a better understanding of summarization tasks, to apply the findings in future language testing research and practice, and to contribute to the accumulation of knowledge. In addition, for the sake of ease of presentation, the next chapter will present the research design in a positivist manner.

CHAPTER FOUR

Research Design and Data Collection

The conceptual organization of Chapter Four was outlined in Figure 3.1 (p.61). In the first section of the chapter (4.1), I will present the research questions and hypotheses. The second section (4.2) details the design of this project, including research participants and protocols, data collection procedures, and methods of evaluating the key quality indicators of student summaries.

4.1 Research questions and hypotheses

This research aims to examine the effects of some key factors in the IFOE framework on students' summarization performances (see also pp.1-2 and Figure 2.1). The items highlighted in italics and bold in Figure 2.1 represent this project's research focus and are summarized as follows (c.f. **4.2**):

- *input*: extended texts of three types presented either on computer screen or in print and available throughout the reading-to-summarize tasks with partial on-site summarization strategy training provided;
- *filter plant*: familiarity with text topics, computer familiarity levels and language skills i.e. reading, translating, and the writing abilities of the student summarizers' in their 20s in terms of producing reader-based summaries;
- *output*: literal and handwritten summaries in English and Chinese, within a given word limit; but handwritten summaries are Word-processed before marking;
- *evaluation*: the expert and the popular scoring templates to evaluate the student written summary protocols in terms of overall quality and specific properties such as lexical density, topographical relationships with the source text, succinctness and content coverage.

Below, I will present the research questions (**4.1.1**) and hypotheses (**4.1.2**) in relation to the four components of the IFOE framework.

4.1.1 Research questions

RQ1 – Expert or popular scoring template

What are the differences in score variances and students' attitudes between using expert and popular templates to evaluate their written summaries?

This research question aims to test the legitimacy and applicability of the two assessment criteria developed from summaries of experts and students respectively (hence *expert template* and *popular template*) to evaluate the quality of students' summaries. It aims to investigate whether and how students value the popular template, and whether and how the two templates contribute to the differences in the scores that students receive for their summaries.

RQ2 – Muddiedness or organicness of summarization tasks, in respect to language abilities

Are students' summarization performances affected by their other linguistic abilities and, if so, to what extent?

This question focuses on the contributions of students' English reading, English and Chinese writing and translation abilities (from English to Chinese), to their summarization performances. It consists of three sub-questions.

RQ2.1
Do traditional summarization tasks measure the same reading comprehension abilities as the standardized reading tests of TOEFL and FCE and, if so, to what extent?

This research question addresses whether traditional summarization tasks (TST) measure as "well" as standardized reading tests of TOEFL and FCE. It aims to unpack the possible relationships between TST and TOEFL or FCE reading tests, in order to re-visit claims in the research literature that summarization tasks involve significant writing ability and are therefore too muddied to be an appropriate measure for reading comprehension. Will this potential muddiedness necessarily exclude the other side of the coin, i.e., the organicness of summarization tasks?

RQ2.2
Does students' general EFL writing ability affect their English summarization performances and, if so, to what extent? Is general EFL writing ability a determining factor in students' English summarization performance?

In a similar approach to RQ2.1, this question aims to investigate whether English summarization tasks require predominantly English writing abilities?

RQ2.3
Does Chinese summarization of English texts involve students' translation abilities (from English to Chinese), and/or Chinese writing abilities and, if so, to what extent?

In contrast to RQ2.2, this question focuses on the students' *Chinese* summarization performances; it investigates the potential confounding effects of the students' translation and/or Chinese writing abilities on their Chinese summarization performances.

RQ3 – Effects of language and language order

What impact does the use of a different language and language order have on summarization performances and measurement of reading comprehension abilities?

This research question aims to understand (a) the effects of language and language order on summarization performances and (b) which activity, English or Chinese summarization, better reflects students' English reading comprehension abilities.

RQ4 – Effects of text presentation mode and students' computer familiarity on their summarization performances

What are the effects of text presentation mode and students' computer familiarity on their summarization performances?

This research question is two-fold, aiming to examine first the effects of text presentation mode, computer or print, and then of students' computer familiarity on their summarization performances.

RQ5 – Effects of text type on summarization performances

What are the effects of text type on students' summarization performances?

This question examines the use of three quite distinct text types to which students are randomly assigned (see **4.2.3**), and the potential effects of text type on students' summarization performances.

4.1.2 Research hypotheses

The research questions above give rise to the following corresponding hypotheses.

RH1
Students' summary protocols judged against the popular template produce higher scores than when they are judged against expert template. Students like the popular scoring template.

RH2.1
The summarization tasks measure reading comprehension abilities differently from the standardized reading tests in TOEFL and FCE.

RH2.2
English writing ability is a determining factor in English summarization performance.

RH2.3
Chinese and English summarization tasks involve different cognitive processing. Chinese summarization performance has a very high correlation with translation and Chinese writing abilities.

RH3
Chinese summarization tasks are better able to reflect the students' reading comprehension abilities than English summarization tasks.

RH4
a) Text presentation mode makes differential effects on students' summarization performances. *b*) Students with higher computer familiarity perform better to summarize the computer presented texts than those within the same reading ability group but of lower computer familiarity.

RH5
Students find some texts more summarizable and therefore easier than others.

The following table summarizes the research questions and hypotheses within the IFOE framework.

IFOE component	Research questions/hypotheses	Key factors to be addressed
Input	◆ RQ/H4 ◆ RQ/H5	◆ Text presentation modes ◆ Text types
Filter Plant	◆ RQ/H2.1 ◆ RQ/H2.2 ◆ RQ/H2.3 ◆ RQ/H4	◆ Participants' English reading abilities ◆ Participants' English writing abilities ◆ Participants' Chinese writing and translation (from English to Chinese) abilities ◆ Participants' computer familiarity
Evaluation of Output	◆ RQ/H3 ◆ RQ/H1	◆ Output languages (English or Chinese) ◆ Assessment criteria (expert or popular)

Table 4.1 A summary of the key factors to be addressed by the five research questions

It is to be noted that RQ/H4 and RQ/H5 investigate the effects of the **input** factors (text types, and text presentation modes, respectively) on students' summarization performances. Within the **filter plant**, RQ/H2.1, RQ/H2.2, RQ/H2.3, and RQ/H4 examines the effects on summarization performances of students' English reading (RQ/H2.1), English writing (RQ/H2.2), Chinese writing and translation (from English to Chinese) abilities (RQ/H2.3), and computer familiarity levels (RQ/H4). For ease of reporting, the last two component of the IFOE model are combined in "**evaluation of output**". RQ/H3 compares English and Chinese summarization performances. RQ/H1 examines the use of the two scoring templates (expert, popular), from two perspectives – whether and how the students value the popular scoring template, and whether and how the use of the two templates contributes to variance in the scores that students receive for their summarization performances.

4.2 Research design

An integrated *quantitative* and *qualitative* approach is taken to study two parallel datasets in this research – students' *actual* summarization performances and their *perceptions* of the summarization tasks. This section reports: (1) details of student participants, (2) research protocols, (3) data collection procedures, and (4) methods of scoring the summaries (see Figure 3.1).

4.2.1 Student participants

1) Recruitment of students

Altogether, one hundred and sixty-seven (167) university undergraduates were recruited from six intact classes[1] in the same department of a Chinese university. However, the data reported hereafter is based only on the participants (N=157) who finished the summarization tasks. After consultation with their teachers, the students' performances were reported as their mid-term examination scores, as a strategy to encourage them to do their best. There were also other benefits of recruiting intact classes instead of volunteers, such as easier administration of experiments, and maximization of similar background knowledge[2].

2) Participants' demographic information

All 157 students are native Chinese speakers, and have learned English as a foreign language for at least eight years in school and university. None of them have received formal education in English-speaking countries. They are predominantly female (n=130, 82.80% female, n=27, 17.20% male), with one class (Class33) exclusively female students (Table 4.2).

participants information: class * gender * year

Count

year			gender		Total
			female	male	
3	class	31	24	3	27
		32	25	3	28
		33	21		21
		34	22	4	26
		41			
		42			
	Total		92	10	102
4	class	31			
		32			
		33			
		34			
		41	19	8	27
		42	19	9	28
	Total		38	17	55

Table 4.2 Students' demographic information

[1] For ease of reporting the data, the classes were coded as Class31, Class32, Class33, Class34 (for Year Three students), Class41, and Class42 (for Year Four students).

[2] However, studying the same subject does not guarantee that individual participants have similar reading habits or background knowledge. A post-summarization questionnaire was designed which included questions on participants' familiarity with the topics of the texts they summarized and whether and how their topic familiarity affected their reading and summarization performances (see 4.2.2).

The participants were in their early 20s (Adams *et al.* 1990; Byrd 1985; Craik & McDowd 1987), without reported disabilities or recent/permanent brain injuries whatsoever (Brookshire & Nicholas 1984; Nicholas & Brookshire 1995; Wegner *et al.* 1984).

3) Participants' EFL abilities

At the initial stage of targeting potential students, the results of TEM-4 tests were used (see Zou *et al.* 1998). The pass rates of TEM-4 for Class41 and Class42 participants were 100% in 2002. The pass rates for Class31, Class32, Class33, and Class34 were 85.18%, 85.71%, 81.95% and 88.46% respectively in 2003[3]. Besides the higher overall pass rate, Year4 students also achieved higher reading scores (Year4: Mean= 21.7, std. deviation= 2.06; Year3: Mean= 18.0, std. deviation= 2.69, see Figure 4.1).

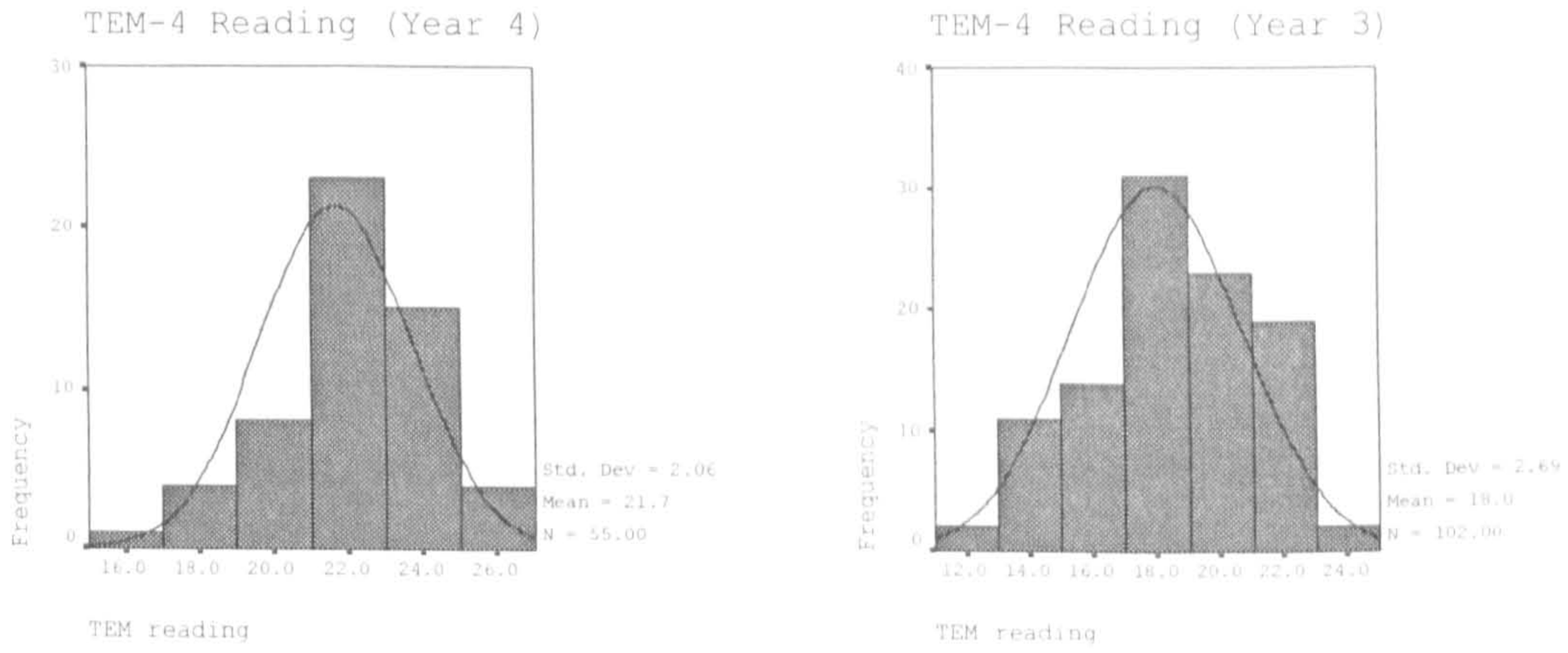


Figure 4.1 Year4 and Year3 students' reading scores in TEM-4

Because data on the possible equivalence between 2002 and 2003 versions of TEM-4 cannot be easily obtained or assumed, students' reading comprehension abilities were further tested using TOEFL and FCE (see **4.2.2** and **5.2**).

This level of language proficiency was chosen because summarization is a demanding task; as Kirkland and Saunders (1991: 108) suggest, "students should not be expected to produce formal, graded academic summaries until they have at least a

---

[3] This data was provided by the university administration. Due to the outbreak of Severe Acute Respiratory Syndrome (SARS) in 2003, TEM-4 was delayed, and therefore no TEM-4 data for Class31, Class32, Class33, and Class34 (i.e. those in Year3) were available before the data collection phase of this research. The TEM-4 for these four classes was actually administered a week before the current research, and the test results were available at the end of 2003.

high intermediate level of proficiency".

4.2.2 Research protocols

This section describes the research protocols in the same order as presented in Figure 4.2.

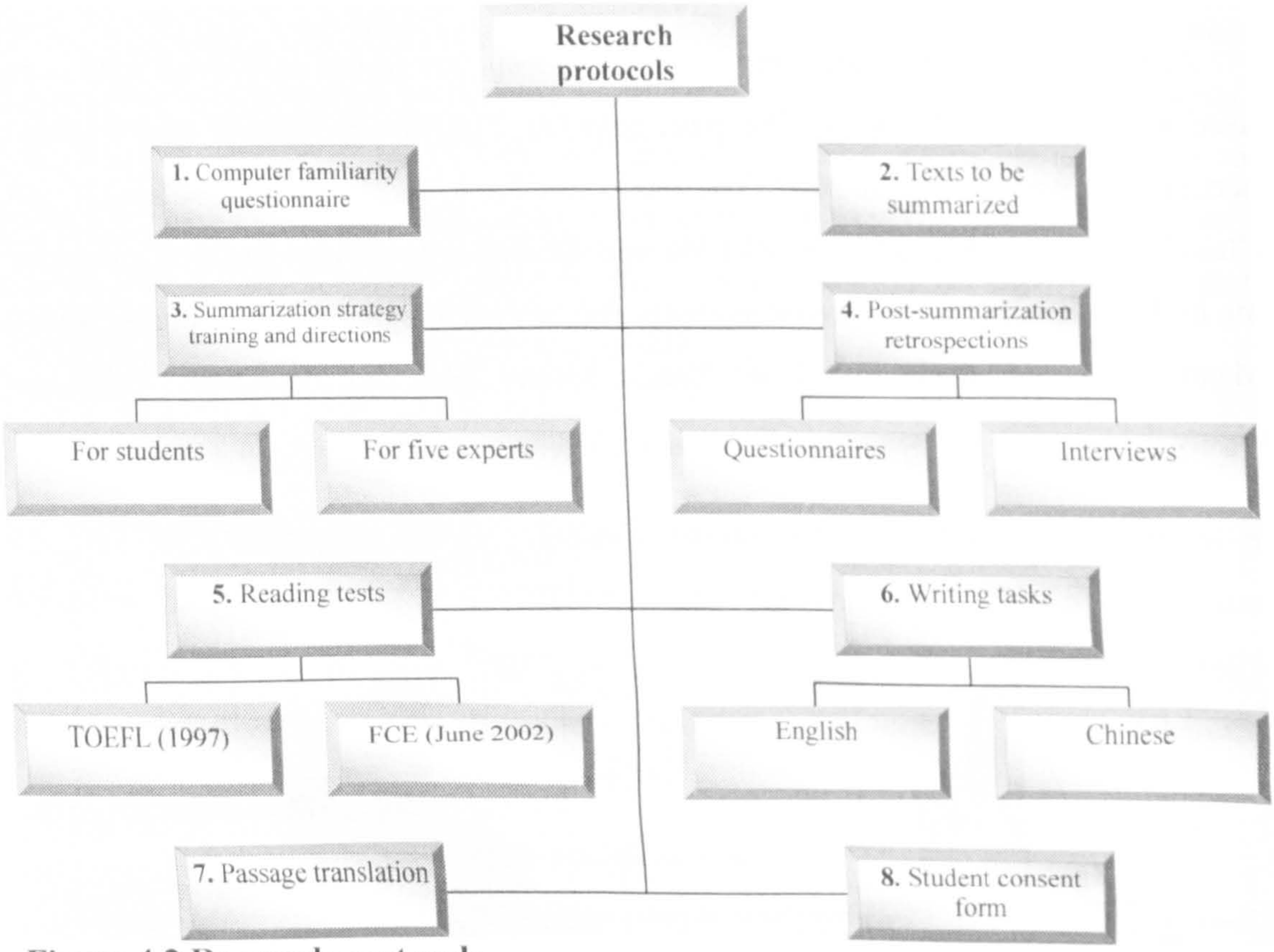


Figure 4.2 Research protocols

1) Computer familiarity questionnaire

As delineated in **4.1.1** above, RQ4 aims to investigate the effects of text presentation mode and computer familiarity on summarization performances. It is therefore essential to gather data on students' computer familiarity. Below, I briefly review measures of computer familiarity and point out four potential limitations of using existing measures. The development of a new computer familiarity questionnaire is discussed.

a) Existing measures of computer familiarity and their limitations

In an extensive review of computer familiarity measures, Kirsch and colleagues (Kirsch *et al.* 1998) identified that definitions of computer familiarity in the literature included the following facets:

- use and experience of computers;
- awareness of related technologies;
- access to computers;
- affective variables, such as attitudes towards and anxiety about computers.

I identify at least four limitations of using existing measures in this project: (i) disagreements over the aspects of computer familiarity, (ii) rapid development of computer friendliness, (iii) lack of opportunities for students to provide constructive responses in their own words, and (iv) discrepancies in end-users.

In their review, Kirsch and colleagues (see Eignor *et al.* 1998; Kirsch *et al.* 1998; Taylor *et al.* 1998; or Taylor *et al.* 1999) noticed that "the concept of computer familiarity is variously defined, based on four related aspects: access, attitude, experience or use, and experience with related technology" (Eignor *et al.* 1998: 2). The lack of agreement on the aspects of computer familiarity poses not only theoretical but also practical and methodological difficulties in designing measures of computer familiarity. Even if there were no such disagreements, until the ETS study in the 1990s, these four aspects had never appeared in combination on a single computer familiarity questionnaire, a lack which the ETS study attempted to rectify.

The ETS measure was the first to incorporate these four aspects into one questionnaire. However, computer technology is becoming more and more user-friendly; the question of "what we use" is becoming less important than the question of "how we use" technology and "how we fix the problems we encounter". Being able to use a computer programme may be less indicative of ability than being able to solve problems when stuck. With widening availability and user-friendliness, the ETS measure designed in the 1990s may not be as appropriate as it used to be. Indeed, as reported by the same group of researchers (Taylor *et al.* 2000), there was an increased use of computers, English Word processing and the Internet in just over 1½ years

(from April 1996 to autumn 1997) among the TOEFL examinees. As Taylor *et al.* pointed out in the 2000 article, "the data reported in this study are already dated", "the situation today can only be extrapolated", and we "should continue to *assess*, rather than *assume*" computer familiarity of today's students (*ibid.*: 584).

In the ETS measure, respondents had no opportunities at all to provide spontaneous comments on their computer familiarity in their own words. Though it might be unrealistic to have open questions in a study as large as that undertaken by the ETS, it is feasible to implement some open questions in the current project, for example, what certificates and training they had in using computers in the past two years.

There are also discrepancies between the intended end-users of different measures. The ETS measure did have Chinese TOEFL test takers as one of its end users; however, there are still several problems. For example, the ETS measure asked about the availability and use of pay-and-display parking meters. This question, I suggest, would not be appropriate for the students in this project, as most of them do not drive and they live in the university accommodation on campus where no such parking meters were available. In addition to such content inappropriateness, the use of the English language could also present some difficulty to some students in understanding technical terms in English.

b) Developing the new computer familiarity questionnaire

Based on the ETS measure, a tailor-made computer familiarity questionnaire (CFQ) was developed, incorporating various indicators of computer familiarity.

i) English and Chinese versions of the computer familiarity questionnaire

The CFQ was originally designed in English and then translated into Chinese and administered to the students in both pilot and main studies. Several back translations were involved. The translations followed the International Test Commission (ITC) guidelines (see Hambleton 2001; Hambleton & De Jong 2003), and were double checked by several Chinese colleagues in linguistics and computer sciences to achieve a faithful translation, semantically and conceptually (Behling & Law 2000; Hambleton & De Jong 2003). It is worth pointing out that the specific ITC guidelines

and the problems of translating one instrument to another language (Behling & Law 2000) may not be fully applicable in this case, because the original English version had already had these students as end-users when it was conceived and designed. It was only for the ease of administration and concern regarding possible misunderstandings of technical terms that the English CFQ was translated into Chinese.

ii) Piloting the Chinese CFQ

The CFQ was piloted in four Chinese universities approximately 13 months before the main study. Altogether 119 students answered the questionnaire. Accidentally, two students from the fourth university, which was the site for the main study, also participated in the questionnaire (ID: 3418, 4128)[4].

iii) Final components of the CFQ

Some slight changes were made to the questionnaire after factor analysis of the data from the pilot study (see **5.1** for detailed procedures). The final version of the CFQ (see Appendix 1) consisted of 33 questions which were grouped according to five categories:

- access/availability to computers (Questions 1 through 5),
- self-assessment of attitude to and ability of using computers (Questions 6 through 12; Question 11 asks the participants the times of examinations already taken using computers) ,
- use of and experience with computers (Questions 16 through 25),
- use of and experience with computer related technology such as ATM (automatic teller machine) and mobile phones (Questions 13 and 14), and
- problem-solving abilities when stuck in using computers (Questions 26 through 31).

---

[4] These two participants' answers in the two versions of the questionnaires were examined in detail, which provided, though accidentally, strong support for the reliability of the questionnaires. Of the 62 (31 x 2) questions, the two students' answers to 31 questions (17 for ID3418 and 14 for ID4128) were identical in the two events. If a participant had given random responses, the probability of getting exactly the same answer for each single question at two events would be 0.25^2; the probability of getting the same answers of half of the 31 questions would be $0.25^{31}(2.1684 \times 10^{-19})$, that is, extremely low.

If a participant had never used a computer, s/he was required to answer only the first 15 questions, as the following questions asking about experience of using computers and problem-solving would be irrelevant. In Questions 1 to 31, four choices were given, such as (a) *very familiar*, (b) *familiar*, (c) *a little familiar*, and (d) *not at all familiar*. Question 32 was "semi-open-ended"; it asked the participants whether they had received any training in their current university studies on how to use computers. If their answer was "Yes", then they were asked to provide further details. Question 33 was an open-ended question; it provided an opportunity for students to provide further information on their computer familiarity such as computer-related examinations they had passed, and attitudes towards using computers.

c) Coding the CFQ answers

Q32 was coded as 1 for *Yes*, and 0 for *No*. Details provided in Questions 32 and 33 were qualitatively analysed. Answers to the remaining questions, except Q30 and Q31, were coded as 4, 3, 2, 1, from the choice on the far left to that on the far right. Q30 and Q31 (Table 4.3) were coded in the opposite direction, i.e., 4 for *never*, 3 for *occasionally*, 2 for *frequently*, 1 for *always*, so that data scoring used the same principle that greater values stand for higher computer familiarity.

How often do you do the following things if you are stuck when using a computer?	always	frequently	occasionally	never
30. turn-off/re-set the computer and start again	○	○	○	○
31. give up	○	○	○	○

Table 4.3 Questions 30 and 31 of the CFQ

2) Texts to be summarized

In this section, details of several key *input* factors of the IFOE framework identified in the literature review (see **2.5.2**) are reported: (*a*) the type and source of the texts, inherent linguistic and organizational properties such as (*b*) length, (*c*) readability, percentage of passivisation, vocabulary density, (*d*) presentation modes, and (*e*) text availability or exposure chances.

a) Text type, source and some necessary changes

Three texts (see Appendix 2) were chosen from quite distinctive sources.

- Text A (Educational History of Southeast Asian Countries, title added by the researcher) is part of an extended review-type article in *The New Encyclopaedia Britannica: Macropaedia: Knowledge in Depth* (volume 18, 15th Edition, 1992, pp.87-88), but it is still a self-contained text even without information from the beginning or the end of the whole review.
- Text B (Let the River Run, original title, with a full colour map of part of the Colorado River system) is from a non-specialized magazine – *New Scientist* (Volume 175, Issue 2362, pp.32-35, 28 September 2002).
- Text C (Work Life Balance Campaign) is from a popular newspaper in the UK – *The Observer* (03 March 2002, p1).

It was assumed, based on my own teaching experience, that the texts were not highly technical or controversial and that these students would have an intermediate familiarity with the topics. Their familiarity with the topics was also obtained through the post-summarization questionnaire and interviews (see **4.2.2.4**).

Although the three texts were from quite distinctive sources: a world-renowned encyclopaedia, a UK popular newspaper, and a non-specialist science magazine and were of different discourse types by intuitive judgement, I am not claiming that they are of absolutely different discourse types. A clear and exact classification of discourse type of these texts is desirable, but not essential for this project, because it focuses only on whether and how a certain text has any potential effects on students' summarization performances (see **RQ5**).

The three texts were kept as close to the original as possible. No changes, except for correcting typing errors in the originals, were made to textB and textC, but several changes were made to the original sub-headings of textA (see **2.5.2.6**). To maximize the comparability of the texts, all subheadings at the page margins of textA were removed. The originally highlighted subheadings of the country names as single lines were submerged immediately in the information of that particular country. A title was added to textA – *Educational History in South-East Asian Countries*, so that all the three texts had a title which has been found effective and informative in enabling summarizers to form a general macrostructure of the source text (Lorch & Lorch 1996; Lorch *et al.* 2001; Sanchez *et al.* 2001).

TextA had a very clear introductory paragraph summarizing the whole text. TextC also had a similar, but less informative and prominent, introductory paragraph. TextB however had no such introductory and summative paragraph. In terms of content organisation, TextA was name organized. In other words, it was organized around the names of the Southeast Asian countries, rather than the attributes of the countries. The content organization of the other two texts was less distinctive than textA. All three texts had titles; but only TextA had subheadings which listed each country by name. The macro-organisational features are summarized in Table 4.4.

	introductory & summative paragraph	content organization	subheadings	heading (title)
textA	yes, and prominent	name organized	yes	yes
textB	No	less distinctive	no	yes
textC	yes, but less informative or prominent than textA	less distinctive	no	yes

Table 4.4 Macro-organisational features of the source texts

b) Text length

The texts were similar in length: textA (educational history) 2270 words, textB (let the river run) 2285, and textC (work life balance campaign) 2120 words. The decision to use this particular text length of around 2200 words was taken for the following five reasons.

Firstly, in the literature most, if not all, of the research into summarization as a language testing task has looked into short texts of usually 400 words (e.g. Kobayashi 2002; Taylor 1996). When the text length is increased, the cognitive loads of summarization tasks may also be increased correspondingly (Kirkland & Saunders 1991). The results obtained by using short texts may not be the same as using longer texts, as the text length influences readers/summarizers' allocation of their time available for reading and for producing the summaries within the given time and word limit of summaries.

Secondly, this length was comparable to the number of words in the reading comprehension section of the standardized tests (see **4.2.2.5**). The FCE reading paper (June, 2001) had 2448 words in the four passages. The TOEFL reading section (1997) had 1615 words in the five passages, plus the questions. To some extent, this comparability of length made it more reasonable to compare students' performances in standardized multiple-choice reading tests and traditional summarization tasks.

However, it should be pointed out that the difficulty of these texts varies.

Thirdly, a 300-350 word summary (i.e., about 15% of 2200 words) also made it comparable with three other written tasks (i.e. English and Chinese writing and translation tasks) used in this research (see **6** and **7**).

Fourthly, in terms of the effects of computer familiarity on summarization performances (RQ4), research studies have found that text length, i.e. the times readers have to scroll the pages, can have dramatic effects on their screen-reading habits, and on differentiating screen from paper reading processes (see **2.5.2.4**). Supposing the text to be summarized fits onto one page on the computer screen, the differences between computer and print reading may well be reduced to a large extent, and therefore attempts to investigate the effects of computer familiarity on the participants' summarization performances could become less meaningful.

Fifthly, in terms of the best approximation between language test and target language use tasks (Bachman 1990), summarizing texts of around 2200 words is more common than summarizing a short text of about 400 words. As an essential skill in university study, summarization is required in almost all aspects of students' academic life. The importance of summarization "can hardly be exaggerated", it is "clearly crucial for education, in pedagogy it is a very common exercise" (Seidlhofer 1995: 2). However, it does not make sense to summarize an already very short text. This approximation between a language test task and target language use task can, to some extent, promote the ecological validity of summarization tasks.

c) Text readability, percentage of passivisation and vocabulary density

The summarizability of a text is partly inherent in a source text *per se* and partly determined by various readers' factors such as their reading and writing abilities, and literacy and cognitive development. To gain deeper knowledge of a text's summarizability, multiple techniques were applied, such as readability indices, percentage of passivisation, and vocabulary density. However, these measures, alone or in combination, still cannot explain fully the summarizability of a text for various reasons, as explained below.

i) Readability and percentage of passivisation

The readability formulae used in this research were Flesch Reading Ease and Flesch-Kincaid Grade Level (F-K Grade Level), both available in MS Word, and also because they were probably the most popular and common readability formulae (see Clapham 1996:92-94; Klare 1974). Besides readability, percentage of passivisation can sometimes add to the "malreadability" of a text (Namukwai & Williams 1988 cited in Clapham 1996: 94).

The F-K Grade level of the three texts is in the region of 11 to 12 (see Table 4.5). This level was chosen for three reasons.

- The readability of some reading passages in FCE and TOEFL was in the region of an F-K Grade Level of 12. It is worth pointing out that the passages in these tests were much shorter ($mean_{FCE}$= 612 words, $mean_{TOEFL}$= 538 words) than the texts for the summarization tasks, and that the readability level of the FCE reading passages were less consistent than TOEFL passages in their readability indices.
- This level of text readability made them comparable to the passage for the translation task.
- A sizable portion of texts in the students' coursebooks for reading comprehension[5] had an F-K Grade Level of around 10.5 to 12.

Another issue of readability for computer presented texts (see **2.5.2**) is the presentation quality and format. The texts were delivered as print layout in Microsoft Word 2000 on recently purchased computers with the Windows 2000 operating system. The font type of the texts was Times New Roman size 12 in the default colour (i.e. black), except that the map in Text A was in full colour. All the computers had a resolution of over 800x600, and were in a very large room/hall – a self-learning centre in the Chinese university.

[5] I measured the readability of the participants' key textbooks in intensive and extensive reading courses, using Flesch-Kincaid Grade Level. The statistics of these measurements are available from the researcher.

ii) Vocabulary density

Although readability formulae such as the ones used in this research do take into consideration vocabulary density (Klare 1984), no single report on vocabulary density is available in these measures. However, vocabulary density is considered crucial for this research, because it reflects the information density of source texts which impose, potentially, the load of information to be processed. In other words, it affects students' decision-making in terms of which information needs to be kept and which not in their summaries.

Several measures of vocabulary density are available, for example the traditional Type Token Ratio (TTR), Guiraud index and D (for a brief review see Daller *et al.* 2003; Jarvis 2002). The recent development of D by Malvern and Richards (1997) relates TTR to token size (N) as a third parameter in the mathematical equation[6]

$$TTR = \frac{D}{N}\left[\left(1+2\frac{N}{D}\right)^{\frac{1}{2}}-1\right]$$

The higher D is, the greater the density of a text. It is beyond the scope of the present research to discuss D mathematically. For the rationale and its mathematical derivations, see Mckee, Malvern and Richards (2000).

The advantages of using D over other measures of lexical density can be summarized as follows (Malvern & Richards 2002: 91):

- it is independent of sample size, thus allowing valid comparisons between varying quantities of linguistic data;
- VOCD takes numerous random samples (without replacement) from the whole set of a text, it takes account of both long-distance and short-distance repetition, and no data remain unused;
- it is more informative because it is representative of the whole of the Type Token Ratio *vs.* token size curve rather than just a single point on it.

Jarvis (2002) compares the accuracy of five commonly used formulae in terms of their ability to model the type-token curves of written narratives produced by EFL learners and English native speakers (Americans), after watching Chaplin's silent film *Modern*

[6] My thanks are due to Professor David Malvern of the University of Reading and Dr Scott Jarvis of Ohio University for confirming that the electronic version of the equation in *Language Testing* was incorrect when I queried it.

Times. He finds that only D (and another measure called U) provides accurate curve-fitting models of lexical diversity, and it seems that Malvern and Richards "have put us on the right track" (p. 82) to measure lexical diversity.

A computer programme VOCD within CLAN (Child Language Analysis) was used to measure D of various texts in the current project. To test the reliability of this measurement, I ran VOCD over 30 times for each text, and found it was reliable (average alpha > .9995).

The readability index, percentage of passivisation and D of the texts (including FCE and TOEFL reading passages) are summarized as follows:

	FCE Reading Passages				TOEFL Reading Passages					Translation Text	Texts to be Summarized		
	1	2	3	4	1	2	3	4	5		TextA	TextB	TextC
Flesch Reading Ease	50.5	73.1	58.3	48.7	46.1	47.4	43.5	40.6	65.4	49.8	14.5	53.3	39.5
Flesch-Kincaid Grade Level	11.8	6.3	9.1	12.0	11.2	11.6	12.0	12.0	10.4	11.9	12.0	10.6	12.0
Percentage of Passivisation*	17	4	12	11	0	6	16	7	23	33	20	12	7
D (vocabulary density)	85.3	151.6	162.8	91.9	89.4	66.9	55.6	80.7	42.7	108.7	84.8	92.2	115.9

Note: the measurement of percentage of passivisation by MS Word was quite unstable. For example, the percentage for "textA" was either 21% (once only) or 20%; that of "textB" was 14% (once only) or 12%; that of "textC" 11% (once only) or 7%.

Table 4.5 Indicators of the summarizability of source texts, FCE and TOEFL passages

Despite undertaking various analyses of the text features, I found it extremely difficult, if not impossible, to have three naturally occurring texts of general interest from three distinctive sources and of the same readability, D, topical structure and signalling devices. The three texts were not claimed to be representative of similar texts.

d) Text presentation modes

The three texts were presented to students in two modes: on computer and on paper, with exactly the same page layout, same font style and size, background colour, and pagination. MS Word was used to display the texts on the computer screen, so that the demands on students' computer familiarity could be minimized because the data from the CFQ piloting indicated that Word was the programme with which the students were most familiar (see also similar finding of O'Sullivan *et al.* 2004 on Chinese university students' computer familiarity). Half of the participants read the texts on computer screen, and half on paper (see **4.2.3**). However, all the summaries were written on paper.

e) Text availability/exposure chances

Throughout the summarization tasks, the source texts were always available, because of two main concerns.

- Firstly, if the text is not available, summarization may be confounded with memory. The texts in this research were much longer than those in previous research. If the texts are absent during summarization tasks, the huge memory load may well prevent students from demonstrating their reading abilities.
- Secondly, removing the text may not be a good representation of real life situations. Students were encouraged to do whatever they liked with the texts, such as underlining, taking notes, and knowledge-mapping, as they would normally do in non-testing events, to enhance the approximation between target language use and test tasks.

3) Directions for the summarization tasks

a) Summarization strategy training as part of the task directions

As part of the task directions, some basic information on what a summary is and how to write a good summary was provided in lieu of explicit summarization strategy training to all participants (see Appendix 3). The six general rules of summarization were specifically highlighted.

- Deletion – delete the trivial and redundant information in the source text
- Superordination – substitute a superordinate term for a list of items, and a superordinate action for a list of subcomponents of that action
- Selection – select the topic sentence that already exists, select the important information
- Invention – invent the topic sentence if it does not exist
- (Re)construction – integrate the important information you've selected and invented into a coherent, concise and self-contained summary that represents and reflects the condensed central ideas or essence of the source text
- Polishing your summary – finish your summary product with best care, make it readable and polished, and faithful to the source text

A Chinese version of this on-site briefing accompanied the English version and students were allocated around 15 minutes to read this information before embarking on the summarization tasks.

b) Other information in the task directions

The directions also explained other requirements on the summarization tasks such as time allowance and length of summary (Appendix 3)[7]. Students were allocated two hours for the first part of the summarization tasks and one hour for the second part. Half of the students summarized the texts first in English and then Chinese, the other half first in Chinese and then English (see **4.2.3**). Both the Chinese and English summaries were to contain 300-350 words, i.e. around 15% of the source texts.

As in the two pilot studies, students in the main study were not told in advance that they were to summarize in two languages. However, they were instructed clearly beforehand that they were to summarize the text they were going to read within time limits. This was considered essential information, based on understanding of the effects of test expectancy on task performance. For example, whether the students were asked to summarize or to recognize the main ideas of meaningful connected discourse would affect their reading process and consequently summarization/recall performances (e.g. Hall *et al.* 1977; Peeck & Knippenberg 1977).

c) Task directions for experts

RQ1 examined the differences in using *expert* and the *popular* templates to evaluate students' summarization performances (see **4.1**). The expert template was generated from summaries written by native speaker experts, common practice in language testing research and practice (see **2.5.1**).

Five English native speaker experts were invited to write summaries. They were all well-educated at prestigious British universities. The decision to invite *five* experts to participate was due to the fact that (*a*) logistically, it is difficult to have more experts, because of limited funding; and (*b*) logically, *five* helps to solve the

---

[7] All these were based on the findings from the small-scale pilot studies in two universities, with 6 Chinese postgraduate students in a British university and 27 undergraduates in a Chinese university which was ranked lower than the "research" university of the main study. The pilot studies investigated (*a*) the effectiveness of the on-site and basic strategy training package, (*b*) the difficulty level of the summarization tasks, (*c*) time needed for the tasks and (*d*) the appropriateness of text features such as readability and topic interest. Details of the research design and findings of the pilot studies are available from the researcher upon request.

disagreements among experts more easily than if an even number were used.

The task directions for the experts were fundamentally the same as those for the students (Appendix 3), except that (*a*) the experts were encouraged to spend as much time as they thought appropriate to both read and summarize the texts; and they were asked to record the time for reading and summarizing the text separately, and (*b*) the experts were required to summarize the three texts, while the students were randomly assigned only one of the three texts; however, the experts were required to read and summarize the texts one by one; (*c*) the experts were only asked to write the summaries in their first language – English; (*d*) the experts were sent both the electronic and print versions of the three texts. It was up to them whether they would like to read and/or summarize on computer or in print.

4) Post-summarization questionnaire and interviews

This project aims to examine both the students' actual performances and their perceptions of the summarization tasks (c.f. **4.1.1**). Two retrospective data collection methods – post-summarization questionnaire and interviews - were used to collect their perceptions. This section presents briefly (*a*) the rationale for using the retrospective data elicitation methods, (*b*) the post-summarization questionnaire (PSQ) and (*c*) the semi-structured post-summarization interviews (PSI).

a) Introspective vs. retrospective elicitation methods

Because impressions of the summarization tasks fade away, it may be ideal to explore the participants' introspections through think-aloud protocols while they are summarizing the texts. Most researchers using think aloud protocols to investigate the test-taking process have been influenced by Ericsson and Simon (1984), who conclude that thinking aloud often has only a negligible effect on test performance. However, some reading researchers find thinking aloud does influence reading performance. For example, Cordon and Day (1996) found thinking aloud has a "significant detrimental effect" (p.288) on students' ability to identify a passage's main ideas. In view of this unclear picture of the effects of thinking aloud processes on reading performance and the practicality of using this approach with over 150 participants in the main study, it was considered that retrospections as early as possible would be the best choice available to look into students' perceptions of the

summarization tasks.

b) Post-summarization questionnaire

The PSQ (Appendix 4) was designed to encompass all the research questions except RQ1. Like the CFQ, two versions of the PSQ were prepared, but only the Chinese version was used. The students filled in the PSQ right after the summarization tasks, without a time limit. The questions mainly focused on students' views on:

- the difficulty of the source texts,
- their familiarity with the general and specific topics of the texts,
- whether and to what extent their topic familiarity helped/hindered their summarization,
- their familiarity with traditional summarization tasks,
- to what extent summarization performances depended on their abilities in EFL reading, English and Chinese writing, English to Chinese translation,
- the use of the two languages to summarize the texts, and which summarization task (English or Chinese summarization) better measured their EFL reading abilities, and
- to what extent their computer familiarity helped with their summarization (only for those who read the texts presented on computer).

There were 30 questions: 17 with five Likert-style choices, 8 with three choices, 2 questions[8] with two choices only, and 3 open-ended questions. Students could select one answer only in the 27 multiple-choice questions.

Answers to the 5-choice questions were coded from 5 for the far left, reflecting *very difficult*, *very helpful* or *very familiar*, to 1 for the far right, reflecting *easy*, *not helpful at all*, or *not familiar at all*. For 3-choice questions (except Q13), they were coded as 1 for the left response, 2 for the middle response, and 0 for the right response.

---

[8] One of the questions (No. 17) was used to check the order of languages used to summarize the text: 4 students (No. 3127, 3406, 3419, 3427) responded incorrectly for their conditions. On checking directly with these four students, their responses were corrected according to their real experimental conditions. Another four students (No. 4305, 4316, 3125, 3408) did not answer Q17: again, by checking directly with these four students, the researcher entered the correct responses. Apart from Q17, all other responses to questions corresponded to the questions in the questionnaire.

Q13 [*which ability contributed most to the Chinese summarization performance?*] was coded as 1 for *English reading ability*, 2 for *Chinese writing ability*, and 3 for *English to Chinese translation ability*. For 2-choice questions, they were coded as 1 for the left response and 2 for the right response. The 3 open-ended questions were analysed using winMAX (Kuckartz 1998, see Appendix 5 for a screenshot of the programme)[9].

c) Post-summarization interviews

To gain more detailed understanding of students' views of the summarization tasks, semi-structured post-summarization interviews (Appendix 6) were conducted as soon as possible after the summarization tasks. In addition to following up the questions in the PSQ, the PSI focused specifically on students' views of using two scoring templates (see RQ1), which were not considered appropriate for elicitation through the PSQ. Twenty-four students were randomly selected (see **4.2.3**) for individual interviews in Chinese, lasting from 30 to 50 minutes. To help interviewees track their memories of doing the summarization tasks, their summaries, original texts (print-outs of the electronic version of the SAVE-AS file) and scrap papers were all available to them during the interviews. The interviews were transcribed and then translated into English for analysis using winMAX (Kuckartz 1998).

5) Reading tests: TOEFL and FCE reading sections/papers

Students' reading abilities were already partly reflected in the results of the TEM-4 tests (see **4.2.1**). However, due to concerns regarding the quality of TEM-4 and the different years in which the students took TEM-4, I used the reading section/paper of TOEFL (1997) and FCE (0100, June 2002) to further measure their reading abilities[10].

TOEFL and FCE were "designed to measure many of the same abilities" (Bachman *et al.* 1995: 15), but represented "radically different approaches to language

---

[9] This programme was also used to analyze the post-summarization interview data, and the expert and student summaries to generate the scoring templates.

[10] The FCE paper was purchased from UCLES, and permission for use was obtained according to the "fair trading" term under Section 29 of the Copyright, Designs and Patents Act 1988. I am grateful to ETS for granting permission to use its 1997 paper free of charge. In order to protect the security of these test papers and the interests of the two organizations, the papers are not listed in the appendices, though I am allowed to do so.

test development" (*ibid.*), reflecting deeper differences between educational measurement traditions in the US and UK. There were some meaningful differences[11] between them on (*a*) lexical knowledge, (*b*) culture contextualization, (*c*) passage length and (*d*) passage genres (Bachman *et al.* 1995: 120-125).

It was hypothesized in the current research that FCE and TOEFL are both measuring reading comprehension abilities, but differently. Correspondingly, when compared with traditional summarization tasks, these differences also exist.

Due to concerns that students may be less familiar with FCE test formats, I provided them with an FCE practice paper, together with the UCLES test-taking recommendations a week in advance. The two tests were administered in separate sessions (see **4.2.3**). The time allowed for FCE was 75 minutes (as in the official test specification), and 65 minutes for TOEFL (10 minutes more than the official TOEFL specification).

6) English and Chinese writing tasks

RQ2.2 and RQ2.3 aim to examine the relationships between summarization performances and English and Chinese writing abilities respectively (see **4.1.1**). In order to measure their writing abilities, students were asked to write argumentative English and Chinese essays of 300-350 words each on topics adapted from the ETS Test of Written English (see Appendix 7): 45 minutes for the Chinese task and 60 for English. The quality of the English and Chinese essays, which were Word-processed before being marked, was evaluated through an augmentation method, according to the same scoring guideline (see Appendix 8).

7) Translation (English to Chinese) task

RQ2.3 also investigates the relationships between summarization performances and translation abilities (English to Chinese). The translation task (Appendix 9) used an English text of 399 words, which had approximately the same readability as the three texts for the summarization tasks (Table 4.5). Students were provided with the

---

[11] In addition to these four meaningful differences, it should be pointed out that the Bachman *et al.* (*ibid.*) study used FCE 1988, which predates major revision in 1996. FCE and TOEFL may have further diverged since 1996.

Chinese equivalents in brackets immediately following five difficult words: anorexia (厌食症), anorexia nervosa (神经性厌食症), psychiatric (精神病学的), syndrome (综合征), and emaciated (消瘦的, 憔悴的).

Students were allowed 70 minutes for the task. The translations, which were Word-processed before being marked, were evaluated in an augmentation method according to:

- LEXICAL MEANING: whether and to what extent the translation faithfully reflected the original passage,
- STYLE, TONE and NUANCES: whether and to what extent the translation adequately reflected the style, tone and nuances of the original passage,
- CHINESE VOCABULARY and SYNTAX: whether and to what extent the translation appropriately and elegantly used Chinese vocabulary and syntax,
- RATER UNDERSTANDING: whether and to what extent the rater experienced difficulty in understanding the translation.

For details of the scoring guidelines and a model translation, see Appendix 10.

8) Student consent form

The student consent form (Appendix 11) served two purposes: to seek the students' written consent in taking part in the research and the proper use of the data to be generated, and to collect demographic data which were cross-checked with the demographic data provided by the university's central administration system. One hundred and sixty-six students signed the consent form. One more student did not sign the form because she was absent when forms were collected, but she sat in some test sessions[12].

9) Summary of research protocols

This research aims to investigate traditional summarization tasks within a proposed IFOE framework, from two parallel datasets: students' *actual* summarization performances and *perceptions* of summarization tasks. Several research protocols

[12] Data from the students (including this one) who did not complete the summarization tasks were excluded from analysis, even though they finished other tasks in the project.

were employed: (*a*) computer familiarity questionnaire, (*b*) three texts for summarization, (*c*) post-summarization questionnaires and interviews, (*d*) standardized reading tests, (*e*) English and Chinese writing tasks, and (*f*) translation task. This section described in detail the design and practical considerations for each research protocol. In particular, the CFQ was developed, taking into account the four potential limitations of using existing computer familiarity measures. Special care was also taken to look into various indicators of a text's summarizability such as its length, type, source, readability, percentage of passivization, vocabulary density, presentation mode, and exposure chances. To ensure the participants, both experts and students, understood what a summary was and how to write a good summary, the directions for summarization tasks included a brief on-site strategy training package. The post-summarization questionnaire and interviews aimed to tap into students' perceptions of the summarization task. Finally the design for the two standardized reading tests, English and Chinese writing and translation tasks, were reported. The procedures of data collection using these research protocols are presented in the next section.

4.2.3 Data collection procedures

In the main study, data were collected from 167 Chinese students who committed about 9 hours each (excluding the time for FCE practice) over a period of 42 days in 7 sessions (Figure 4.3). For details of task directions for experts, see Appendix 3.

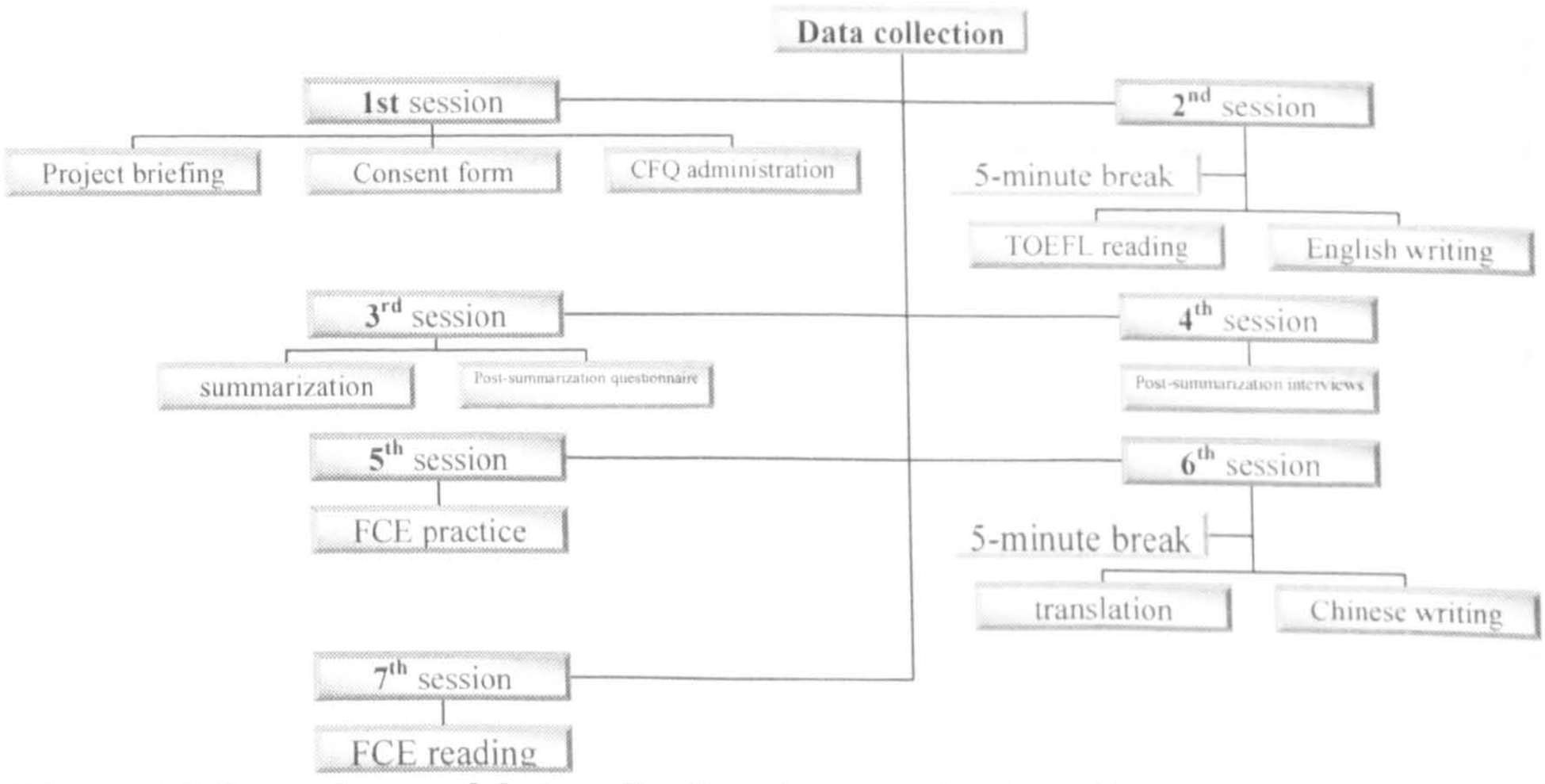


Figure 4.3 Procedures of data collection from student participants

1st session Project briefing, consent form, and CFQ administration.

2nd session All participants took the TOEFL reading test (65 minutes) and then undertook the English writing task (60 minutes). There was a five-minute quiet notional break.

3rd session Two days after the 2nd session, the participants undertook the 3-hour summarization tasks according to the pre-defined allocation of experimental conditions (Table 4.6). About half of the participants (n=82) did the summarization tasks in the computer room and the other half (n=75) in ordinary classrooms. Immediately after the summarization (see Appendix 3 for task directions), the students were asked to fill in the PSQ without a time limit.

TEXT A (Edu. History)				**TEXT B (River)**				**TEXT C (Work-Life)**			
Class 32		Class 41		Class 42		Class 31		Class 34		Class 33	
COMPUTER		PRINT		COMPUTER		PRINT		COMPUTER		PRINT	
1/2 E/C	1/2 C/E	1/2 E/C	1/2 C/E	1/2 E/C	1/2 C/E	1/2 E/C	1/2 C/E	1/2 E/C	1/2 C/E	1/2 E/C	1/2 C/E
13	13	12	15	16+2*	12	15	12	14	12	10	11
2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews	2 for interviews

Note: these two students were actually from Class 32 and were assigned to text13 due to difficulties arranging the seats in the computer room.
E/C=English then Chinese; C/E=Chinese then English

Table 4.6 Conditions of the summarization tasks

4th session Twenty-four students randomly selected were interviewed individually in Chinese (2 from each summarization condition, see Table 4.6). Some interviews were held before the 5th session; others after the 5th and the 6th sessions. All interviews were completed before the 7th session.

5th session The second day after the 3rd session, all students were given the FCE practice paper (reading paper only) and the UCLES test-taking recommendations. They were asked to familiarise themselves with the reading test formats.

6th session In the 6th session (5 days after the 3rd), all students undertook the translation task (70 minutes) and the Chinese writing task (45 minutes). There was a five-minute quiet notional break.

7th session In the final session, all students sat the FCE reading test within 75 minutes.

4.2.4 Evaluating the quality of students' summaries

1) Key quality indicators

Several *independent* and *integrated* quality indicators are referenced to evaluate students' summaries (see **2.5.1**). Summaries of textA and textC were assigned, according to both the *expert* and the *popular* template, the following four scores: (*a*) right statement credit (RSC), (*b*) wrong statement penalty (WSP), and (*c*) summary and source text relationship score (SSS), (*d*) 5% bonus/penalty on its (lack of) conciseness. Based on the RSC, WSP, SSS scores and the 5% bonus/penalty, these summaries were then assigned a holistic score (HS) for overall quality (see Appendix 12 and 13). Summaries of textB were assigned HS only according to the expert templates (see Table 4.8). In addition to these five scores, the lexical density of a student's summary (D-SS) was measured by VOCD.

a) Independent quality indicators: RSC, WSP, SSS, and 5%

To generate the RSC scoring templates, I used winMAX (see Appendix 5) to code all the English summaries produced by the five experts and the 157 students. The guidelines for coding the summaries were inspired by, but not confined to, Kroll's conceptualisation of idea units (1977, p. 90, cited in Johns and Mayes, 1990). This project used "statement" instead of idea unit to signify a chunk of information which can contain complete clause(s) and/or sentence(s). Therefore, in terms of the amount of information, a statement is bigger than an idea unit. A coded statement can include several idea units which convey fundamentally the same general concept within complete clause(s) or/and sentence(s). I first read all the English summaries written by the students in order to get an overview of the content of the students' summaries. Then the summaries were Word processed and imported into winMAX. In the winMAX programme, I read the summaries iteratively and assigned codes to the statements of the summaries according to the guidelines described above. All sentences were coded. Each code was attached a memo which described briefly the meaning of the code. Memos were also occasionally attached to some coded statements to record and remind myself why a particular code was assigned to a statement. Coding the English summaries into statements was an iterative, accumulative and laborious process.

An assistant then read all the printed "codes", "memos" of the "codes" and "coded segments" of each English summary, pointing out any disagreement with my codings. The differences were then resolved through email and telephone communications. This linear rather than independent coding process was employed to avoid the potential problems of very low agreement between two independent coders, as evidenced in Cohen's (1993; 1994) and Sawaki's (2003) studies.

The first 10 most frequently occurring statements were then used to generate the popular RSC scoring templates (Appendix 12). These statements were later translated into Chinese to create the corresponding templates for evaluating Chinese summaries of the same source texts. In exactly the same manner, the expert RSC scoring templates were created from the summaries written by the five English native speakers.

An unweighted partial credit system was used to assign RSC, without taking into consideration the level of importance of the statements:

2 points for a complete inclusion of one of the 10 statements
1 point for inclusion but of an incomplete statement
0 point for exclusion of a statement
Note: recurring statements were credited **ONLY ONCE**

Therefore, there was a maximum raw score of 20 for RSC (minimum 0). See the rating procedures in 2) below.

If all that is looked at is whether a key statement is included or not, some other important features reflecting the quality of a summary may well be lost. For example, one summary might include more wrong statements than another; one summary might have more direct copying from the source than another; one summary might be more concise than another. These three possible situations were taken into account by assigning WSP and SSS scores, plus or minus 5%, to a summary.

An *obviously wrong* statement was penalized as a wrong additional unit (1 point for each wrong statement). Some reasonable inferences were acceptable, and not penalized. The minimum score for WSP was 0 and there was no maximum score.

The relationship between summary and source text (SSS) was assigned

according to a 3-point scheme developed from Stein and Kirby's (1992: 224) scale on the depth of the summarization process:

3 A score of 3 would be assigned to summaries that were predominantly written in the summarizer's own words, rearranged the order of the statements in a logical manner, had examples of integration and connectives, and had global interpretation of the source text.

2 A score of 2 would be assigned to summaries that had some indication of the summarizer's own words being used, re-ordering of the statements (though still linear in their presentation), and an attempt to integrate or use connectives.

1 A score of 1 would be assigned to summaries that were predominantly copied verbatim, followed the original order of the statements in the source text, showed no instance of integration and connectives, and were not global in their interpretation of the source text.

A 5% score (5%+, 5%0, or 5%-) was used to acknowledge and reflect differences in the succinctness of summaries.

5%+ Those with clear evidence of a succinctly written summary, and with a low percentage of non-important statements, judged holistically by the raters, received a 5% bonus.

5%0 Those with no clear evidence of the quality of a succinctly written summary, and with several instances of non-important statements, judged holistically by the raters, did not receive the bonus (in a sense it is also a penalty);

5%- Those with a clear absence of the quality of a succinctly written summary, and with a higher percentage of non-important statements than important statements, judged holistically by the raters, received a 5% penalty.

I have been using 5% scores to signify the succinctness of summaries in my own language testing practice. Similar practices can also be found in the literature on summarization studies, for example, Bensoussan and Kreindler (1990: 59). However, the process in this study differs slightly from theirs (see **2.5.1**). The 5% score used in this project was based on the scores students had already earned, instead of a full 5% score. In other words, student having earned a grade of 60%, and also having written a succinct summary would be awarded a final grade of 63% [i.e. (60*5%+60) / %], not 65%. The raters only needed to indicate whether a 5% score should be added or not (5%+, 5%0, 5%-).

b) Integrated overall quality of student summaries: HS

Based on its RSC, WSP, SSS and 5%, each summary was also assigned HS to reflect its overall quality, using an augmentation method. The scale (Appendix 13) focused on the following four salient features of a summary:

- faithfulness to the source text, in terms of the percentages of the right and wrong statements in the summary,
- the topological relationships between the summary and the source text, with emphasis on the use of the summarizer's own language and language from the source text, the integration and connections of the statements,
- the conciseness, coherence and logicality of communication of meaning,
- the overall difficulty that the raters might encounter.

2) Evaluating the quality of summaries

a) Raters and rating procedures

Three postgraduates registered in the MEd TESOL at the University of Bristol marked the summaries, as well as other written products (English and Chinese writings, translation). In all rating tasks, each rater was randomly assigned ⅔ of the summaries within one class. Table 4.7 illustrates this overlapping design for the six summaries of Class31. Therefore, all raters were connected and all summaries were double marked.

examples	Rater 1	Rater 2	Rater 3
Summary 3101	√	√	
Summary 3102		√	√
Summary 3103	√		√
Summary 3104	√	√	
Summary 3105		√	√
Summary 3106	√		√

Table 4.7 Rater assignments for rating the summaries

If the number of students in one class was not divisible by 3, the remaining 1 or 2 summaries were again randomly assigned to one of the three conditions (Condition 1 = Raters 1 and 2, Condition 2 = Raters 2 and 3, Condition 3 = Raters 3 and 1).

The order of the four scoring templates was randomly assigned, as was the order of source texts and summaries. Summaries of the first two randomly selected texts were judged against four scoring templates and assigned subjectively RSC, WSP, SSS, 5%, and HS scores, while those of the third text were assigned only HS according to the expert templates (Table 4.8). As shown in the Table, all summaries of textC and textA were assigned scores for RSC and HS, but SSS was only assigned to the first English summaries (be it against the expert or popular scoring template). There was

no SSS score for Chinese summaries for the obvious reason that all Chinese summaries were written in the summarizers' own words (see SSS guidelines above). Scores for SSS and WSP were only assigned to the first Chinese or English summaries (be it against the expert or popular scoring template). However, the scores for WSP or SSS were considered the same, whether a summary was judged against an expert or a popular scoring template, because what was wrong in one scoring template was also considered wrong in another scoring template. Therefore, they were shared across different scoring templates.

Text Order	Order of the Scoring Templates	Scores Assigned	Scores Measured
Work-life Balance (TextC)	Expert English (EE)	RSC, WSP, 5%, SSS, HS	D-SS (vocabulary density of English summaries)
	Expert Chinese (EC)	RSC, WSP, 5%, HS	
	Popular English (PE)	RSC, 5%, HS	
	Popular Chinese (PC)	RSC, 5%, HS	
Educational History (TextA)	Popular English (PE)	RSC, WSP, 5%, SSS, HS	
	Expert Chinese (EC)	RSC, WSP, 5%, HS	
	Popular Chinese (PC)	RSC, 5%, HS	
	Expert English (EE)	RSC, 5%, HS	
Let the River Run (TextB)	Expert English (EE)	HS	
	Expert Chinese (EC)	HS	

Table 4.8 Rating procedures and scores assigned

The raters were not told the experimental conditions in which the summaries were produced, students' performances in other tasks, whether they were using the expert or popular templates. All written scripts were anonymised and Word-processed before being marked. No attempt was made to correct spelling, punctuation, or grammatical mistakes in the originals. Special care was also taken to randomise the order of the summaries that the raters received within each cohort. In this way it was hoped that the "halo effect", "test-to-test carryover effects", and "order effects" (see Hopkins 1998: 191-192) could be minimized.

The raters received brief training in moderation sessions. Five summaries in each scoring sequence were marked and moderated by all the raters and myself (see Table 4.8); differences among us were discussed so that we could have better and similar understandings of the assessment criteria (Appendices 12 & 13). The same moderation procedures were also applied to the marking of the translation and English and Chinese writings.

b) Negotiating differences

It is inevitable that there will be differences between the scores assigned to the same script. In order to best reflect the quality of students' summaries, English and Chinese writings and translation, several principles were established such that the inevitable differences could be negotiated:

i) HS and RSC

HS was assigned through an augmentation method (see Appendix 13). The final HS score for a summary was the average of two numerical scores independently assigned if the difference between them was ≤ 3. When the difference between the two numerical scores was > 3, a third rater re-marked the summary in question, without knowing the previous two scores.

- In the case that the third score was the average of the first two scores, the third score was then reported. For example, if the first two scores were 12 and 16, and the third score 14, then the final score would be 14.
- In the rare case that the difference between any two of the three scores was still greater than 3, the three raters negotiated face to face to assign a proper score for the summary in question. For example, if the first two scores were 8 and 16, the third score might be 12.

The differences between the first two independent ratings of RSC for summaries, holistic scores for English and Chinese writings and translation, were all negotiated and resolved in this way.

ii) SSS, 5%, and WSP

Procedures to negotiate differences in SSS, 5%, and WSP were slightly different from the above principles. SSS could be either 1, 2 or 3. If the difference was 1, then the average of the two independent SSS was reported as the final score; if the difference was 2 (i.e. when $SSS_1=1$, $SSS_2=3$), then the summary in question was re-marked by a third rater. After re-marking, there were three possibilities of SSS combinations, and the final SSS was reported as follows:

SSS_1=1, SSS_2=3, SSS_3=1 → SSS=1
SSS_1=1, SSS_2=3, SSS_3=2 → SSS=2
SSS_1=1, SSS_2=3, SSS_3=3 → SSS=3

A 5% score can be either +, 0, or –, and therefore there were six possible score combinations in the first round of rating. The procedures to report the final 5% score were as follows:

+ and + → 5% score = +
+ and 0 → 5% score = ½
0 and 0 → 5% score = 0
0 and – → 5% score = - ½
– and – → 5% score = –
+ and – → re-marked by a third rater (see below).

When a third rater re-marked the summary in question, there would be three possible score combinations, and the final 5% score was reported as follows:

+ and – and 0 → 5% score = 0, ½ , or - ½ (negotiated by the raters)
+ and – and + → 5% score = +
+ and – and – → 5% score = –

The final WSP score was reported as the average of the first two WSPs when the difference was equal to or less than two (≤ 2). If the difference was > 2, a third rater assigned a new WSP for the summary in question. Then, the average of the two adjacent WSPs was reported as the final WSP in a similar way as that for negotiating the difference for HS.

4.3 Summary

This chapter has made transparent the research questions and hypotheses and research design of the study which investigates in an organic manner the four components (*input*, *filter plant*, *output* and *evaluation*) of the IFOE framework (see Figure 3.1). Over 150 Chinese university undergraduates and 5 English native speaker experts participated in the project. The experts were asked to produce summaries of three extended texts of about 2200 words each and their summaries were used to generate the expert scoring templates. The students were asked to summarize, in both their first language (Chinese) and a foreign language (English), one of the three English texts, either computer mediated or paper presented. The summarization tasks for the students were in a factorial design of 3 text types x 2

text presentation modes (*computer* vs. *print*) x 2 language orders (*English then Chinese* vs. *Chinese then English*), and their written summaries were subjected to both expert and popular assessment criteria. The key quality indicators of a student's summary included its RSC, WSP, SSS, 5% and HS, plus D-SS. The five research questions addressed the effects on students' summarization performances of (i) text type, (ii) text presentation mode (computer vs. paper) and computer familiarity, (iii) EFL reading, writing (English and Chinese) and translation (from English to Chinese) abilities, (iv) languages used for writing the summaries (Chinese and English), and (v) assessment criteria (expert vs. popular template). The impacts of these factors on students' summarization performances were analysed from two parallel datasets: (i) students' *actual* performances in the summarization tasks and other measurements such as computer familiarity, reading, writing and translation abilities, and (ii) their *perceptions* of the summarization tasks through post-summarization questionnaire and interviews.

In the following Part IV, I will first report the basic statistics at a micro-level (Chapter 5) and then the analyses and findings relating to the five key research questions in Chapters 6 to 10.

PART IV

CHAPTER FIVE

Basic Statistics at Micro-Level

The data were analysed at micro- and macro-levels. This chapter reports the basic statistics at micro-level of (i) computer familiarity, (ii) EFL reading, (iii) English and Chinese writing, (iv) translation, and (v) summarization performances (WSP, 5%, RSC, SSS, HS, D-SS and Lengths). The macro-level analyses in specific relation to the five research questions are reported in Chapters 6 to 10.

5.1 Computer familiarity

5.1.1 Assessing the factorability of the CFQ data

The main purpose of factor analysing the CFQ data was to establish a more reliable computer familiarity scale for better discrimination. The factorability of the CFQ data (see Appendix 14 for descriptive statistics) was assessed through examining the sample size, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser 1970, 1974) and Bartlett's test of sphericity (Bartlett 1954), and anti-image and Pearson correlations matrices.

Though only about half of the students (n=82) summarized the computer-mediated texts, data from all the students (N=157) were used in the factor analyses, based on the understanding that:

- the sample size would be too small for statistical faithfulness in the results of factor analyses if only data from 82 students were used[1]; and

[1] Taking into account the number of variables to be used in the factor analyses (32), a sample size of around 170 was needed to give statistical faithfulness in the findings of the factor analyses (Baggaley 1982; Cattell 1952; Child 1990). Baggaley came up with a very useful ratio of the number of participants to variables, by which the sample size (N) can be estimated if the number of variables (p) is known and a rough estimate of the average correlation (Q) between all the variables can be made. In this research, to play safe, it was assumed that the average correlation coefficient between the 32 variables was 0.10. According to Baggaley's ratio, N/p = 5.78 (p=32 was not in Baggaley's table: the closest is 30). Given the value of p (32), the sample size (N) therefore needs to be around 170.

- it was assumed that the two groups of students (computer vs. print) had similar computer familiarity because they were randomly assigned to these two experimental conditions.

Inspections of the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (.809) which was above .6 – the minimum value for a good factor analysis (Tabachnick & Fidell 2003) and Bartlett's test of sphericity ($p < .0005$) indicated the appropriateness of factor analysing the data.

The anti-image correlation matrices were also examined to see whether there were variables that did not seem to fit with the structure of the other variables. Q5 (.383) and Q30 (.468) were found to have MSA (measures of sampling adequacy) below .50, and were therefore dropped in further factor analyses. Finally, the Pearson correlation matrix for the remaining 30 questions showed that a significant number of correlations were ≥.3, meeting the recommendation by Tabachnick and Fidell (2003). The average correlation was .2253.

5.1.2 Factor analysing the CFQ data: methods and results

The CFQ data were subjected to exploratory principal component analyses. Nine components had eigenvalues > 1 (see Appendix 15). Applications of Cattell's scree test (1952) and Kaiser's criterion are among the most popular for factor analysts to determine the number of factors to be extracted. Cattell suggested that "Kaiser's criterion is probably most reliable when the number of variables is between 20 and 50" (Child 1990: 38), and this method is "particularly suitable for principal components designs" (*ibid*: 37).

Another critical decision to make is the selection of rotation method for subsequent solutions. The debates between American and British psychologists on whether to use orthogonal or oblique rotations suggest both are laudable on different grounds.

> Many have argued that correlated factors are much more reasonable to assume in most cases..., and therefore oblique rotations are quite reasonable. (Stevens 2002: 392)

Pedhazur and Schmelkin (1991:615, cited in Stevens 2002: 392) also argued that:

> From the perspective of construct validation, the decision whether to rotate factors orthogonally or obliquely reflects one's conception regarding the structure of the construct under consideration. It boils down to the question: Are aspects of a postulated multidimensional construct intercorrelated? The answer to this question is relegated to the status of an assumption when an orthogonal rotation is employed... The preferred course of action is, in our opinion, to rotate both orthogonally and obliquely. When, on the basis of the latter, it is concluded that the correlation among the factors are negligible, the interpretation of the simpler orthogonal solution becomes tenable.

In this project, oblique rotation was favoured, because it was assumed that the behavioural characteristics of computer use were so interrelated that this should be taken into consideration in the selection of rotation methods (see Child 1990: 52). However, following Pedhazur and Schmelkin's advice, I used both orthogonal and oblique rotation methods in a series of extractions (see Appendix 15).

5.1.3 Creating the computer familiarity scale

Fifteen variables/questions having loadings ≥.3 in the first factor of the two-factor promax oblique rotation (Kappa= 4) were used to define the scale of computer familiarity. The sum of a student's answers to the 15 questions – Q1, Q2, Q3, Q4, Q8, Q15-19, Q21-25 was his/her computer familiarity score[2]. In the case of a missing value in any of the 15 questions, a minimum score of 1 was assigned for that question. The theoretical maximum of computer familiarity was therefore 60. The intraclass correlation coefficient indicated a respectably high reliability for the computer familiarity scale (alpha=0.8425, standardized item alpha=0.8658, see Appendix 15).

5.1.4 Students' computer familiarity

Generally speaking, these students' computer familiarity was moderate to high (mean=39.7771, s.d.=6.92598, min=21, max=53). The final open-ended question (Q33) provided further evidence for this level of computer familiarity. All students had had basic training in the university's compulsory courses, and had passed either national or provincial examinations on both the theoretical knowledge and practical use of computers. All the participants, except ID3419 who had a computer familiarity score of 37, had positive attitudes towards computers.

---

[2] The use of the variables/questions having loadings ≥.3 of the first factor in the two-factor varimax rotations (see Appendix 14) did not make much difference in students' computer familiarity scores, therefore only the results from the promax rotations were reported and used to create the scale.

As anticipated, no significant difference in computer familiarity was found between students of different summarization conditions (*text presentation mode*, *text type* and *language order*, see Table 4.6), which means that the experimental conditions did not artificially create differences among the groups.

The computer familiarity scale was able to discriminate students from different Years. As anticipated, Year4 students (n=55, mean=43.46, std. deviation=5.86) were significantly more familiar with computers than Year3 students (n=102, mean=37.79, std. deviation=6.67; t=-5.291, df =155. sig.<.0005). The magnitude of the difference was large (η^2=0.153) in Cohen's terms (1988). Although a significant difference was also noted among the six classes ($F_{5,151}$=8.38, sig. <0.0005), it was mainly attributable to the significant difference between class32 and class41/class42, and between class33 and class41/class42 (Figure 5.1). The magnitude of difference was also large (η^2=0.2172).

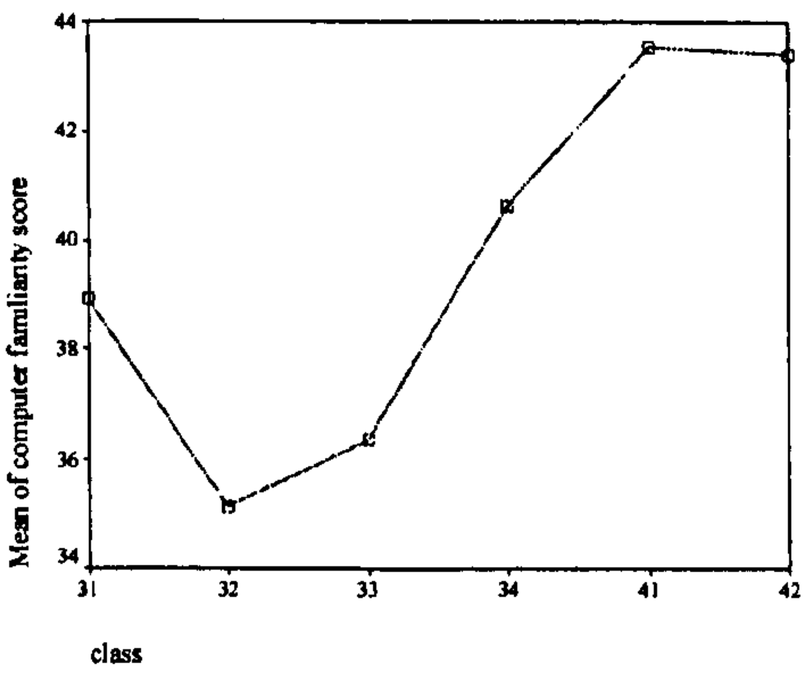


Figure 5.1 Means plot of computer familiarity of six classes

These results illustrate with some confidence that the scale is able to discriminate students' computer familiarity, but is not contaminated (un)favourably by the artificially created summarization conditions.

5.2 TOEFL-R and FCE-R

TOEFL-R and FCE-R were used to measure students' reading comprehension abilities. Some basic statistics are provided in Table 5.1.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TOEFL Reading	156	24.00	48 00	36.2949	4.76034
FCE Reading	154	17.00	48 00	35.1753	6 08989

Table 5.1 Descriptive statistics of students' performances in TOEFL-R and FCE-R

The correlation between TOEFL-R and FCE-R performances was moderate (r=0.40, sig.<0.0005), which reflected to some extent the difference between the two tests and strongly supported the use of *two* tests in the current project (see **4.2.2**).

As expected, no significant difference in TOEFL-R or FCE-R was found between different summarization conditions (*text presentation mode*, *language order*). However, there were significant differences in TOEFL-R between *text type* groups ($F_{2,153}$=5.842, sig.<0.0045, η^2=0.071), but not in FCE-R ($F_{2,151}$=0.958, n.s.). This significant difference in TOEFL-R was attributable to the difference between students on textA and textB (sig.<0.0255), and between textA and textC (sig.<0.0095). In fact, this difference was partly anticipated, because significant differences in reading abilities between different Years and between different classes are assumed (see below), leading to the strong possibility of difference between *text* groups because the classes students were in partly determined the text they summarized (see Table 4.6).

The independent samples *t*-tests of TOEFL-R and FCE-R scores demonstrated significant differences between Year3 and Year4 (TOEFL: t=-5.472, df=154, sig.<0.0005, mean of Year3=34.90, mean of Year4=38.93; FCE: t=-3.964, df=152, sig.<0.0005, mean of Year3=33.81, mean of Year4=37.70).

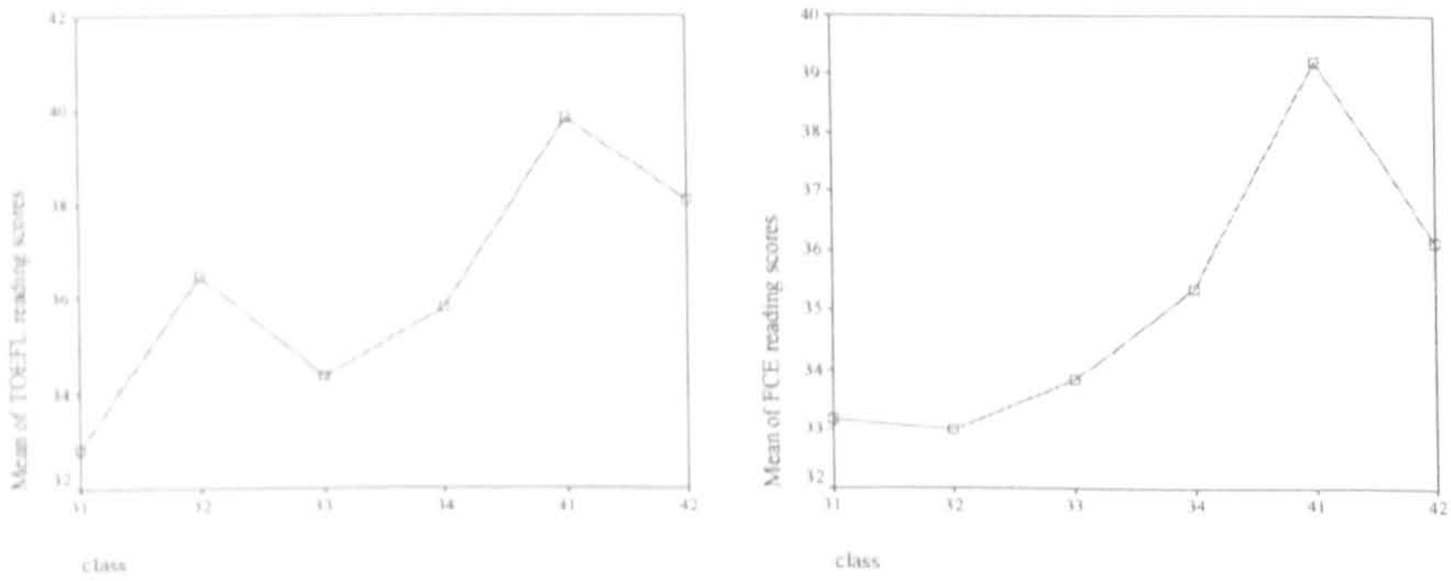


Figure 5.2 Means plots of TOEFL-R and FCE-R scores by six classes

One-way ANOVA also indicated that there was significant difference among the six classes ($F_{5,150}$=9.158, sig.<0.0005, η^2=0.2339 for TOEFL; $F_{5,148}$=4.475, sig.<0.0015, η^2=0.1313 for FCE), however this significant difference was mainly attributable to the substantial differences between the classes of different Years (see Figure 5.2).

5.3 English and Chinese writings

5.3.1 English writing

The number of English essays on which there was a score difference greater than three between the first two raters was nine between Rater1 and Rater2 (17.31%, n=52), six between Rater2 and Rater3 (11.54%, n=52), and five between Rater3 and Rater1 (9.62%, n=52). In total, 12.82% of the 156 English essays were re-marked in the second round. After the second round of rating, no score difference was greater than 3.

The inter-rater reliability indices (standardized item alpha[3]) were 0.8531 between Rater1 and Rater2 (n=48), 0.8316 between Rater2 and Rater3 (n=51), and 0.8970 between Rater1 and Rater3 (n=57)[4]. The average inter-rater reliability was 0.8606, acceptably high in language testing research and practice. However, Rater1 was consistently more lenient than Rater3 and Rater2.

The descriptive statistics for students' English writing performance are shown in Figure 5.3 below.

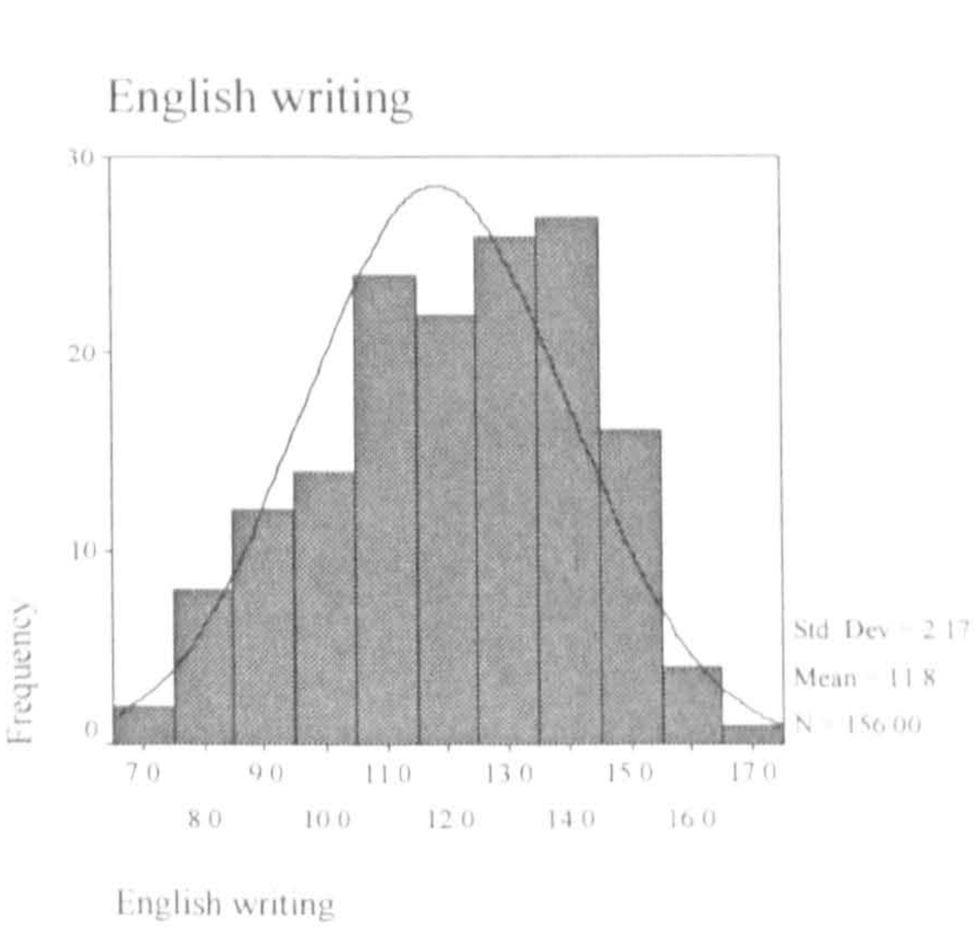


Figure 5.3 English writing performances

[3] This research took a cautious approach in investigating not only the standardized item Alpha, but also the average measure intraclass correlation (ICC), using a two-way random effect model with absolute agreement definition. ICC is slightly lower than the standardized item Alpha, because ICC reduces the effects of the agreements between the raters merely by chance. It was found that differences between standardized item Alpha and average measure ICC were negligible in this research and therefore only the standardized item Alpha is reported.

[4] Please note the number for each pair of raters differs from that in the first round of rating. This is because in some cases that the third score was reported as the average/final score in the second round of rating, i.e. when the third score was equidistant from the first two scores, the new pair of raters was then made up of the third rater with either of the first two raters, the latter selected randomly.

A series of ANOVA or *t*-tests, where appropriate, indicated that there was no significant difference in students' English writing abilities under different summarization conditions (*text type, text presentation mode, summarization language order*). Interestingly, the English writing task did not seem to be able to discriminate between students of different Year or Class groups.

5.3.2 Chinese writing

The number of Chinese essays on which there was a score difference greater than three between the first two raters was two between Rater1 and Rater2 (4%, n=50), eleven between Rater2 and Rater3 (21.57%, n=51), and nine between Rater3 and Rater1 (17.31%, n=52). On average, 14.38% of the 153 Chinese essays were re-marked. After the second round of rating, there was one essay on which the difference between the three scores were still greater than three (ID3408 rated by Rater2 and Rater3 in the first round). The raters then negotiated and agreed on a fourth score.

The inter-rater reliability indices were 0.8723 between Rater1 and Rater2 (n=61), 0.6178 between Rater2 and Rater3 (n=43), and 0.8014 between Rater1 and Rater3 (n=48). The average inter-rater reliability was 0.7637.

The descriptive statistics for students' Chinese writing performance are illustrated in Figure 5.4 below.

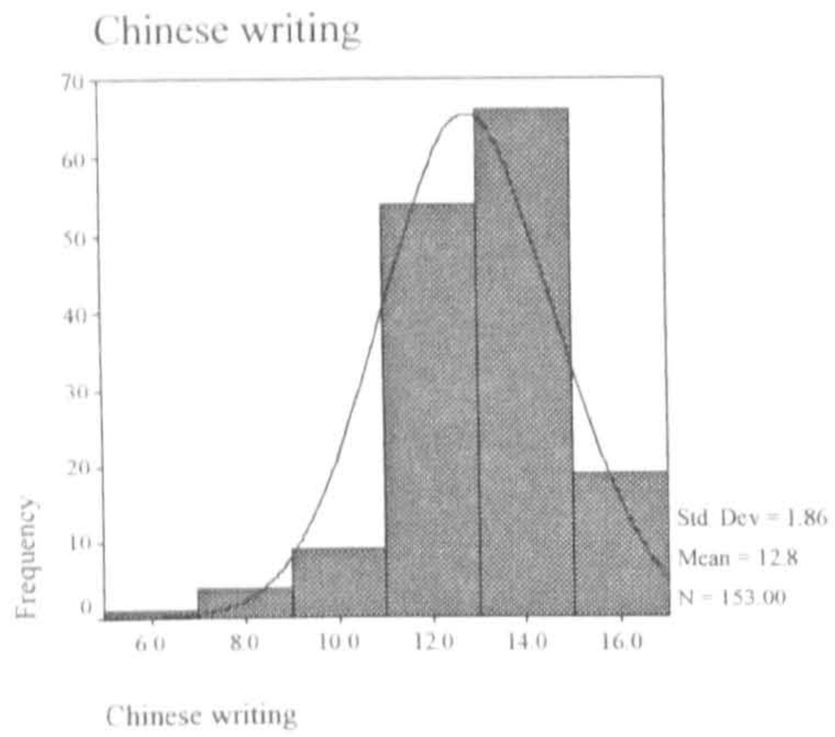


Figure 5.4 Chinese writing performances

A series of ANOVA or *t*-tests, where appropriate, indicated there were no significant means differences in the Chinese writing abilities of the students under different summarization conditions (*text type, text presentation mode, summarization language order*), nor in different Year or Class groups.

5.4 Translation

The number of translation scripts on which there was a score difference greater than three between the first two raters was four between Rater1 and Rater2 (7.84%, n=51), seven between Rater2 and Rater3 (13.21%, n=53), and ten between Rater3 and Rater1 (20%, n=50). After the second round of rating, no score difference was greater than 3.

Inter-rater reliability was 0.8732 between Rater1 and Rater2 (n=58), 0.8684 between Rater2 and Rater3 (n=52), and 0.8893 between Rater1 and Rater3 (n=44). The average inter-rater reliability was 0.877.

The descriptive statistics for the students' performances are shown in Figure 5.5 below.

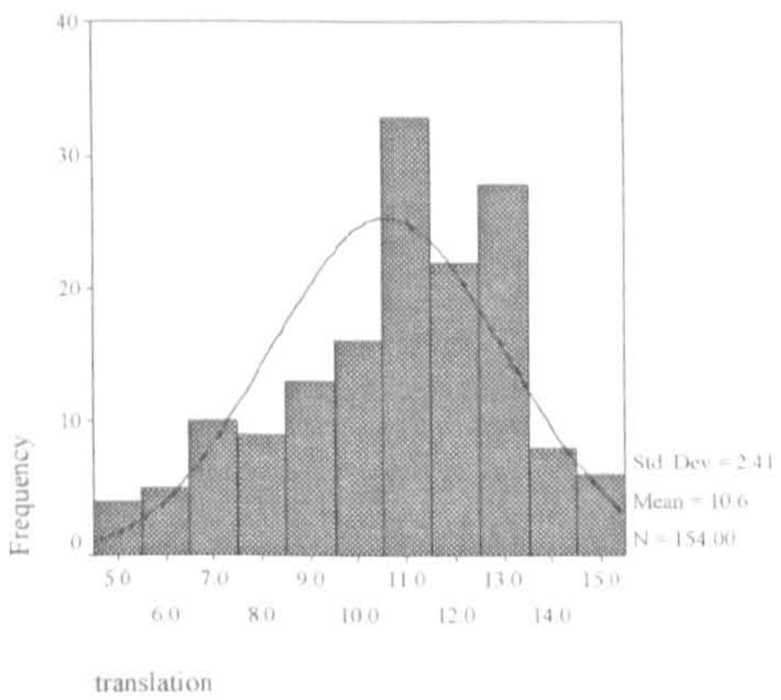
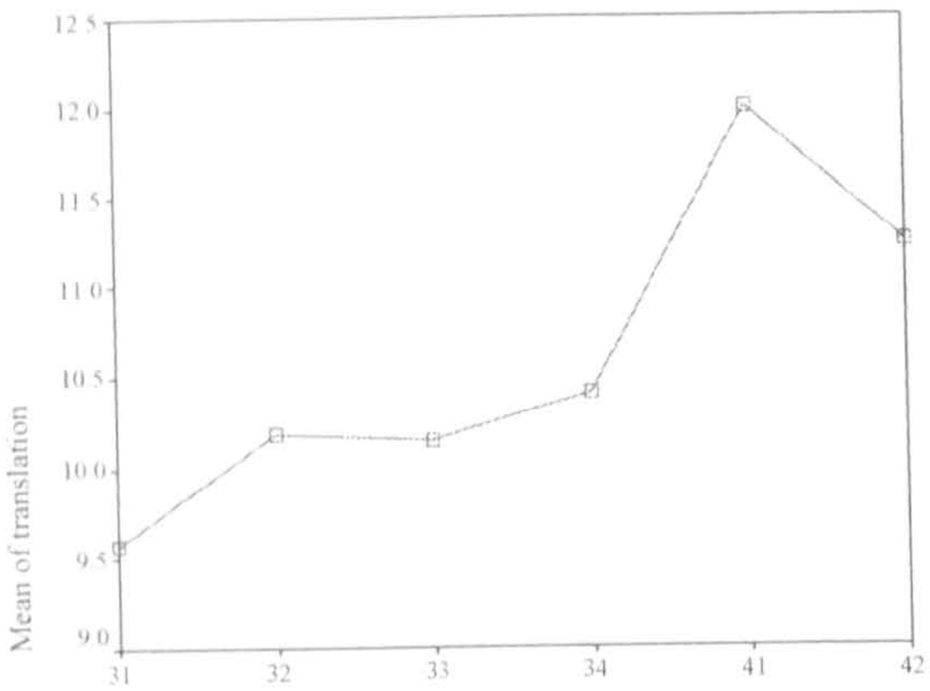


Figure 5.5 Translation performances

A series of ANOVA or *t*-tests, where appropriate, indicated there were no significant means differences in the translation abilities of students under different summarization conditions (*text type*, *text presentation mode*, *summarization language order*). There were significant differences in the translation abilities between students of different Years (t=-4.019, df=152, sig.<.0005, η^2=.0961) and Class ($F_{5, 148}$=3.864, sig.<.0035, η^2=.1155). The post-hoc Scheffe test indicated that the significant difference between Classes was mainly attributable to the difference between Class31 and Class41. Differences between other classes were not significant (Figure 5.6).



translation

Scheffe[a,b]

class	N	Subset for alpha = .05	
		1	2
31	27	9.5741	
33	21	10.1429	10.1429
32	28	10.1786	10.1786
34	25	10.4000	10.4000
42	26	11.2500	11.2500
41	27		12.0000
Sig.		.249	.150

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 25.438.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Figure 5.6 Means plot and subset of translation scores by Class

In summary, the experimental designs did not create (un)favourably artificial significant differences in the students' computer familiarity, TOEFL-R, FCE-R, English and Chinese writing, or translation between different summarization conditions. The reliability indices of these measurements were high.

5.5 Summarization performances

5.5.1 Inter-rater reliability of RSC and HS scores

Inter-rater reliability for RSC and HS was analysed from two perspectives: a) *by three texts as a whole group* and b) *by individual text*. It was found that inter-rater reliability using intraclass correlation coefficients was respectably high in almost all cases (see Appendix 16). It was also noted that:

- the mean inter-rater reliability for RSC was slightly higher than that for HS ($mean_{RSC}$=0.9087, $mean_{HS}$=0.8805); This may be due to the fact that when assigning RSC, raters had more detailed and tangible guidelines to refer to and these were therefore probably more helpful than the guidelines for assigning HS (see Appendices 12 & 13);
- inter-rater reliability was at the same level for both Chinese and English summaries, and for both *expert* and *popular* templates, and for all the three source texts;
- the highest inter-rater reliability Alpha reached 0.9815; however, there were three cases where inter-rater reliability was below 0.80 (CPRSC, 0.7112; CPHS, 0.6695; EPHS, 0.7849, see the highlighted area in Appendix 16).

Generally speaking, the respectably high inter-rater reliability attained in this project disproved the claim in the literature that marking written summaries is extremely difficult and high inter-rater reliability is almost impossible to achieve.

5.5.2 Brief overview of students' summarization performances

1) WSP

As reported in the frequencies of WSP scores (-2.0, -1.5, -1.0, -0.5, 0) for the summaries of textA and textC (Table 5.2), the vast majority of the summaries (70% for English, and 74% for Chinese) did not contain significantly wrong statements and therefore did not receive any penalty in this respect.

WSP	TextA				TextC				Total			
	English summary		Chinese summary		English summary		Chinese summary		English summary		Chinese summary	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%	Freq.	%
-2.0	-	-	-	-	1	2.1	-	-	1	1	-	-
-1.5	-	-	1	1.9	2	4.3	3	6.4	2	2	4	4
-1.0	2	3.8	-	-	3	6.4	4	8.5	5	5	4	4
-.5	9	17.0	10	18.9	13	27.7	8	17.0	22	22	18	18
.0	42	79.2	42	79.2	28	59.6	32	68.1	70	70	74	74
Total	53	100	53	100	47	100	47	100	100	100	100	100

Table 5.2 Frequency of WSP for textA and textC summaries

The Chinese and English summaries shared a similar pattern of WSP score distributions, as did the two texts. Only one English summary in the whole cohort was penalized for containing two significantly wrong statements.

2) 5%

A 5% score (5%+, 5%0, or 5%-) was used to acknowledge and reflect whether a summary was written succinctly (see **4.2.4**). It was found that there was no significant difference in 5% scores of summaries judged by the *expert* and the *popular* templates. However, English summaries of textA received more penalties (56.6% for EE, 35.8% EP) than their Chinese counterparts (26.4%, 18.9%, see Appendix 17). In other words, Chinese summaries were more likely to receive the bonus than English ones (see RSC adjustments below).

3) RSC

In the macro-level analyses (see Chapters 6 to 10), the RSC score for a summary

was adjusted according to its WSP and 5% scores (see Appendix 18), and reported in percentages (Table 5.3). Because only a small number of summaries were penalized for containing significantly wrong statements (Table 5.2), the WSP adjustments did not make much difference. The subsequent 5% adjustments made no significant difference in the RSC of English summaries, but did make a significant difference in the RSC of Chinese summaries (t=-3.167, df=99, sig.<.0025 for CERSC; t=-4.333, df=99, sig.<.0005 for CPRSC).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
EERSC	100	21.0	80.0	46.265	11.9611
EPRSC	100	23.5	81.5	52.940	11.2925
CERSC	100	7.5	79.5	42.605	14.8200
CPRSC	100	9 0	79.5	46.850	12.5621

Table 5.3 Descriptive statistics for RSC scores after WSP and 5% adjustments

4) SSS

SSS was assigned only to the English summaries of Texts A and C. As reported in Table 5.4, only a very small number of English summaries were considered *very much in the summarizer's own words and organization.* If an equal distribution of scores were expected, each score would be assumed to be 20%. Chi-square indicated a statistically significant difference from expectations for the aggregate scores of both texts (χ^2=27.2, df=4, sig.<0.0005). In the separate analyses for each text, the chi-square tests also indicated a statistically significant difference from expectations (χ^2=37.4, df=4, sig.<0.0005 for textA; χ^2=37, df=4, sig. <0.0005 for text C).

SSS	TextA		TextC		Total	
	Frequency	Percent (%)	Frequency	Percent (%)	Frequency	Percent (%)
1.0	15	28.3	9	19.1	24	24
1.5	13	24.5	19	40.4	32	32
2.0	19	35.8	9	19.1	28	28
2.5	3	5.7	9	19.1	12	12
3.0	3	5.7	1	2.1	4	4
Total	53	100	47	100	100	100

Table 5.4 Frequency of SSS for textA and textC summaries

5) HS

Summaries of textA and textC were assigned HS scores according to both the expert and the popular templates, while textB summaries were evaluated only

according to the expert templates. Table 5.5 below shows the basic statistics for HS.

Rating Criteria	HS			
	EE	EP	CE	CP
Mean	10.478	11.315	9.535	10.21
Standard deviation	2.0680	1.9894	2.3779	2.1998
Min. – Max.	6-15	6-16.5	3-16	2.5-15
K-S z^*	0.811	0.971	1.011	1.024
Sig.	0.5275	0.3025	0.2585	0.2455
No. cases	157	100	157	100

Note: * Kolmogorov-Smirnov test (2-tailed) of normal distribution.

Table 5.5 HS scores of students' summaries

6) Lengths

A word count was also conducted for each summary to allow further data analyses at macro-levels because it was assumed that longer summaries were likely to contain more information and therefore had better chance of including right statements (see Appendix 12) than short ones. Overall, the Chinese summaries were significantly longer than English ones (t=19.067, df=156, sig.<.0005), and there was greater variation among the Chinese summaries themselves (Table 5.6).

Summaries	English	Chinese
Mean	310.61	507.46
Standard deviation	75.898	149.896
Min. – Max.	147-624	238-1245
Kolmogorov-Smirnov Z^*	0.898	1.405
Sig.	0.3965	0.0395
No. of summaries	157	157

Note: * Kolmogorov-Smirnov test (2-tailed) of normal distribution.

Table 5.6 The lengths of English and Chinese summaries

As shown in Table 5.6, the lengths of the Chinese summaries were not normally distributed, which would violate some assumptions of normal distribution in the macro-level analyses. Removal of the univariate outliers improved the normality. In the subsequent macro-level data analyses, these outliers were therefore removed (e.g. see Table 10.1).

7) Vocabulary density

The vocabulary density of the English summaries was measured using VOCD in the CLAN programme. Table 5.7 reports the descriptive statistics and the comparisons in D of students' summaries and the source texts.

	D-SS (students' summaries)		
	textA	textB	textC
Mean	79.86	71.55	90.87
s.d.	11.41	10.81	13.48
Min-max.	50.61-106.25	52-94.14	66.56-122.87
Kolmogorov-Smirnov Z	0.475	0.423	1.016
Sig. <	0.9785	0.9945	0.2535
No. of summaries	53	57	47

Note: D was 84.76, 92.2, and 115.9 for textA, textB, and textC respectively.

Table 5.7 Vocabulary density of English summaries written by the students

The mean D of students' summaries was lower than that of source texts. However, only two experts (No. 2 & 5 in Table 5.8) produced summaries of lower D than the source texts, the other three wrote summaries that had substantially higher D than the source texts.

Expert ID	textA summaries	textB summaries	textC summaries
1	100.08	95.80	153.50
2	71.50	66.90	104.16
3	117.64	92.75	150.34
4	113.76	93.67	121.98
5	62.75	66.64	113.22

Table 5.8 Vocabulary density of English summaries written by the experts

5.6 Summary

This chapter reported (*a*) the basic statistics of the *filter plant* variables such as students' computer familiarity and reading, writing and translation abilities and (*b*) inter-rater reliability of RSC and HS scores. An overview of the students' summarization performance was also presented. In the next chapters (6-10), I will report the macro-level analyses and findings in the order of the five research questions (see **4.1**).

CHAPTER SIX

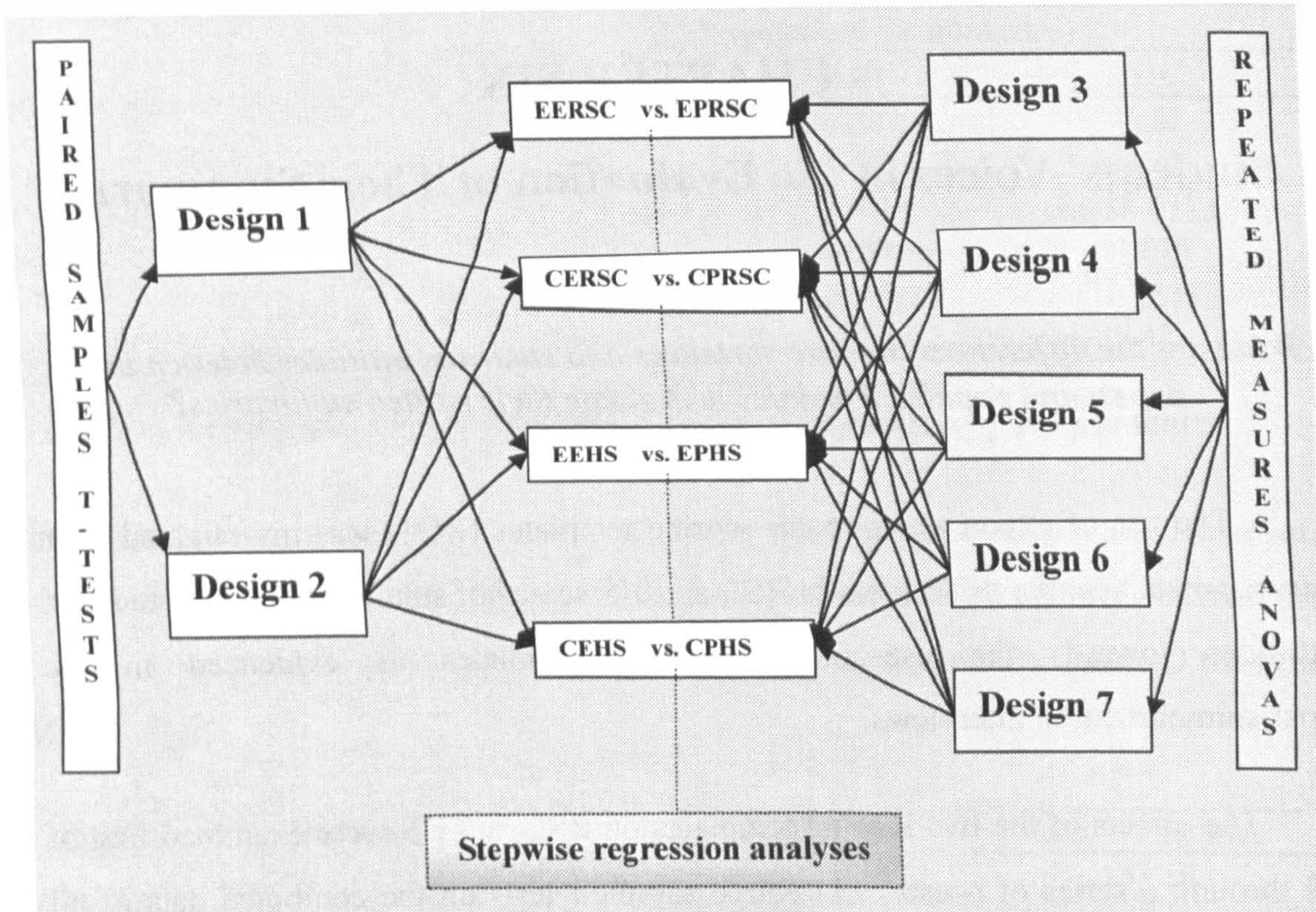
Students' Voices in the Evaluation of Their Summaries

What are the differences in score variances and students' attitudes between using expert and popular templates to evaluate their written summaries?

The use of expert and popular scoring templates (RQ1) was investigated from two perspectives: the differences in RSC and HS scores of summaries and in students' attitudes towards the use of these two templates as evidenced in the post-summarization interviews.

The effects of the two scoring templates on RSC and HS were examined first of all through a series of *t*-tests[1]: (1) paired sample *t*-tests on the combined data of all summaries, (2) paired sample *t*-tests on data from summaries of each individual text; and then in a series of ANOVA: (3) one-way repeated measures ANOVA by *text type*, (4) two-way repeated measures ANOVA by both *text type* and *level of TOEFL-R* scores, (5) one-way repeated measures ANOVA by *text type* and with TOEFL-R as a covariate, (6) one-way repeated measures ANOVA by *text type* and with *summary lengths* as a covariate, and (7) two-way repeated measures ANOVA by *text type* and *level of summary lengths*. Stepwise regression analyses were also conducted to understand which scoring template could better predict TOEFL-R and FCE-R. The procedures for the quantitative data analyses are presented in Figure 6.1. Students' attitudes towards the use of the two templates were analysed qualitatively (see **6.2**).

[1] For the sake of simplicity in presenting the results, *t*-tests can provide a brief overview of the effects of interest. However, they can not incorporate simultaneously the effects of other experimental factors such as *text type* and *presentation mode* and therefore can not present as complete and detailed a description of the phenomenon under investigation as multivariate statistics (Stevens 2002: 174-175). In this project, both *t*-tests and multivariate statistics were reported so that readers can have first of all a brief overview and then a detailed description of the effects of interest. See also **8.1.2**, **9.1.2**, and **10.1.2**.



Note:

- Designs 1 and 2 are paired samples *t*-tests; Designs 3 to 7 are repeated measures ANOVAs.
- Design 1 = paired samples *t*-tests on the combined data of all summaries (textA and textC)
- Design 2 = paired samples *t*-tests on the summaries of each individual text
- Design 3 = one-way repeated measures ANOVA by *text type*
- Design 4 = two-way repeated measures ANOVA by *text type* and Level of TOEFL Reading scores
- Design 5 = one-way repeated measures ANOVA by *text type* and with covariate of TOEFL Reading scores
- Design 6 = one-way repeated measures ANOVA by *text type* and with covariate of Summary Length
- Design 7 = two-way repeated measures ANOVA by *text type* and Level of Summary Length
- The stepwise regression analyses used TOEFL and FCE Reading scores as dependent variables and the four pairs of RSC and HS scores above as independents (see also Chapter 7 on the relationships between summarization performances and other language abilities).

Figure 6.1 Plan for the statistical analyses on the effects of scoring templates

6.1 Summarization performances

6.1.1 Expert and popular templates in multivariate ANOVAs

HS and RSC scores *before* and *after adjustments* (see Appendix 18) were subjected to the analyses[2]. Box's M statistics were checked, with no serious violations of the assumptions noted. It was found that scoring templates had significant main effects on both RSC and HS, in the absence of any interaction effects between the scoring templates and other factors in the models such as *text type* and levels of

[2] Although it was noticed that there were no significant differences between RSC before any adjustments and RSC after adjustments for the English summaries, there were significant differences for the Chinese summaries. Because this was the first macro-level data analysis, to be thorough it was decided that both RSCs would be analyzed (i.e., RSC *before* and *after* adjustments for WSP and 5%).

students' TOEFL-R and levels of the lengths of their summaries (high, medium, low). The effect sizes using partial η^2 were very large according to Cohen (1988). The results of these analyses are summarized in Table 6.1 (p.122).

The significant main effects of scoring templates held true for both the English and Chinese summaries (except in models No.5&6). When a summary was judged according to the popular templates, it had a significantly higher RSC and HS than when it was judged against the expert templates, regardless of *text type*, though *text type* had significant main effects on RSC and HS. TextA summaries received significantly higher RSC and HS scores than summaries of textC (see Chapter 10 *Effects of Text Type*). However, the main effects of scoring templates were no longer significant when the effects of TOEFL-R raw scores (No.5) and the lengths of summaries (No.6) as covariates were taken into account. Besides these key findings, it was also noted that:

- the effect size for RSC of the English summaries was almost twice as that for the Chinese summaries, except when the length of summaries was a covariate in the repeated measures ANOVA.
- there was not much difference in effect sizes between HS of the English and Chinese summaries except for textC summaries in Model 2 where the effect size on the English summaries (.1882) were much smaller than on the Chinese summaries (.2581).
- for the English summaries, the effect sizes on RSC were much bigger than those on HS, except for textA summaries in the separate paired samples t-tests (η^2=0.1789 for RSC and η^2=0.1969 for HS - a very small difference though);
- for the Chinese summaries, the effect sizes on RSC and HS were approximately at the same level, but with slightly bigger effect sizes for HS, except for (i) textA summaries where the effect size for RSC was slightly bigger than HS (η^2=0.1395 for RSC and η^2=0.112 for HS) and (ii) for textC summaries where the effect size for HS was almost twice as for RSC (η^2=0.1163 for RSC and η^2=0.2581 for HS).

	English summaries			Chinese summaries		
	RSC		HS	RSC		HS
	Before adjustments	After WSP and 5% adjustments		Before adjustments	After WSP and 5% adjustments	
1	$mean_{EERSC}$=9.455 $mean_{EPRSC}$=10.74 t=-5.618, df=99, sig.<0.0005 η^2=0.2417	$mean_{EERSC}$=46.265, $mean_{EPRSC}$=52.94 t=-5.544, df=99, sig.<0.0005 η^2=0.2369	EEHS<EPHS, t=-4 842, df=99, sig.<0.0005, η^2=0.1965 $mean_{EEHS}$=10.205, s.d.=2.0878, $mean_{EPHS}$=11.315, s.d.=1.9894, r=0.369, sig.<0.0005	$mean_{CERSC}$=8.625 $mean_{CPRSC}$=9.43 t=-3.719, df=99, sig.<0.0005, η^2=0.1226	$mean_{CERSC}$=42.605 $mean_{CPRSC}$=46.85 t=-3.753, df=99 sig.<0.0005, η^2=0.1246	CEHS<CPHS, t=-4.606, df=99, sig.<0.0005, η^2=0.1765 $mean_{CEHS}$=9.200, s.d.=2.5156, $mean_{CPHS}$=10.210, s.d.=2.1998, r=0.575, sig.<0.0005
2	**TextA**: EERSC<EPRSC, mean difference=-.991 t=-3.366, df=52, sig.<0.0015, η^2=0.1789 **TextC:** EERSC<EPRSC, mean difference=-1.638, t=-4.578, df=46, sig.<0.0005, η^2=0.3130	**TextA**: EERSC<EPRSC, mean difference=-5.217, t=-3.28, df=52, sig.<0.0025, η^2=0.1714 **TextC**: EERSC<EPRSC, mean difference=-8.319, t=-4.579, df=46, sig.<0.0005 η^2=0.3131	**TextA** t=-3.57, df=52, sig.<0.0015, η^2=0.1969 $mean_{EEHS}$=10.557, $mean_{EPHS}$=11.604 **TextC** t=-3.266, df=46, sig.<0.0025, η^2=0.1882 $mean_{EEHS}$=9.809, $mean_{EPHS}$=10.989	**TextA**: CERSC<CPRSC, mean difference=-.708 t=-2.903, df=52, sig.<0.0055, η^2=0.1395 **TextC:** CERSC<CPRSC, mean difference=-.915, t=-2.46, df=46, sig <0.0185, η^2=0.1163	**TextA**: CERSC<CPRSC, mean difference=-3.858, t=-2.89, df=52, sig.<0.0065, η^2=0.1384 **TextC**: CERSC<CPRSC, mean difference=-4.681, t=-2.474, df=46, sig.<0.0175, η^2=0.1174	**TextA** t=-2.561, df=52, sig.<0.0135, η^2=0.112 $mean_{CEHS}$=9.981, $mean_{CPHS}$=10.745. **TextC**: t=-4, df=46, sig.<0.0005, η^2=0.2581 $mean_{CEHS}$=8.319, $mean_{CPHS}$=9 606
3	Wilks' Lambda=0.750, F=32.721, sig.<0.0005, partial η^2=0.250	Wilks' Lambda=0.756, F=31.701, sig.<0.0005, partial η^2=0.244. No significant interaction effects	Wilks' Lambda=0.808, F=23.311, sig.<0.0005, partial η^2=0.192	Wilks' Lambda=0.876, F=13.883, sig.<0.0005, partial η^2=0.124	Wilks' Lambda=0.874, F=14 072, sig.<0.0005, partial η^2=0.126 No significant interaction effects	Wilks' Lambda=0.817, F=21.896, sig.<0 0005, partial η^2=0.183
4	Wilks' Lambda=0.762, F=29.428, sig.<0.0005, partial η^2=0.238	Wilks' Lambda=0.768 F=28.326, sig.<0.0005, partial η^2=0.232	Wilks' Lambda=0.828, F=19.516, sig.<0.0005, partial η^2=0.172	Wilks' Lambda=0.88, F=12.868, sig.<0 0015, partial η^2=0.12	Wilks' Lambda=0 88, F=12.799, sig.<0.0015, partial η^2=0.12	Wilks' Lambda=0.815, F=21.341, sig.<0.0005, partial η^2=0.185
5	n s.	n s.	n s.	n s.	n.s.	n s.
6	Wilks' Lambda=0.988, F=1.166 sig.<0.2835, **n.s.**, partial η^2=0 012	Wilks' Lambda=0.985, F=1.467, sig.<0.2295, **n.s.**, partial η^2=0.015	Wilks' Lambda=0.990, F=1.023, sig.<0.3145, **n.s.**, partial η^2=0.01	Wilks' Lambda=0.963, F=3.693, sig.<0 0585, **n.s.** partial η^2=0.037	Wilks' Lambda=0.955, F=4.623, sig.<0 0345, partial η^2=0.045	Wilks' Lambda=0.941, F=6.032, sig.<0.0165, partial η^2=0.059
7	Wilks' Lambda=0.751, F=31.243, sig.<0.0005, partial η^2=0.249	Wilks' Lambda=0.753, F=30.77, sig.<0 0005, partial η^2=0.247 No significant interaction effects, but level of summary length did have significant between-subjects main effects (F=7.868, sig.<0 0015, partial η^2=0 143)	Wilks' Lambda=0.802, F=23.199, sig.<0.0005, partial η^2=0.198	Wilks' Lambda=0 879, F=12.993, sig.<0.0015, partial η^2=0.121	Wilks' Lambda=0.876, F=13.257, sig.<0.0005, partial η^2=0.124 No significant interaction effects, but level of summary length did have significant between-subjects main effects (F=16.774, sig.<0 0005, partial η^2=0.263)	Wilks' Lambda=0 827, F=19.614, sig.<0 0005, partial η^2=0 173

Table 6.1 Summary of the effects of expert and popular scoring templates on RSC and HS scores

Note: **1**=paired samples t-tests, **2**=paired samples t-tests for individual text, **3**=one-way repeated measures ANOVA by text type, **4**=Two-way repeated measures of ANOVA by text type and levels of TOEFL reading abilities, **5**=One-way repeated measures ANOVA by text type with TOEFL raw scores as a covariate, **6**=One-way repeated measures ANOVA by text type with summary length as a covariate, **7**=Two-way repeated measures ANOVA by text type and levels of summary length

6.1.2 Expert and popular templates in regression analyses

Correlations between the 8 scores of summaries (4 RSC and 4 HS) and FCE or TOEFL were examined. Further stepwise regression analyses[3] were conducted with FCE or TOEFL as the dependent variable and EERSC/EPRSC, CERSC/CPRSC, EEHS/EPHS, and CEHS/CPHS as independent variables in the models respectively. Assumptions for regression analyses were also checked[4], with no serious violations noted. It was found that it was always the scores from the expert templates that had a better chance of being retained in the models, except for RSC scores of the English summaries (Table 6.2) when the scores from the two templates were competing for entrance in the stepwise regression analyses (TOEFL or FCE as dependent variables). However, in four of these models, neither of the two independent variables was able significantly to predict the outcomes. The relationships between students' summarization performances and their language abilities will be further discussed in Chapter 7.

Dependent variable	Independent variables							
	EERSC	EPRSC	CERSC	CPRSC	EEHS	EPHS	CEHS	CPHS
FCE	.022 n.s.	.048 n.s.	.221 Sig.<.0145	.181 Sig.<.0375	.028 n.s.	-.009 n s	.182 Sig.<.0375	.113 n s
	Neither*		CERSC*, F (1, 96)=4.915, sig.<.0295		Neither*		Neither*	
TOEFL	.253 Sig.<.0065	.294 Sig <.0025	.257 Sig.<.0055	.233 Sig.<.0105	.181 Sig.<0.0365	.164 n s	.250 Sig <.0065	.216 Sig.<.0155
	EPRSC*, F (1, 98)=9.256, sig.<.0035		CERSC*, F (1, 98) = 6.930 Sig <.0105		Neither*		CEHS*, F (1, 98)=6.540 Sig.<.0125	

Note: Variable(s) kept in the stepwise regression analyses

Table 6.2: Correlation between the eight scores (4 RSC, 4 HS) and TOEFL/FCE, and results from the stepwise regression analyses

The following section **(6.2)** analyzes the students' attitudes towards the use of expert and popular templates.

---

[3] It should be noted that a slight difference in the correlation coefficients between an independent variable and the dependent variable can be significant in determining whether the independent variable can be kept in the stepwise regression analyses where "the order of entry of predictors into the regression equation is determined via a mathematical maximization procedure" (Stevens 2002: 121). For the purpose of this enquiry (i.e. which can better predict FCE or TOEFL score), the stepwise regression analysis is considered appropriate, although it is indeed the case that the two comparative scores from the templates have quite similar correlation coefficients with FCE and TOEFL. Most of the time, when one score has a significant correlation with FCE or TOEFL, the correlation between the other score and FCE or TOEFL is also significant.

[4] For the procedures of assumption checking, please see Chapter 7.

6.2 Post-summarization interviews

The 22 interviewees[5] were given a full explanation of how the expert and popular scoring templates were to be generated and how they would be used as criteria to evaluate the students' summaries, most of them (13 out of 22) strongly preferred the expert templates, 7 participants expressed a preference for the popular templates, and 2 participants would accept both templates. The interview data were analyzed using winMAX (Kuckartz 1998). The reasons all the interviewees argued for or against particular templates focused mainly on **1)** the degree of experience and language abilities of students and experts in understanding and summarizing the source text, **2)** the stereotypical status and common practice of using students and experts in educational assessment, and **3)** the dialectical interpretations of "quantity" and "quality".

6.2.1 Experience and language abilities of students and experts

The pro-expert students thought experts were more *authoritative, knowledgeable, learned, able, older and experienced,* and therefore could have deeper and fuller (or correct) understanding of the source text. They were also more advanced in using summarization strategies. Their expertise in the topic, simply because they were experts, would also help them to write better summaries. Furthermore, experts were more advanced in language abilities, which made them an unquestionable authority which was not only associated with their expertise but also their language abilities, and their age. Although I did not mention the age of the experts, the connotation of the term "expert" (*zhuan jia*, in Chinese) seemed to endow the experts with superiority over the students associated with age. These students also thought that it would be rare for experts to make mistakes. Experts did the summarization tasks seriously because they were experts and they were invited properly by the researcher. Expert summaries were therefore more reliable, just and would be fairer to the students to evaluate the students' summaries. On the other hand, the student template was considered less authoritative, less convincing, and less persuasive, because the students' experience and language abilities were not

[5] Altogether, 24 students were interviewed. Unfortunately, one student was not asked to comment on the use of two scoring templates, and another student's response to this question was inaudible and therefore could not fully be transcribed.

comparable to those of the experts and they would make mistakes. Although the students considered that there could also be some top students whose summaries would be close to the experts', they felt that unfortunately these would only be the minority. In the following extract, Alice Zhang[6] eloquently expressed her concerns regarding the use of a popular template and stressed the gap in abilities between experts and students. She frequently commented "authority is authority, after all."

> ... I think I would like to have the expert scoring template to judge my summaries. I am not sure how many, the percentage of very top students in our group of over 160 students. Authority is authority, after all. It would be less possible for them to make mistakes than for us students. It is not because I believe in experts or authorities without any suspicion. I don't think the student scoring template would have any authority in this respect. Our overall language abilities, I mean, we, the students in *this* department, may not be good enough to have confidence in generating a scoring template from our summaries to judge our own summaries, even if from the top students in our group..... Among the 160 students, there might be a couple of students who may be able to produce some good statements which would be very close to the experts' versions, but they could only represent a minority and will not be able to be included in the popular scoring template simply because they are the minority. Among the 160 students, there are good and not so good students, it would be extremely difficult for them at the same status [of language abilities]. It is possible that very good, and right statements from top students could exist but they only occur occasionally, and may not be included. ... Authority is, after all, authority. They are so experienced, full of many years of experience. At least, they are much better than us in language abilities. .... It is possible that some top students may be approaching their [experts'] level of language abilities, experience, and so on, but it would be rare and [their statements] be drowned in the popular statements of other students which may be included in the scoring templates simply because they are the majority, but these statements may not be as appropriate as the statements from the minority of top students. Authority is authority, after all. This doesn't mean that we are accepting authority without any suspicion. Good and top students are always the minority, however.
>
> **Extract 6.1**: Alice Zhang

Similarly, several other students challenged me during the interviews about the gap between experts and students, particularly in terms of their English language abilities and other expertise. Expert templates were far more convincing and authoritative than the popular templates, they argued.

> ... We can't compare our abilities with experts' abilities, can we?! ... If I want to know my real language abilities, I would like to use the expert template to evaluate my own proficiency, my level/status of proficiency, according to their stricter and higher and more objective standard.
>
> **Extract 6.2**: Helen Zhang

---

[6] To anonymize the interviewees, they were given Western first names and the same surname of Zhang, arguably the most common Chinese surname. Their names are in alphabetic order: Alice, Ben, Cindy, Daniel, Elyn, Fred, Grace, Helen, Ian, Jake, Katie, Louis, Michael, Nancy, Ollie, Peter, Quentin, Rachel, Simon, Tom, Ulysses, Victoria, Wendy, Yvonne. The names are not necessarily indicative of their actual gender.

> I prefer the expert template, because they are called experts, and because it has been a common practice that it is always from experts that the "*biao zhun da an*" (the standard answer) is created. ... We are all students; no student is more authoritative than another, no one is more persuasive than another.
>
> **Extract 6.3**: Ulysses Zhang

> ... Experts' are more authoritative, I don't think I could write a better summary than experts. ... In our mind, we seem to have such a strong "belief" that experts are experts and we use them as a scale to judge students, although there do exist some top students.
>
> **Extract 6.4**: Wendy Zhang

The pro-popular students viewed experience and language abilities from another perspective. They thought experts were *too* experienced and *too* familiar with the source text. The expert template would be so authoritative that the students could be pressurized. These students thought that there would be too great a discrepancy in experience and language abilities between experts and students and also among experts themselves, while the young students would form a similarly homogeneous group of language abilities and experience and have quite similar views/understandings of a source text, which rendered the popular template more able to reflect faithfully the students' current situation and would be fairer to students. The pro-popular students also argued that they could have their own unique understanding of a source text; and sometimes much better understanding than experts because "old experts may have generation gaps with young students, and can't fully understand the texts which young students can fully appreciate".

> I prefer the popular template. It is to test the students' understanding of the text, but some experts may have spent lots of time in "researching" the text and have deeper understandings of it, which are not at the same level of students' understandings: students may not have ever thought of them. Students' understandings of a text may be something we students can all understand and accept and consider them very important which expert may well ignore. .... I noticed the test direction that our summaries would also be judged according to our own summaries. I was shocked at first glance. It is too strange. How could it be? Won't there be any sample summary or reference to judge our summaries? But now looking back, I think the popular template is more flexible and could better reflect our students' situation. ... There is another issue of level of understanding, level of reading comprehension abilities. Since they are experts in a particular area, they have deeper understanding of the text. Sometimes, students may have much difficulty in understanding something abstract, even if it is written in Chinese. ... Every one has her own understandings or views of a text, for example, one expert may be in favour of one idea which students may be against, or don't like the idea. There are differences. Experts may list all the supporting details for that idea, but students may list other details to be against the expert's idea, vice versa. ... This is not to say which is better than the other; it is only an issue of difference. Experts can have their advantages, but students can have their advantages, too. It is particularly

> true if a text is suitable for the students' age. Old experts may have generation gap with young students, and can't fully understand the texts which young students can fully appreciate.
>
> **Extract 6.5** Fred Zhang

> I think the majority of us would prefer the student template, because we (students) are almost at the same level of (language) proficiency. But the experts are more authoritative... It (expert template) is something like "*biao zhun da an*" (the standard answer) in all the examinations in Chinese educational system, but it also makes us feel pressurized. If it (expert template) is used and you do not answer correctly, we feel we are completely defeated even if we make one wrong step only, something like in a chess game. However, students may have our own unique understandings of a text, and we are at almost the same level of proficiency, hence, it (student template) would be closer to our own understandings, it is not a so-called "*biao zhun da an*".
>
> **Extract 6.6** Quentin Zhang

6.2.2 Stereotypical status and common practice of using students and experts in educational assessment

In the interview data, experts earned and strengthened their indisputable authority and superiority over students not only through their experience and language abilities as some connotations of *zhuan jia*, but also by their stereotypical status and the common practice of using students and experts in educational assessment. Students were to be evaluated, not be valued, while experts were appreciated as able to create the standard answer (*biao zhun da an*, in Chinese), as some pro-expert participants commented.

> You can't use students to judge students themselves. You need a proper reference to judge the students' summaries, and this reference is the expert template.
>
> **Extract 6.7** Victoria Zhang

> I would accept the expert template. Perhaps, it is because we are in arts and social sciences department and we are so used to having such a similar scoring scheme in examinations for so many years – if you mention one thing listed in a model answer, and then you will be awarded a point. We are used to this kind of scheme for many years now.
>
> **Extract 6.8** Ollie Zhang

> Although their understandings may be different from ours to a large extent, I still prefer expert template because it has always been like this – we use expert views, we have no choice.
>
> **Extract 6.9** Grace Zhang

Even the pro-popular students found it strange and unbelievable that both expert and popular templates would be used (see Extract 6.5 Fred Zhang).

6.2.3 Dialectical interpretations of "quantity" vs. "quality"

Dialectical interpretations of "quantity" vs. "quality" were only explored by the pro-expert participants in interviews. They argued that some good and proper statements of the few top students may well not be included in the popular template simply because they were the minority and their statements could be *drowned* in the not-so-good statements of the majority (see Extract 6.1: Alice Zhang).

The **pro-expert** participants also argued that being popular did not necessarily mean being right.

> … It is not necessarily the popular statements are right because more people agree to such a statement. Are you sure then a popular statement must be right?
>
> **Extract 6.10** Cindy Zhang

> I trust experts, because I think my summary cannot be better than experts'. … Sometimes, the truth is usually in the hands of the minority. …
>
> **Extract 6.11** Peter Zhang

And sometimes too many versions of summaries of a text could also make it extremely difficult to reach an agreement.

> I like the expert template because the differences among students can be too wide. There are less differences and deviations in expert's summaries because there are only a few experts, and also because they are experts. Too many students. Even though the popular template may be close to our own understanding, but if we have to choose one template, experts' would be better because their understandings would be more focused than ours…
>
> **Extract 6.12** Grace Zhang

They also interpreted a very famous Chinese saying from a different perspective. In the Chinese saying, *san ge chou pi jiang* (three ordinary people) can be equal to *yi ge zhu ge liang* (one wizard). However, the pro-expert participants thought *san ge chou pi giang* could never be equal to *yi ge zhu ge liang* if they were not good enough.

> … Sometimes *san ge chou pi jiang* is not equal to *yi ge zhu ge liang*, if the three *chou pi jiang* are not good enough.
>
> **Extract 6.13** Elyn Zhang

> Further the students' advantage over experts only lies in the number of students, but the quantity of students does not necessarily improve the quality of the popular template, it is only a sort of accumulation, because most of the students are at approximately the same level of abilities. I would suggest choosing only the elite students or with appropriate proportions of students of different abilities to create the popular template. The number of students does not matter. I don't think it will make much difference from keeping all students.
>
> **Extract 6.14** Louis Zhang

6.2.4 Two "indifferent" students

Two participants who would accept both scoring templates reached their conclusion from quite different perspectives. One participant said:

> I think it won't make too much difference whether it was judged by expert or popular scoring templates. These two templates would be approximately the same. What experts would list in their summaries as key statements, we can also include them in our summaries. The difference may lie in the ways we express these statements, but not in frequencies of these statements.
>
> **Extract 6.15** Ben Zhang

The other participant thought it was not his job to decide which scoring template to use. He considered this the job for testers. However, when asked which one he would choose if he had to, he was very much inclined to the expert template on similar grounds as other pro-expert participants, as discussed above.

> I would accept both scoring templates because my job is to finish the tasks, but it is you testers' job to decide how to judge my summary. … If I have to choose one, I think I would like the expert scoring template, because experts are authority anyhow and they are more learned than us. … If you choose us, students in this department, it is definitely not right to generate such a popular scoring template! To tell the truth, the English language proficiency of students in our Year is not very good. If you let them summarize the text, there must be many parts of the text that the students don't fully understand and their summaries are not satisfactory. I think experts (teachers) must be invited.
>
> **Extract 6.16** Daniel Zhang

6.2.5 Students' further suggestions

One pro-expert participant also suggested using both scoring templates and comparing the difference, which was exactly one of the aims of this research.

> In fact, both of them are OK, you can use both to judge our summaries and compare the results. …
>
> **Extract 6.17** Helen Zhang

Another pro-expert participant suggested that the potential reader of the summaries, as stated in the test directions, evaluate the summaries based on whether s/he could understand the summaries.

> In fact, I think scoring criteria should not be so rigid. It does not really mean that a summary is necessarily good if it includes one particular statement. … You said in the test direction that the summary was written for a friend who had not read the text, so I think the best way to judge the quality of summaries is to ask the friend to read and compare two summaries to see which summary he can understand better. … Whether one point is important or not in a source text is really up to the readers: what one reader thinks important may not be important to another reader. But if you do have to have a scoring template, it is undoubtedly an expert template. Experts must be better than students who may make the same mistakes.
>
> **Extract 6.18** Ulysses Zhang

I now turn to a summary of the findings relating to RQ1.

6.3 Summary of findings relating to RQ1

The main purpose of this research question was to investigate (*a*) potential variances in the RSC and HS due to the use of the expert and the popular scoring templates and (*b*) students' views on their possible contributions to the development of such assessment criteria.

With regard to (*a*), the quantitative data of students' summarization performances clearly and consistently demonstrated that the use of two scoring templates could make substantial differences to the RSC and HS scores a summary could be assigned (Table 6.1). A summary received much higher RSC and HS scores when it was evaluated according to the popular templates than the expert templates. The effect sizes were very large on both RSC and HS, albeit with different magnitudes, and also on the English and Chinese summaries. For the English summaries, the effects on RSC were much bigger than on HS; for the Chinese summaries, the effect sizes on RSC and HS were approximately at the same level. Although summaries received higher scores when evaluated according to the popular templates, it was the scores according to the expert templates that had a far better chance of predicting the students' TOEFL-R or FCE-R. This was particularly true for scoring of the Chinese summaries using expert templates (Table 6.2).

With regard to (*b*), it seemed that the majority of the students strongly preferred the expert templates, arguing quite convincingly from several perspectives, in particular (i) their inferior experience and English language and summarization abilities, compared to English native-speaker experts, (ii) stereotypical status and the common practice of using experts, rather than students, to create "*biao zhun da an*" (the standard answer) in educational assessment, and (iii) "quantity" did not necessarily guarantee "quality". The majority of the students interviewed were used to and ready to accept the common practice of using experts' authoritative standards to evaluate students' performances.

Further discussion of the findings is reported in **11.2.1.**

CHAPTER SEVEN

Summarization Performances and Other Language Abilities

Are students' summarization performances affected by their other linguistic abilities and if so, to what extent?

This research question (RQ2) aims to examine whether the students' summarization performances were mainly attributable to their English reading comprehension or other language abilities such as English and Chinese writing and translation from English to Chinese, which were claimed to be muddying students' summary writing (see **2.5.4**). To put it in another way, three sub-questions were asked:

- *Do traditional summarization tasks measure students' reading comprehension abilities as TOEFL-R and FCE-R and, if so, to what extent?*
- *Does students' general EFL writing ability affect their English summarization performances and, if so, to what extent? Is general EFL writing ability a determining factor in students' English summarization performances?*
- *Does Chinese summarization of English texts involve students' translation abilities (from English to Chinese), and/or Chinese writing abilities and, if so, to what extent?*

The relationships between students' summarization performances and their other language abilities were first of all examined in a snapshot-like approach, using one-way ANOVA with factors of TOEFL-R, FCE-R, English Writing, Chinese Writing and Translation (Low, Medium, and High). Their performances were then subjected to more detailed stepwise regression analyses (see Figure 7.1). Following the stepwise regressions on the first block of independents (language abilities), sequential regressions were then conducted with *text type* dummy variables in the second block of independents (see also Chapter 10).

Students' perceptions of these relationships were analysed qualitatively and are reported in Chapter 8, since these are very much intertwined with the effects of *language* and *language order* for the summarization tasks.

7.1 The mean differences in summarization performances between language ability groups

In order to have a brief overview of the relationships between students' summarization performances and their other language abilities, several one-way ANOVAs were conducted in relation to three categories (Low, Medium, and High) of the language abilities: *TOEFL-R*, *FCE-R*, *English Writing*, *Chinese Writing* and *Translation*.

With reference to the three levels of abilities as measured by *FCE-R*, *English Writing*, *Chinese Writing* and *Translation*, no statistically significant difference was found in any of the 8 scores for the summarization performances. Only *TOEFL-R* seemed to be able to differentiate EPRSC ($F_{2,97}$=4.605, sig.<.0125). The post hoc Scheffe test found that the significant difference in EPRSC was mainly attributable to the big difference between Low and High *TOEFL-R* summaries (mean difference=-7.476, sig.<.0165). However, no statistically significant difference was observed in the other 7 scores for summaries between the three levels of *TOEFL-R*.

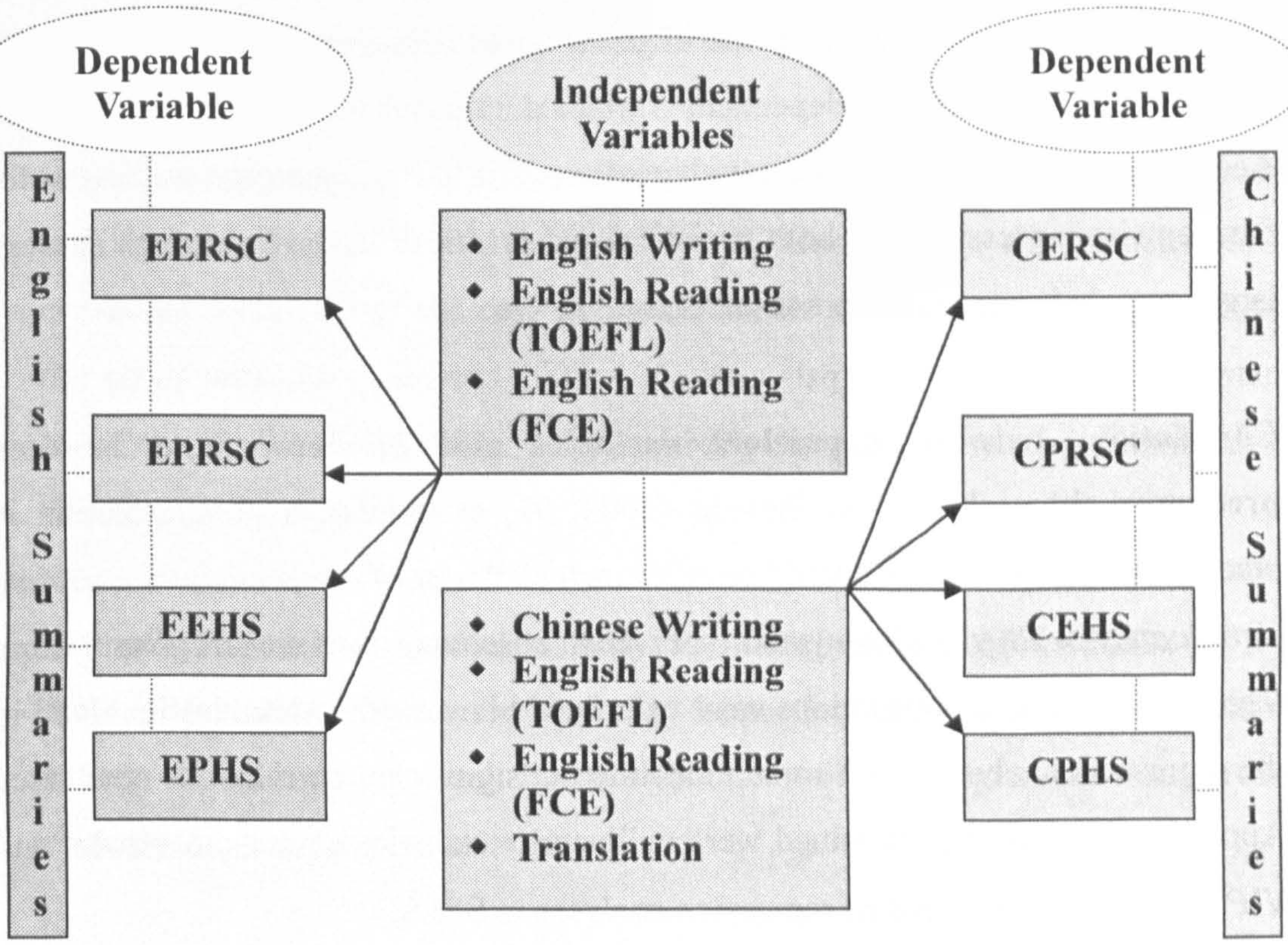
7.2 Multiple regression analyses

Students' performance data were further subjected to multiple regression analyses following two procedures (see Figure 7.1). In Procedure A, only one block of independent variables was subjected to stepwise regressions (*TOEFL-R*, *FCE-R* and *English Writing* for English summarization performances; *Chinese Writing*, *TOEFL-R*, *FCE-R* and *Translation* for Chinese summarization performances). In Procedure B, sequential regressions were conducted with *text type* as dummy variable[1] entered in the second block of independent variables to see if *text type* might make a significant additional contribution to the variances of the dependents. Within the second block of independent variables, the stepwise method was also used.

In this section, the procedures used to check the assumptions for multiple regressions are first reported. Findings from the analyses are reported separately for the English and Chinese summarization performances (English: **7.2.2**; Chinese: **7.2.3**).

---

[1] The *text type* dummy variables were set up in the following way. When textA was 1 (dummy variable one), the other two were 0; when textB was 1 (dummy variable two), the other two were 0; when textC was 1 (dummy variable three), the other two were 0.



Notes:

- Assumptions for multiple regressions were checked for each analysis.
- Stepping method criteria using probability-of-F-to-Enter ≤ .050, Probability-of-F-to-remove ≥ .100.

Figure 7.1 Plan for the multiple regression analyses on summarization performances and other language abilities

7.2.1 Checking assumptions of multiple regressions

Some key assumptions of multiple regression analyses were checked such as sample size, correlations between dependent and independent variables and between independent variables themselves, outliers and influential data points, normality, linearity, homoscedasticity and independence of residuals. No serious violations of the assumptions were noted.

- **Sample size**: According to Stevens (2002: 88), for social science research, about 15 subjects per predictor are needed for a reliable equation. Tabachnick and Fidell (2003: 117), quoting Green (1991), suggest a more conservative formula for calculating sample size requirements ($n \geq 50 + 8m$ for testing multiple correlation

and $n \geq 104 + m$ for testing individual predictors, where m=number of predictors). More cases are needed for stepwise regression, as Tabachnick and Fidell (2003) suggest, and a cases-to-independent variables ratio of 40 to 1 is reasonable. According to these criteria, the number of cases in this project was sufficient for the stepwise regressions for both statistical and practical reasons, although an even larger sample is needed for cross validation.

- **Correlations between dependent variables and predictors and between predictors themselves:** As Stevens (2002: 91) comments, a good situation in practice to obtain a high R would be one in which "*most of our predictors correlate significantly with y and the predictors have relatively low correlations among themselves*." Simple correlations were examined between the variables involved in the regression analyses, with some moderate but significant correlations noted (see Appendix 19). Further examined were collinearity statistics such as tolerance[2] and VIF[3], as part of the output of regression analyses in SPSS.

- **Outliers and influential data points[4]:** Multiple regression can be "very sensitive to data points that 'split off' or are different from the rest of the points, that is, to outliers" (Stevens 2002: 125); "just one or two such points can affect the interpretation of results" (*ibid.*).

To identify cases that were very different from the rest of the sample on the ***set of***

---

[2] Tolerance is $1-R^2$ of that predictor on all the other predictors, ignoring the dependent. The higher the intercorrelation of the predictors, the more likely the tolerance will approach zero. As a rule of thumb, if the tolerance is less than .20, a problem with multicollinearity is indicated. When the tolerance is close to zero there is a high multicollinearity of that predictor with other predictors and the b and beta coefficients will be unstable. The greater the multicollinearity and the lower the tolerance, the greater the standard error of the regression coefficients. Stevens (2002: 92) points out three reasons why multicollinearity could cause a serious problem for multiple regression: (1) it severely limits the size of R, because the predictors are going after much of the same variance on the dependent; (2) it makes determining the importance of a given predictor difficult as the effects of the predictors are confounded due to the high correlations among them; (3) it increases the variances of the regression coefficients and can make the prediction equation unstable.

[3] The variance inflation factor – VIF is simply the reciprocal of tolerance. VIF for a predictor indicates whether there is a strong linear association between it and all the remaining predictors which would cause concern for multicollinearity when "any VIF exceeds 10, there is reason for at least some concern; then one should consider variable deletion or an alternative to least squares estimation to combat the problem" (Myers 1990: 369, as cited in Stevens 2002: 92-93)

[4] "There is a distinction between the two because a point that is an outlier (either on *y* or for the predictors) will *not necessarily* [original emphasis] be influential in affecting the regression equation." (Stevens 2002: 125).

predictors, the hat elements (i.e. leverage values) were examined as part of the output from the regression analyses in SPSS. According to Hoaglin and Welsch (1978, as cited in Stevens 2002: 126), centered leverage values greater than $3p/n$ (p=*number of predictors+1; n=number of subjects*) should be a cause for concern and be carefully examined[5]. Cook's distance (Cook 1977), which measures the ***combined influence of the case being an outlier on dependent variable and on the set of predictors*** (Stevens 2002: 126), was also evaluated for Cook's distance >1 (Cook & Weisberg 1982). Mahalanobis's distances were also examined. No violations were noted.

- **Normality, linearity, homoscedasticity, and independence of residuals:**
  These assumptions were checked by inspecting the standardized residuals scatterplots and the Normal Probability Plots of the regression standardized residuals. No serious violations were noted.

7.2.2 English summarization performances

1) RSC

In Procedure A, it was found that only TOEFL-R could explain a small, but significant, proportion of the variances of EERSC (R^2 change=.054, $F_{1, 96}$=5.533, sig.<0.0215). The other two predictors (FCE-R and English Writing) were excluded due to the very low partial correlations with the dependent (-.061 for FCE-R and .075 for English Writing). However, in Procedure B where the second block of independent variable (text type dummy variable one)[6] was added in the sequential regression analysis, TOEFL-R and text type *together* made a significant contribution to EERSC (R^2= .106, $F_{2,95}$=5.619, sig.<.0055). R^2 change of the second model (=.051, $F_{1,95}$ = 5.449, sig.<0.0225) was as big as the R^2 of the first model (=.054).

---

[5] Hoaglin and Welsch (1978) actually suggested 2p/n may be considered large; Stevens (2002: 126) suggested a less strict formula because 2p/n "can lead to more points then [sic] we really would want to examine."

[6] There were three text types altogether. However, RSC scores were only assigned to textA and textC summaries (see Table 4.8), therefore only one dummy variable was involved in the sequential regression analyses when RSC scores were the dependent variables. As for HS scores of expert templates, two dummy variables were included in the second block of independent variables.

The standardized coefficients (Beta) in the second model indicated that *text type* had larger impacts on EERSC than *TOEFL-R* (Table 7.1). Students who summarized textA seemed to have been advantaged to a considerable extent. In this model, *TOEFL-R* was no longer significant (t=1.624, n.s.) due to the substantially greater impacts of *text type*.

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	24.303	9.506		2.557	.012	5.434	43.171		
	TOEFL reading	.603	.256	.233	2.352	.021	.094	1.111	1.000	1.000
2	(Constant)	27.789	9.412		2.952	.004	9.104	46.474		
	TOEFL reading	.425	.262	.165	1.624	.108	-.095	.945	.916	1.092
	txt type dummy variable 1	5.641	2.417	.237	2.334	.022	.843	10.439	.916	1.092

a. Dependent Variable: EERSC

Table 7.1 EERSC and TOEFL-R

Similar results were obtained from the analyses on EPRSC in Procedure A. *TOEFL-R* predicted a small, but significant, amount of EPRSC (R^2=.072, $F_{1,96}$=7.431, sig.<.0085; EPRSC=29.326+0.65**TOEFL-R*). The higher TOEFL-R a student had, the higher his/her EPRSC was predicted. The two excluded predictors had very low partial correlations with EPRSC (-.046 for FCE-R and .08 for English Writing). However, in Procedure B analyses, it was found that *text type* did not make significant additional contribution. In other words, it was still *TOEFL-R* that was statistically significant to predict EPRSC. *Text type*, *FCE-R* and *English Writing* were excluded in the sequential regression analysis (Table 7.2).

Excluded Variables [b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance	VIF	Minimum Tolerance
1	FCE reading all parts	-.047[a]	-.448	.655	-.046	.888	1.126	.888
	English writing	.077[a]	.778	.438	.080	.990	1.010	.990
	txt type dummy variable 1	.083[a]	.806	.422	.082	.916	1.092	.916

a. Predictors in the Model: (Constant), TOEFL reading

b. Dependent Variable: EPRSC

Table 7.2 Excluded variables in the sequential regressions on EPRSC and language abilities and text type

2) HS

In the stepwise regression analyses, none of the three independent variables were able to predict a statistically significant amount of EEHS or EPHS. In order to gain an

overview of the differential contributions of each predictor, regression analyses using the enter method were also conducted (EEHS: R^2=.028, $F_{3, 149}$=1.406, n.s., EPHS: R^2=.033, $F_{3, 94}$=1.079, n.s.). Comparatively speaking, it was still TOEFL-R that could probably better predict EEHS and EPHS than FCE-R and English Writing, although not to a statistically significant extent.

In the follow-up sequential regression analysis, it was found that *text type* (dummy variable 2) was able to predict a significant amount of EEHS, although quite small (R^2=.033, $F_{1,151}$=5.107, sig.<.0255; EEHS=10.24+0.778**text type* dummy variable 2). Summarizers of textB were slightly advantaged. Among the excluded predictors, TOEFL-R had the highest partial correlation with EEHS (Table 7.3). The effect of *text type* on EPHS was not significant.

Excluded Variables [b]

Model		Beta In	t	Sig	Partial Correlation	Collinearity Statistics: Tolerance	VIF	Minimum Tolerance
1	TOEFL reading	.139[a]	1.738	.084	.141	.989	1 012	989
	FCE reading all parts	.111[a]	1.391	.166	.113	995	1 005	.995
	English writing	-.058[a]	-.722	.471	-.059	.985	1.015	985
	txt type dummy variable 1	.159[a]	1 675	.096	135	.703	1 423	703

[a] Predictors in the Model. (Constant), txt type dummy variable 2

[b] Dependent Vanable EEHS

Table 7.3 Excluded variables in the sequential regressions on EEHS and language abilities and text type

3) Summary

TOEFL-R seemed to be the only statistically significant predictor (among FCE-R and English Writing) of RSC scores on the English summaries. FCE-R and English Writing were far less capable of accounting for the variances in the RSC scores. The higher the TOEFL-R a student achieved, the greater the likelihood of getting higher EERSC and EPRSC. Although statistically significant, TOEFL-R could only explain a very small proportion of the variances (5.4% for EERSC and 7.2% for EPRSC). In terms of the overall quality of summaries (EEHS and EPHS), none of the three predictors were able to explain a statistically significant amount of the variances. However, comparatively speaking, it was still TOEFL-R that could probably better predict the overall quality scores than FCE-R or English Writing.

When *text type* was added in the follow-up sequential regression analyses, it explained statistically significantly an additional 5% and 3% of the variances of EERSC and EEHS respectively. The additional 5% and 3% were almost at the same level as that accounted for by TOEFL-R alone for EERSC and that accounted for by the three predictors of language abilities together for EEHS. TextA students were advantaged in terms of receiving higher EERSC, and textB students in terms of higher EEHS, language abilities held constant. However, *text type* did not seem to make significant difference on EPRSC and EPHS (scores of English summaries according to the popular template). This to some extent supports not only the findings of (*a*) the differential effects of the expert and the popular templates (see Chapter 6) and (*b*) the overall effects of *text type* on summarization performances (see Chapter 10).

7.2.3 Chinese summarization performances

1) RSC

In the stepwise regression analyses of the first block of independents (TOEFL-R, FCE-R, Chinese Writing and Translation), it was found that only TOEFL-R was able to predict a statistically significant but small amount of the RSC scores of Chinese summaries (CERSC: R^2=.058, $F_{1,95}$=5.898, sig.<.0175; CPRSC: R^2=.042, $F_{1,95}$=4.201, sig.<.0435). The follow-up sequential regression analyses were conducted with *text type* dummy variable in the second block of independents. It was found the R^2 changes (CERSC: R^2 change=.081, $F_{1,94}$=8.907, sig.<.0045; CPRSC: R^2 change=.098, $F_{1,94}$=10.744, sig.<.0015) were statistically significant. The overall R^2 in the second models was increased to around 14% and was statistically significant (CERSC: $F_{2,94}$=7.648, sig.<.0015; CPRSC: $F_{2,94}$=7.688, sig.<.0015).

Higher TOEFL-R students tended to receive higher CERSC and CPRSC, regardless of which scoring templates were used. If the source text happened to be textA, the chance of getting higher RSC was further significantly boosted (CERSC=20.124 + 0.489**TOEFL* + 8.826**text type* dummy one; CPRSC = 32.089 + 0.29*TOEFL + 8.118*text type dummy one). A difference of 8.826 or 8.118 was quite large against the means of the RSC in the region of 47. Furthermore, *text type* seemed

to be able to "dwarf" the effects of TOEFL-R on summarization performances. They were no longer significant when *text type* was a predictor in the models.

FCE-R, Chinese writing, and translation were all excluded in the models. These three independent variables were far less able to predict RSC of Chinese summaries than *TOEFL-R* or *text type*. The partial correlations of these excluded predictors with CERSC and CPRSC were all very small, ranging from -0.011 to 0.162 in the models.

2) HS

a) CEHS

In the stepwise regression analysis, TOEFL-R predicted a small, but statistically significant, amount of CEHS (R^2=.043, $F_{1,148}$=6.601, sig.<.0115). The other three predictors were excluded (see Table 7.6). When the data was subjected to sequential regression analyses with *text type* dummy variables (one and two) in the second block of independents where stepwise method was used, *text type* dummy variables exerted significant additional impact on CEHS (Model 2: R^2 change =.047, F change $_{1,147}$=7.510, sig.<.0075; Model 3: R^2 change=.054, F change $_{1,146}$=9.228, sig.<.0035). The summary of the three models is presented in Table 7.4 below. The best Model 3, where TOEFL-R and the two dummy variables were included, was able to explain around 14% of the variances of CEHS. In Model 1, TOEFL-R alone accounted for about 4% of CEHS; in Model 2, TOEFL-R together with *text type* dummy variable one accounted for about 9% of CEHS.

Model Summary [d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig F Change
1	.207[a]	.043	.036	2.3298	.043	6.601	1	148	.011
2	.299[b]	.089	.077	2 2802	.047	7.510	1	147	007
3	.379[c]	.143	.126	2 2190	.054	9 228	1	146	003

[a] Predictors (Constant), TOEFL reading

[b] Predictors (Constant), TOEFL reading, txt type dummy vanable 2

[c] Predictors (Constant), TOEFL reading, txt type dummy vanable 2, txt type dummy vanable 1

[d] Dependent Vanable CEHS

Table 7.4 Model summary of regressions on CEHS and TOEFL-R, FCE-R, Chinese writing and translation and text type dummies

The ANOVA statistics for the Rs in the regressions (Models 1, 2, & 3) are reported below (Table 7.5). All the three Rs are statistically significantly different from zero.

ANOVA [d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	35.829	1	35.829	6.601	.011[a]
	Residual	803.364	148	5.428		
	Total	839.193	149			
2	Regression	74.877	2	37.438	7.200	.001[b]
	Residual	764.317	147	5.199		
	Total	839.193	149			
3	Regression	120.314	3	40.105	8.145	.000[c]
	Residual	718 879	146	4.924		
	Total	839 193	149			

a Predictors: (Constant), TOEFL reading

b. Predictors (Constant), TOEFL reading, txt type dummy variable 2

c. Predictors (Constant), TOEFL reading, txt type dummy variable 2, txt type dummy variable 1

d Dependent Variable CEHS

Table 7.5 ANOVA statistics of CEHS regressions

The regression equations of the three models are:

- Model 1: CEHS = 5.798+0.104*TOEFL
- Model 2: CEHS = 4.942+0.117*TOEFL+1.075*text type dummy 2
- Model 3: CEHS = 5.232+0.08836*TOEFL+1.798*text type dummy 2 +1.412*text type dummy 1

As the three equations demonstrate, the higher the TOEFL-R, the greater the likelihood of having higher CEHS, although the magnitude of the importance of TOEFL-R varied across the three models. In addition to the effects of TOEFL-R, *text type* played a significant role in CEHS. TextC summarizers seemed to be disadvantaged. TextB summarizers were slightly more advantaged than textA summarizers. According to Model 3:

- if a participant summarized textA, CEHS is predicted as 5.232 + 0.08836*TOEFL + 1.142;
- if a participant summarized textB, CEHS is predicted as 5.232 + 0.08836*TOEFL + 1.798;
- if a participant summarized textC, CEHS is predicted as 5.232 + 0.08836*TOEFL.

In other words, a particular source text could probably make a difference of up to 1.798 in CEHS. This difference is meaningfully large, taking into account that the mean of CEHS is quite low (9.5).

Examination of the statistics of the excluded variables in the three models indicated that students' *translation*, *Chinese writing* and *FCE reading* abilities had very low partial correlations with CEHS (Table 7.6).

Excluded Variables [d]

Model		Beta In	t	Sig	Partial Correlation	Collinearity Statistics: Tolerance	VIF	Minimum Tolerance
1	FCE reading all parts	.133[a]	1.525	.129	.125	.843	1.186	.843
	Chinese writing	.127[a]	1.580	.116	.129	.989	1.011	.989
	translation	-.024[a]	-.265	.791	-.022	.800	1.250	.800
	txt type dummy variable 1	.086[a]	1.036	.302	.085	.932	1.073	.932
	txt type dummy variable 2	.217[a]	2.740	.007	.220	.986	1.015	.986
2	FCE reading all parts	.147[b]	1.719	.088	.141	.840	1.190	.837
	Chinese writing	.102[b]	1.280	.203	.105	.974	1.026	.971
	translation	-.017[b]	-.188	.851	-.016	.799	1.251	.793
	txt type dummy variable 1	.284[b]	3.038	.003	.244	.671	1.491	.671
3	FCE reading all parts	.143[c]	1.729	.086	.142	.840	1.190	.671
	Chinese writing	.093[c]	1.193	.235	.099	.973	1.028	.670
	translation	-.028[c]	-.328	.743	-.027	.798	1.254	.670

a. Predictors in the Model: (Constant), TOEFL reading

b. Predictors in the Model: (Constant), TOEFL reading, txt type dummy variable 2

c. Predictors in the Model: (Constant), TOEFL reading, txt type dummy variable 2, txt type dummy variable 1

d. Dependent Variable: CEHS

Table 7.6 Excluded variables in the regressions of CEHS and TOEFL-R, FCE-R, Chinese writing, translation and text type dummies

b) CPHS

The stepwise regressions found that none were able to predict a significant amount of variance in CPHS[7] and therefore no statistics were produced in SPSS. In order to gain an overview of the contribution of each individual independent, regression analysis using the enter method was also conducted (R^2=.033, $F_{4, 91}$=0.773, n.s.). The statistics of the standardized coefficients (Beta) seemed to indicate that TOEFL-R was probably the best predictor of CPHS, with the effect approaching the statistical significance level (Table 7.7).

Coefficients [a]

Model		Unstandardized Coefficients: B	Std. Error	Standardized Coefficients: Beta	t	Sig	95% Confidence Interval for B: Lower Bound	Upper Bound	Collinearity Statistics: Tolerance	VIF
1	(Constant)	7.786	2.435		3.198	.002	2.949	12.622		
	TOEFL reading	8.759E-02	.050	.198	1.753	.083	-.012	.187	.831	1.203
	FCE reading all parts	-1.121E-02	.037	-.033	-.301	.764	-.085	.063	.874	1.144
	Chinese writing	2.542E-02	.124	.022	.205	.838	-.221	.271	.968	1.033
	translation	-5.767E-02	.106	-.061	-.542	.589	-.269	.154	.843	1.187

a. Dependent Variable: CPHS

Table 7.7 Coefficients of TOEFL-R, FCE-R, Chinese writing and translation in the regressions of CPHS using enter method

In the sequential regression analysis to examine the additional impact of *text type* on CPHS, it was found that *text type* dummy variable one alone was able to account

[7] The casewise diagnostics indicated that one student (No.82, student No.=3406) had an extremely low CPHS = 2.5, an outlier outside three standard deviations. The analyses were therefore based on data excluding this student.

for an additional 5% of CPHS variances (R^2=.048, $F_{1,94}$=4.703, sig.<.0335). TextA seemed easier to summarize than textC. Students were advantaged by an extra score of 0.892 if they summarized textA, language abilities held constant.

3) Summary

Without taking into consideration the effects of *text type*, TOEFL-R was the only independent (among FCE-R, Chinese writing and translation) which was able to predict a small, but statistically significant, amount of the variances of CERSC (6%), CPRSC (4%) and CEHS (4%). The higher a student's TOEFL-R, the better the Chinese summary s/he would produce. However, CPHS could not be predicted by any of the four independent variables of language abilities. Overall, FCE-R, Chinese writing and translation were far less capable of predicting Chinese summarization performances than TOEFL-R.

In the follow-up sequential regression analyses, the additional effects of *text type* seemed evident, even with larger effects than TOEFL-R. Text type was able to contribute an additional 5% to 10% in explaining the variances of CERSC, CPRSC, CEHS and CPHS. Students who summarized textA and textB were advantaged over those who summarized textC. They had better chance of getting higher scores, language abilities held constant.

7.3 Summary of findings relating to RQ2

This research question examined (a) the contribution of students' language abilities, such as English reading, English and Chinese writing and translation, towards their summarization performances, and (b) the students' perceptions of such relations. It was found that TOEFL-R was the only language-related independent variable capable of predicting small, but statistically significant, amounts of the variances in both English and Chinese summarization performances (cf. **8.1.3**). However, not all of the eight scores were able to be predicted significantly by TOEFL-R. All of the four RSC scores (EERSC, EPRSC, CERSC, CPRSC) were predicted significantly by TOEFL-R, while three of the four HS scores were not (EEHS, EPHS, CPHS). It seemed that TOEFL-R was far more capable of predicting RSC than HS.

None of the other language abilities were able to predict significant proportions of the variances in students' summarization performances. It seemed that their abilities in writing (English or Chinese) and translation were not contributing to summarization performances as much as their reading comprehension abilities measured by TOEFL-R. It is interesting to note that TOEFL-R was better able to predict summarization performances than FCE-R.

Text type also made a small, but significant, additional impact on summarization performances. Those who summarized textA were advantaged over other students, and had a greater chance of producing better summaries, their language abilities held constant. In some situations (e.g. CERSC), *text type* even dwarfed the contributions of TOEFL-R. Furthermore, these additional effects, according to the statistics of the sequential regression analyses, were more prominent on the Chinese summaries than the English. Text type predicted significantly all the four indicators of the Chinese summarization performances, but only two quality indicators of the English summarization performances (EERSC and EEHS, i.e. those according to the expert template). This sheds further light on the significant effects of scoring template (expert vs. popular) on the scores a summary could be assigned (see Chapter 6) and the differential effects of the use of English and Chinese for the summarization tasks (see Chapter 8).

Students' views on the relationships between their summarization performances and other language abilities are reported in the next chapter.

Further discussion of these findings is reported in **11.2.2**.

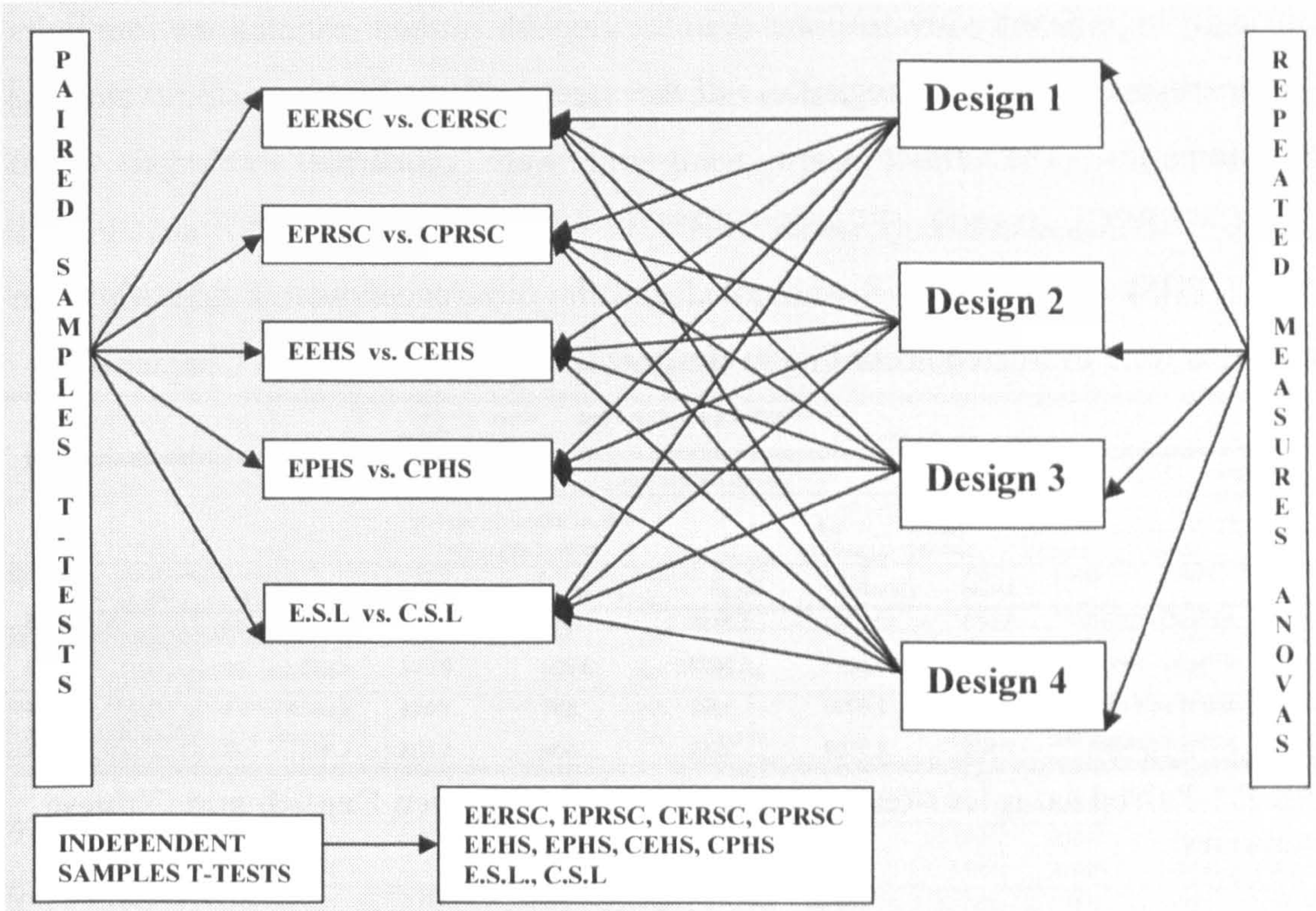
CHAPTER EIGHT

Language and Language Order

What impact does the use of a different language and language order have on summarization performances and measurement of reading comprehension abilities?

This research question (RQ3) was investigated first through statistical analyses of students' *actual* summarization performances in terms of the means differences in (a) RSC, HS and Length between the English and Chinese summaries and between summaries written in the order of *English then Chinese* and *Chinese then English* and (b) the relationships between the three quality indicators and students' language abilities. Further investigated were students' *perceptions* of the impact of different language and language order on their *actual* summarization performances.

Data were subjected to *t*-tests and then a series of repeated measures ANOVAs (Figure 8.1) by between-subjects factors of *text type* (TXT), *text presentation mode* (PRESMODE), and *summarization language order* (LANGORD) with *language* (LANG) as the within-subjects factor. Multiple regressions were also employed to examine which summarization performance (English or Chinese) could better predict TOEFL-R (see also Chapter 7).



Notes:

- Paired-samples *t*-tests: to examine the differences between the use of different language
- Independent samples *t*-tests: to examine the differences in each score between the use of different language order
- Repeated measures: within-subjects factor = Language (LANG) for the following four designs

Design 1=Intercept+LANGORD+TXT+LANGORD*TXT
Design 2=Intercept+LANGORD+PRESMODE+LANGORD*PRESMODE
Design 3=Intercept+TXT+PRESMODE+TXT*PRESMODE
Design 4 =
Intercept+LANGORD+TXT+PRESMODE+TXT*PRESMODE+TXT*LANGORD+PRESMODE*LANGORD+TXT*PRESMODE*LANGORD

Figure 8.1 Plan for the statistical analyses on the effects of language and language order on summarization performances

8.1 Students' actual summarization performances

As showed in Figure 8.1 above, the students' actual summarization performances were subjected to *t*-tests, repeated measures ANOVA and multiple regressions to investigate the effects of the use of different language and language order on their summarization performances.

8.1.1 T-tests

1) RSC and HS

In the paired samples *t*-tests of EERSC and CERSC, EPRSC and CPRSC, EEHS and CEHS, and EPHS and CPHS, it was found that the Chinese summaries received

statistically significantly lower scores than the English summaries in all the four pairs of comparisons (Table 8.1), regardless of the assessment criteria used for evaluating the summaries. The effect sizes using η^2 were moderate to large: 0.066 (EERSC/CERSC), 0.1809 (EPRSC/CPRSC), 0.1364 (EEHS/CEHS) and 0.1631 (EPHS/CPHS). Furthermore, it was noted that the differences were larger when the summaries were evaluated according to the popular templates (see also Chapter 6).

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	EERSC - CERSC	3.660	13.8358	1.3836	.915	6.405	2.645	99	.009
Pair 2	EPRSC - CPRSC	6.090	13.0226	1.3023	3.506	8.674	4.676	99	.000
Pair 3	EEHS - CEHS	.943	2.3798	.1899	.568	1.318	4.963	156	.000
Pair 4	EPHS - CPHS	1.105	2.5159	.2516	.606	1.604	4.392	99	.000

Table 8.1 Paired samples *t*-tests on the differences between English and Chinese summaries

However, the *t*-tests of summaries of each individual text did not fully bear out the same significant differences between the pairs (Table 8.2). In textA summaries, only the difference between EPHS and CPHS was statistically significant. The magnitude of differences also varied across different source texts. For example, textA summaries did not differ significantly between EEHS and CEHS; however textB and textC summaries did, but with a different magnitude (η^2=0.1153 for textB, η^2=0.2814 for textC summaries).

		Paired Differences							
Text	Pair	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Lower	Upper	t	df	Sig. (2-tailed)
A	EERSC/CERSC	2.160	13.2994	1.8268	-1.505	5.826	1.183	52	.242
	EPRSC/CPRSC	3.519	13.1854	1.8112	-.115	7.153	1.943		.057
	EEHS/CEHS	.575	2.3685	.3253	-.077	1.228	1.769		.083
	EPHS/CPSH	.858	2.5275	.3472	.162	1.555	2.473		.017
B	EEHS/CEHS	.833	2.3287	.3084	.215	1.451	2.702	56	.009
C	EERSC/CERSC	5.351	14.3712	2.0963	1.132	9.571	2.553	46	.014
	EPRSC/CPRSC	8.989	12.3387	1.7998	5.367	12.612	4.995		.000
	EEHS/CEHS	1.489	2.4058	.3509	.783	2.196	4.244		.000
	EPHS/CPSH	1.383	2.5005	.3647	.649	2.117	3.792		.000

Table 8.2 Paired samples *t*-tests on summaries of each individual text

These raise questions concerning (i) the possible effects of text type (TXT) on summarization performances and (ii) the insufficiency of *t*-tests in generating a fuller picture of the impact of *language* in association with other factors such as *text presentation mode* and *language order*.

To investigate the effects of language order (*English then Chinese*, *Chinese then English*) on RSC and HS, independent samples *t*-tests were conducted on the data of all the summaries (Table 8.3), as well as the summaries of each individual text. No significant difference was found in both cases. Please note that research design for this project deliberately controlled for the effects of language order (see Table 4.6).

Independent Samples Test

		Levene's Test for Equality of Vanances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig	t	df	Sig (2-tailed)	Mean Difference	Std Error Difference	Lower	Upper
EERSC	Equal vanances assumed	026	873	.300	98	765	721	2 4038	-4 0493	5 4911
	Equal vanances not assumed			300	97 944	765	721	2 4030	-4 0478	5 4895
EEHS	Equal vanances assumed	084	773	295	155	768	098	3314	- 5569	7524
	Equal vanances not assumed			.295	154 476	768	098	3309	- 5560	7514
CERSC	Equal vanances assumed	1 814	181	-1 545	98	126	-4 548	2 9441	-10 3900	1 2948
	Equal vanances not assumed			-1 550	96 717	.125	-4 548	2 9348	-10 3727	1 2775
CEHS	Equal vanances assumed	.081	.776	-1.067	155	.288	- 405	3798	-1 1554	3450
	Equal vanances not assumed			-1 065	152 597	288	- 405	3804	-1 1567	3463
EPRSC	Equal vanances assumed	358	551	1 716	98	089	3 839	2 2371	- 6003	8 2786
	Equal vanances not assumed			1 715	97 479	090	3 839	2 2386	- 6035	8 2818
EPHS	Equal vanances assumed	3 150	.079	1 838	98	.069	723	3933	- 0575	1 5033
	Equal vanances not assumed			1 832	93 959	070	723	3946	- 0606	1 5064
CPRSC	Equal vanances assumed	7 933	006	- 566	98	573	-1 427	2 5216	-6 4306	3 5774
	Equal vanances not assumed			- 571	84 518	570	-1 427	2 4995	-6 3966	3 5435
CPHS	Equal vanances assumed	6 601	012	155	98	877	068	4422	- 8092	9460
	Equal vanances not assumed			156	86 830	876	068	4387	- 8036	9404

Table 8.3 Independent samples *t*-tests on the differences in RSC and HS between *English then Chinese* and *Chinese then English*

2) Lengths of summaries

The *t*-tests identified that students produced significantly longer Chinese summaries than English summaries (mean difference=194.99, t=20.092, df=150, sig.<0.0005)[1]. Correlation between the English and Chinese summary lengths was moderate (0.416, sig.<0.0005). In addition, the English summaries written in the order of *English then Chinese* were significantly longer than those written in the order of *Chinese then English*. No such effects of language order were observed on the lengths of the Chinese summaries (Table 8.4). Separate *t*-tests for each *text type* also yielded similar results.

[1] Excluding the six outliers: ID: 4102, 4107, 4215, 3205, 4118, 3118.

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
E.S.L	.768	.382	3.059	150	.003	31.58	10.322	11.182	51.971
C.S.L	1.772	.185	-.485	153	.629	-10.52	21.706	-53 406	32.359

Note: The univariate outliers (ID: 4102, 4107, 4215, 4118, 3205 for E.S.L; ID: 4102, 3118 for C.S.L) were excluded in the analyses.

Table 8.4 Independent samples *t*-tests on the differences in the lengths of English and Chinese summaries between *English then Chinese* and *Chinese then English*

It is also noted that the mean length of the English summaries (302.48, std. deviation=64.312, std. error mean=6.057, N=151) was within the word limit (300-350) as specified in the test directions (see Appendix 3). However, students tended to write much longer Chinese summaries than required (mean=497.46, std. deviation=130.68, std. error mean=11.963), and their lengths also tended to vary to a greater extent (std. deviation=130.68, almost twice that of English summaries).

3) Summary of findings from *t*-tests

In summary, the language (English or Chinese) that students used to summarize the source texts made significant differences in the *RSC*, *HS* and *Length* of their summaries. Regardless of the assessment criteria (expert or popular templates), English summaries consistently received higher *RSC* and *HS* than Chinese summaries, though the Chinese summaries were considerably longer than the English summaries (η^2=0.729). The much-longer Chinese summaries did not seemed to have "helped" the students to receive higher RSC or HS, although common sense would dictate that they would do so because longer summaries had better chance of covering more information from the source texts and would lead to higher RSC and HS. Furthermore, it was found that the effect sizes of *language* on HS were much larger than on RSC; and larger when using popular than expert templates (see also Chapter 6).

Language order (*English then Chinese*, or *Chinese then English*) did not significantly affect the RSC and HS scores, as anticipated. However, it did have significant effects on the lengths of the English summaries. The English summaries produced in the order of *English then Chinese* were significantly longer than those written in the order of *Chinese then English* (η^2=0.0483). In other words, the

English summaries tended to be longer when they were produced *before* rather than *after* the Chinese summaries. However, the effect of *language order* on the lengths of the Chinese summaries was not significant.

These differential effects of *language* and *language order* on some aspects of summarization performance might also be related to the students' higher familiarity with English than Chinese summarization tasks (see also **8.2.1**).

8.1.2 Repeated Measures ANOVA

As demonstrated above, the *t*-tests provided a quick but incomplete picture of the effects of language and language order on students' summarization performances. However, they were not able to incorporate factors such as *text type*, *text presentation mode*, *summarization language and language order* simultaneously, all of which may have affected summarization performances to various degrees. Use of repeated measures ANOVA is of advantage in this case, as Stevens (2002: 492) points out: "*In repeated measures designs,...variability among the subjects due to individual differences is completely removed from the error term* (original emphasis)". In other words, repeated measures allow comparison of the variance caused by the independent variable to a more accurate error term which has had the variance caused by differences in individuals removed from it. Besides the increased precision, another distinct advantage of repeated measures design is economy in terms of the number of students required, although such designs potentially have serious disadvantages, in particular those caused by the order in which the treatments are administered, unless care is taken (Stevens 2002: 495). However, in this project, the order of treatments is not a major issue because (*a*) the students' summarization abilities were only measured twice (English and Chinese) and (*b*) the order of the summarization tasks (English then Chinese, Chinese then English) was randomised.

Assumptions for repeated measures analysis such as independence of the observations and multivariate normality were checked. Another key assumption of repeated measures analysis is sphericity (or circularity) which tests the null hypothesis that the error covariance between pairs of tests is equal (Stevens 2002: 501). Because there is only one pair of measures (i.e. measured only twice), sphericity is not

considered an issue in this project. The value of Mauchly's W=1, so should be the three commonly used estimates of adjustment of E (epsilon) =1 (Greenhouse-Geisser, Huynh-Feldt, and Lower-bound). No serious violation of the assumptions was noted in repeated measures analyses of all the designs (see Figure 8.1).

As demonstrated in Figure 8.1, the within-subject factor was *language* of two levels (i.e. English and Chinese) and the within-subject variables were therefore the pairs of English and Chinese RSC, HS, and Lengths of summaries. The between-subjects factors were *language order*, *text type* and/or *presentation mode* in the following four models:

- Design 1=Intercept+LANGORD+TXT+LANGORD*TXT
- Design 2=Intercept+LANGORD+PRESMODE+LANGORD*PRESMODE
- Design 3=Intercept+TXT+PRESMODE+TXT*PRESMODE
- Design 4=Intercept+ LANGORD + TXT + PRESMODE +TXT*PRESMODE +TXT*LANGORD +PRESMODE*LANGORD+TXT*PPESMODE*LANGORD

1) RSC

As in the *t*-tests above, the within-subjects factor LANG (*language*) was found to have significant main effects on RSC in all the four designs (see Figure 8.1), with effect sizes using partial η^2 ranging from 0.077 to 0.105 for RSC of expert templates and from 0.191 to 0.222 for RSC of popular templates (see Appendix 20). As indicated in the values of partial η^2, the effects of LANG were much more prominent on RSC of the popular than the expert templates. In the four designs for RSC of expert templates, the mean differences between English and Chinese summaries ranged from 3.861 to 4.290. On the other hand, the mean differences in RSC of popular templates between English and Chinese summaries ranged from 6.164 to 6.469. The between-subjects factor LANGORD (*language order*) did not have significant main effects in any of the four designs[2] [for the significant interactive effects between LANG and LANGORD on RSC of expert templates (Designs 2 and 4) for, and on RSC of popular templates (Design 2), see below].

Besides the significant main effects, LANG was also found to have significant interactive effects on RSC of expert templates (EERSC vs. CERSC) with associated

[2] Text type, another between-subjects factor in the four designs, also had significant main effects, and will be discussed in Chapter 10.

between-subjects factors. In particular,

- **With LANGORD** (Design 2: F=4.73, sig.<0.0325, partial η^2=0.047; Design 4: F=4.196, sig.<0.0435, partial η^2=0.044). The difference in RSC between the English and Chinese summaries was particularly large when the summaries were written in the order of *English then Chinese* (E/C) rather than in the order of *Chinese then English* (C/E). In addition, the English summaries produced in the order of E/C had slightly higher RSC than in the order of C/E; while the Chinese summaries produced in the order of E/C had much lower scores of RSC than those produced in the order of C/E (Figure 8.2)

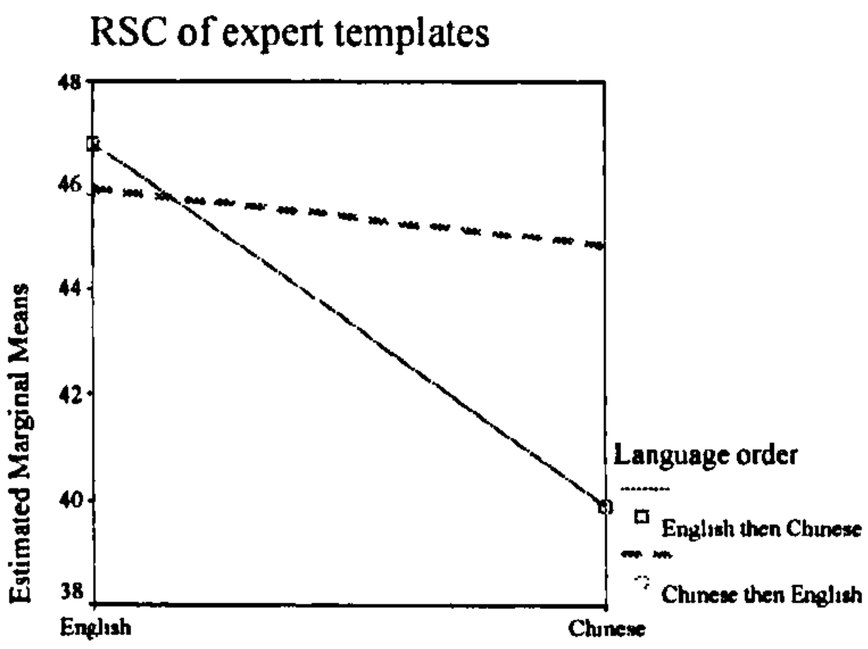


Figure 8.2 Interactive effects on RSC of expert templates between LANG and LANGORD (Design 2)

- **With PRESMODE** (Design 2: F=4.376, sig.<0.0395, partial η^2=0.044; Design 3: F=4.59, sig.<0.0355, partial η^2=0.046; Design 4: F=5.743, sig.<0.0195, partial η^2=0.059). The difference in RSC between the English and Chinese summaries was larger for summaries of paper presented texts than that of computer presented texts. In addition, the English summaries of paper presented texts received higher RSC than those of computer presented texts, while the Chinese summaries of paper presented texts received lower RSC than those of computer presented texts (Figure 8.3). See Chapter 9 for further details.

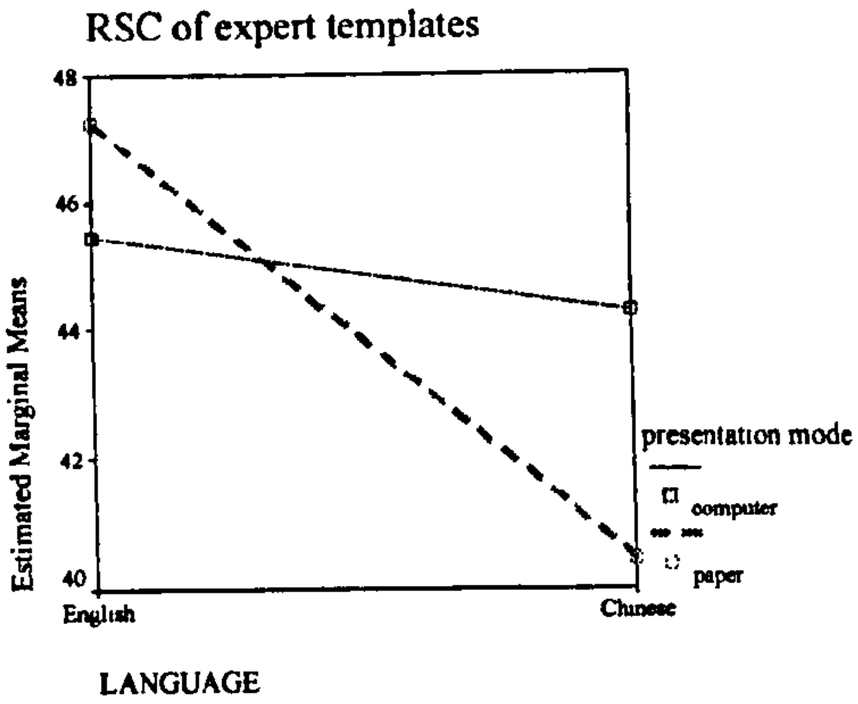


Figure 8.3 Interactive effects on RSC of expert templates between LANG and PRESMODE (Design 2)

- **With TXT*PRESMODE** (Design 3: F=5.192, sig.<0.0255, partial η^2=0.051; Design 4: F=4.875, sig.<0.0305, partial η^2=0.05). The difference in RSC between English and Chinese was particularly prominent for summaries of textC when presented on paper. The difference between the English and Chinese summaries of paper-presented textC may account for most of the difference between the two languages (Figure 8.4)

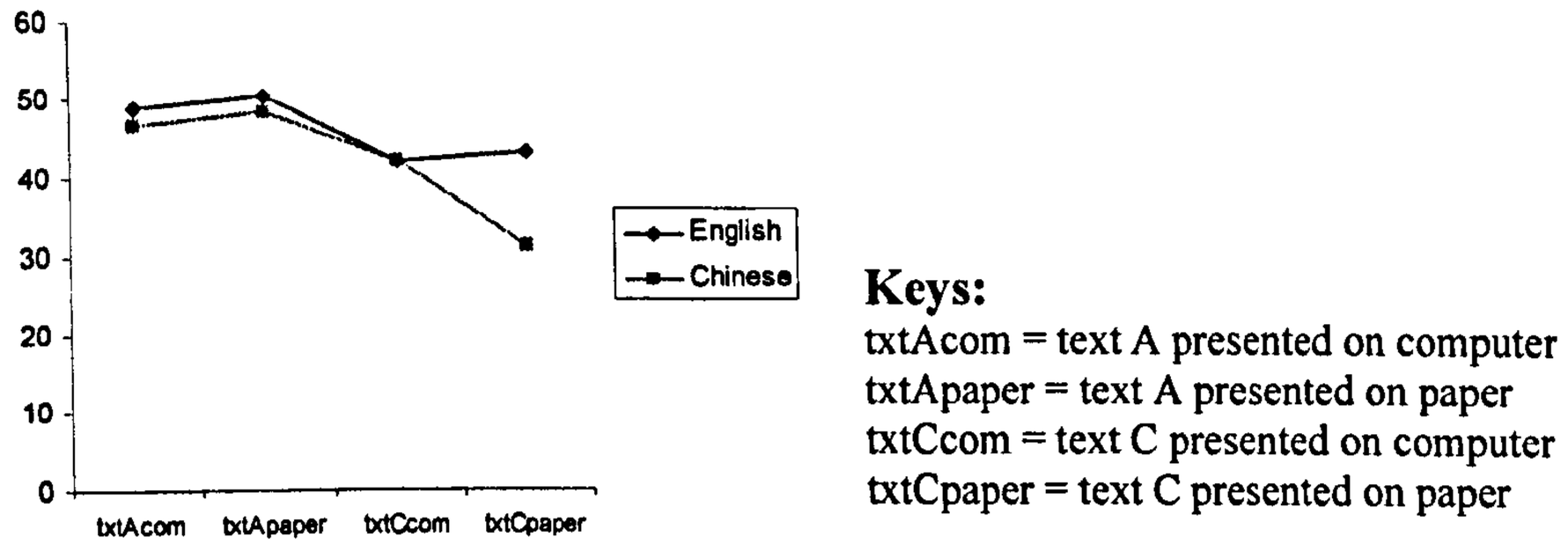


Figure 8.4 Interactive effects on RSC of expert templates between LANG and TXT*PRESMODE (Design 3)

Similarly, LANG also had significant interactive effects on RSC of popular templates with TXT and LANGORD:

- **With TXT** (Design 1: F=4.308, sig.<0.0415, partial η^2=0.043; Design 3: F=5.009, sig.<0.0285, partial η^2=0.05; Design 4: F=4.439, sig.<0.0385, partial η^2=0.046). The difference between the English and Chinese summaries of textA was smaller than that of textC summaries. Furthermore, the difference in RSC of the English summaries was smaller than that of the Chinese summaries (Figure 8.5).

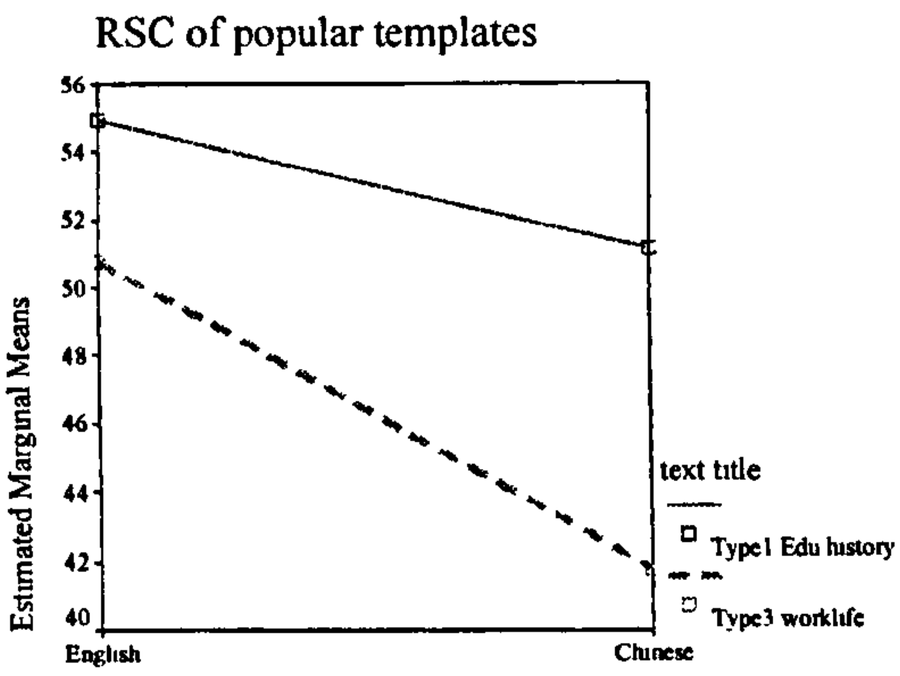


Figure 8.5 Interactive effects on RSC of popular templates between LANG and TXT (Design 1)

- **With LANGORD** (Design 2: F=4.306, sig.<0.0415, partial η^2=0.043). The difference in RSC between the English and Chinese summaries was larger when the summaries were produced in the order of *English then Chinese* than *Chinese then English*. In addition, the English summaries had higher RSC when produced in the order of *English then Chinese* than in the order of *Chinese then English*; similarly the Chinese summaries had higher RSC when produced in the order of *Chinese then English* than in the order of *English then Chinese* (Figure 8.6, also c.f. Figure 8.2).

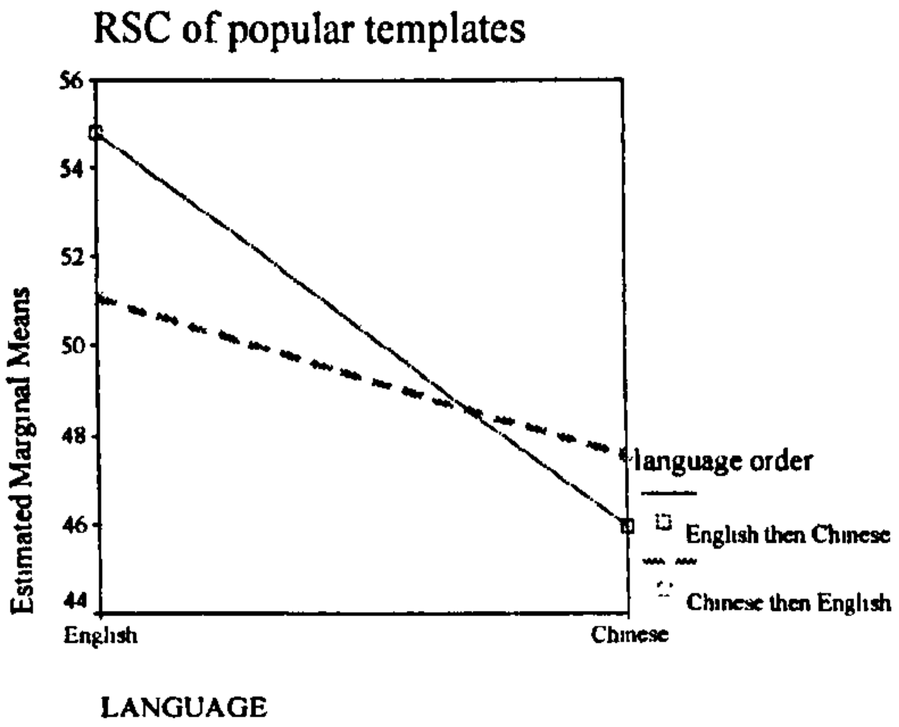


Figure 8.6 Interactive effects on RSC of popular templates between LANG and LANGORD (Design 2)

2) HS

The same procedures were undertaken to analyze HS scores using repeated measures (see Figure 8.1 and Appendix 21). In all the four designs, LANG was found to have significant main effects on HS, with partial η^2 ranging from 0.138 to 0.186. English summaries received significantly higher HS than Chinese summaries. The

effect sizes using partial η^2 were slightly larger when the summaries were evaluated according to the popular templates (from 0.17 to 0.186) than the expert templates (from 0.138 to 0.172). The only significant interactive effect of LANG was with TXT*PRESMODE on HS of expert template (Figure 8.7) in Design 3 (F=3.75, sig.<0.0265, partial η^2=0.047) and Design 4 (F=3.468, sig.<0.0345, partial η^2=0.046). As shown in Figure 8.7, the difference in HS of expert templates was much larger for paper-presented textC than the other 5 combinations. There was no such interactive effect on HS of popular templates. Interestingly, it is exactly the same phenomenon as in RSC for textC summaries (see Figure 8.4).

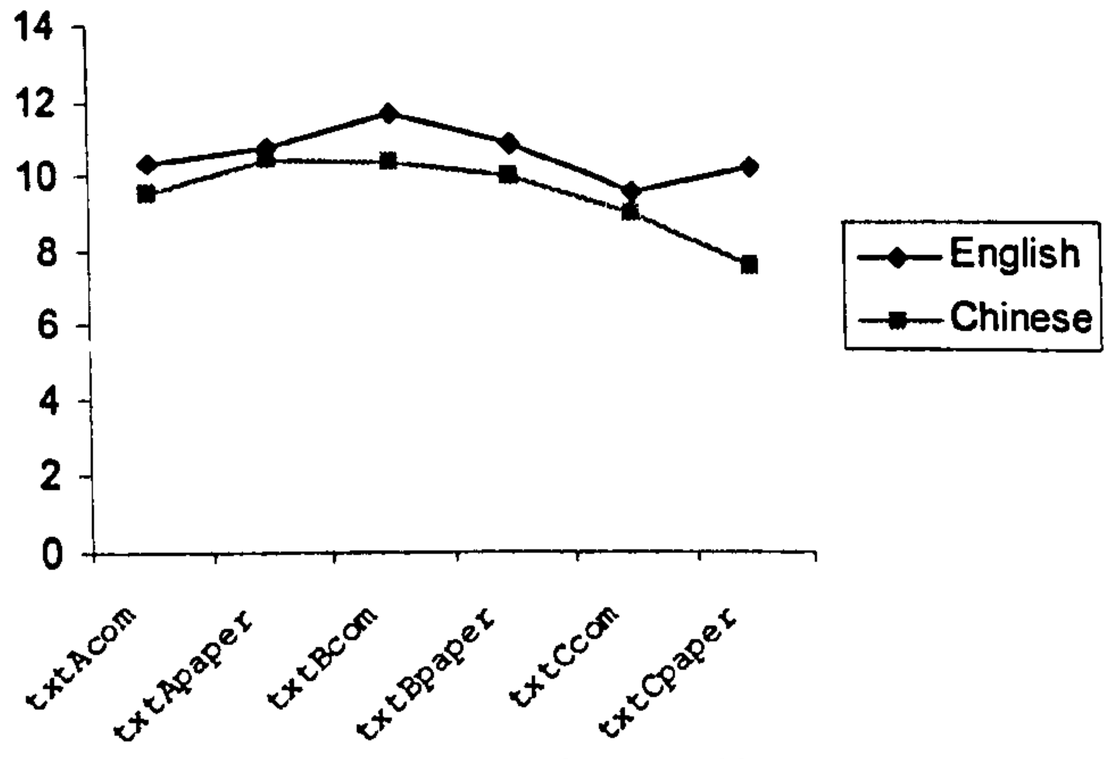


Figure 8.7 Interactive effects on HS of expert templates between LANG*TXT*PRESMODE (Design 3)

Whether the students summarized the source texts in the order of *English then Chinese* or *Chinese then English* did not have significant main effects on HS. Among all the three between-subjects factors, text type was the only one that had significant main effects on HS of both the expert and the popular templates (see Chapter 10 for further details). It also had significant interactive effects with PRESMODE*LANG (see Figure 8.7 above) on HS of expert template, and with LANGORD on HS of the popular templates in Design 1 (F=4.637, sig.<0.0345, partial η^2=0.046) and Design 4 (F=4.668, sig.<0.0335, partial η^2=0.048). The summarization language order (*English then Chinese*, *Chinese then English*) did not make much difference in the HS between the English and Chinese summaries of textA; however, for textC summaries, the difference was much larger (Figure 8.8)

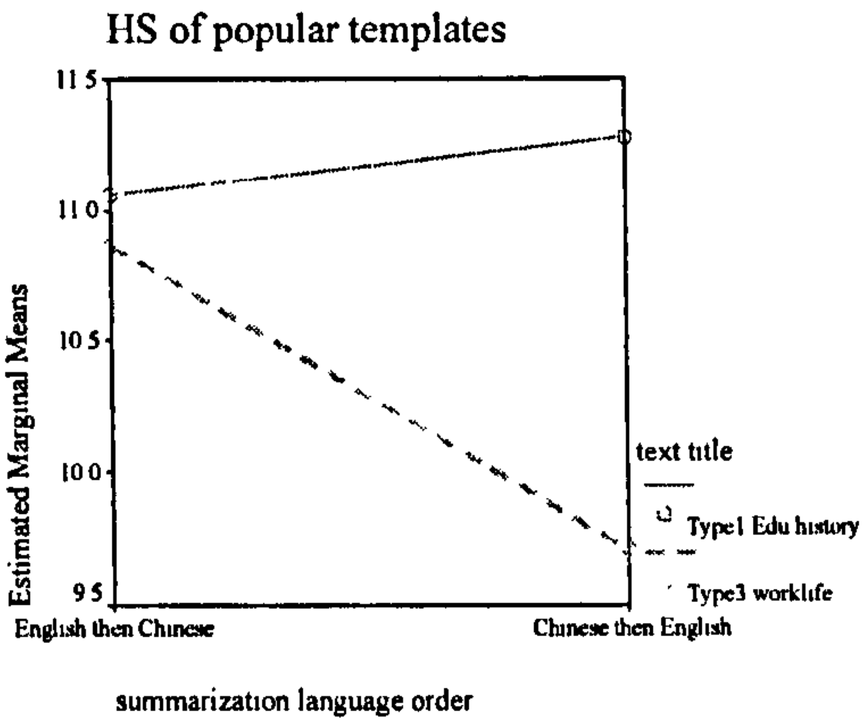


Figure 8.8 Interactive effects on HS of popular templates between TXT*LANGORD (Design 1)

3) Lengths of summaries

What language (LANG) the students used to summarize the source texts had considerable effects on the length of their summaries. The English summaries were considerably shorter than the Chinese summaries (partial η^2 ranging from 0.743 to 0.771). Besides these significant main effects, LANG was also found to have some significant interactive effects with LANGORD, TXT, LANGORD*TXT and PRESMODE respectively (see Appendix 22 for full details of the statistics).

- **With LANGORD** (Design 1: F=4.276, sig.<0.0405, partial η^2=0.029; Design 2: F=5.711, sig.<0.0185, partial η^2=0.037; Design 4: F=4.837, sig.<0.0305, partial η^2=0.034). The English summaries were longer when produced in the order of *English then Chinese* than *Chinese then English*; while the difference in the lengths of the Chinese summaries was very small between the two language orders (Figure 8.9).

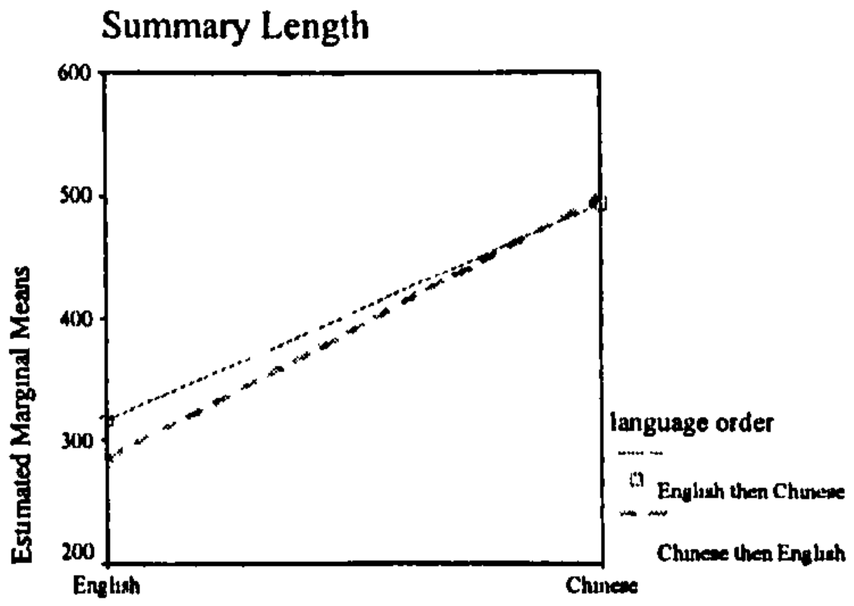


Figure 8.9 Interactive effects on the lengths of summaries between LANG and LANGORD (Design 1 as an exemplary visual representation)

- **With TXT** (Design 1: F=6.291, sig.<0.0025, partial η^2=0.08; Design 3: F=7.281, sig.<0.0015, partial η^2=0.091; Design 4: F=6.901, sig.<0.0015, partial η^2=0.090). The English summaries of textA and textB were approximately of the same length, but slightly longer than textC summaries, while the Chinese summaries of textA were much longer than those of textB and textC (Figure 8.10)

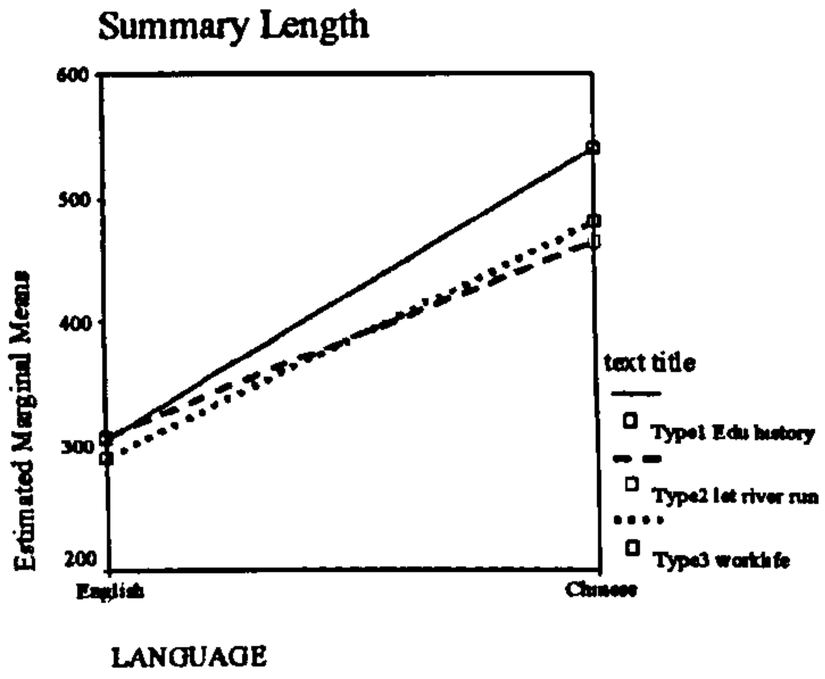


Figure 8.10 Interactive effects on the lengths of summaries between LANG and TXT (Design 1)

- **With TXT*LANGORD** (Design 1: F=3.83, sig.<0.0245, partial η^2=0.05; Design 4: F=4.012, sig.<0.0205, partial η^2=0.055). Overall, the English summaries were much shorter than the Chinese summaries of any text type. In addition, there were also significant interactions between language*text*language order. Both the English and Chinese summaries of textB and textC produced in the order of *English then Chinese* (E/C) were longer than those in the order of *Chinese then English* (C/E). However, the English summaries of textA were **longer** when produced in the order of E/C than those in the order of C/E; while the Chinese summaries of textA were **shorter** (Figure 8.11).

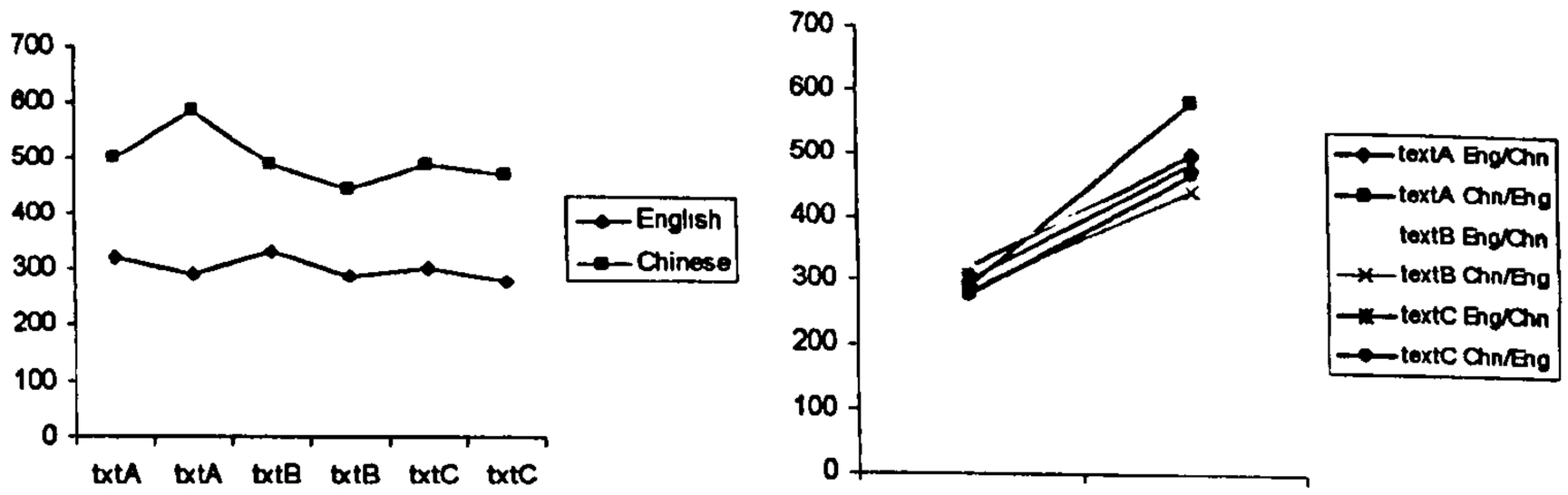


Figure 8.11 Interactive effects on the lengths of summaries between language, language order and text (Design 4)

- **With PRESMODE** (Design 2: F=7.755, sig.<0.0065, partial η^2=0.05; Design 3:

F=8.295, sig.<0.0055, partial η^2=0.054; Design 4: F=9.174, sig.<0.0035, partial η^2=0.062). The differences in length between the English and Chinese summaries were larger when the source texts were presented on computer than on paper. In addition, the difference in the length of the English summaries was smaller than that of the Chinese summaries between the two kinds of text presentation mode (Figure 8.12).

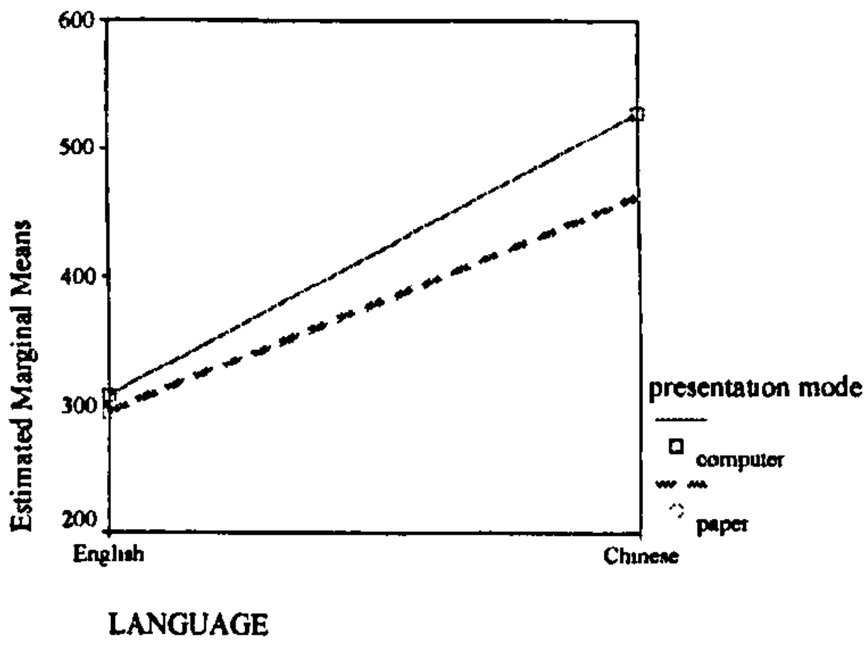


Figure 8.12 Interactive effects on the lengths of summaries between language and presentation mode (Design 2)

Apart from the interactive effects with LANG and LANG*TXT, the between-subjects factor LANGORD did not have any other significant effect on the length of summaries (see Appendix 22).

4) Summary of findings from the repeated measures analyses

The repeated measures analyses further supported the main findings from the *t*-tests (see **8.1.1**), but provided a fuller and more complex picture of the effects of *language* and *language order* on RSC, HS and Length of summaries. The English summaries consistently received higher RSC and HS, although they were considerably shorter than the Chinese summaries. This was true for both scoring templates (*expert* and *popular*), although the effect sizes were slightly larger when RSC and HS scores were assigned according to the popular templates than the expert templates (see also Chapter 6).

In addition to the significant main effects on RSC, HS and length of summaries, LANG was also found to have significant interactive effects, to a varying extent, with some of the associated between-subjects factors. In particular,

- LANG*LANGORD interactive effects.

It held true for RSC, be it assigned according to the expert or the popular templates, that the difference in RSC between the English and Chinese summaries was particularly large when the summaries were written in the order of *English then Chinese* (E/C) than when in the order of *Chinese then English* (C/E). In addition, the English summaries produced in the order of E/C had slightly higher RSC than in the order of C/E, while the Chinese summaries produced in the order of E/C had much lower RSC than in the order of C/E. This raised the question of how the initial summarization of the source English texts, be it in English or in Chinese, helped or hindered to some extent the follow-up summarization process and product (see **11.2.2** for further discussions).

The English summaries were considerably shorter than the Chinese summaries; furthermore, this difference was larger when the summaries were produced in the order of *Chinese then English* than *English then Chinese*. In addition, the English summaries were longer when produced in the order of *English then Chinese* than *Chinese then English*, while the difference in Chinese summary length was very small between the two language orders.

This kind of interactive effect was not present on HS.

- LANG was also found to have significant interactive effects with TXT (on RSC of popular templates, summary length), PRESMODE (on RSC of expert templates, summary length), TXT*PRESMODE (on RSC and HS of expert templates), and TXT*LANGORD (HS of popular templates, summary length). The difference in RSC and HS between the English and Chinese summaries of textC was particularly prominent and larger than the other source texts, especially when textC was presented on paper. This was further complicated by the interactive effects with *language*. The English summaries of paper presented texts tended to receive higher RSC scores than computer presented counterparts, while the Chinese summaries of paper presented texts tended to receive much lower RSC scores.

Apart from the significant interactive effects with *language* (LANG, see above),

language order (LANGORD) did not have any other significant effects on RSC, HS and length of summaries. Whether the students summarized the source texts in order of *English then Chinese* or *Chinese then English* did not make significant differences on RSC, HS and length of summaries, as anticipated according to the research design (see Table 4.6).

8.1.3 Multiple regressions on summarization performances and TOEFL

In Chapter 7, it was established that TOEFL-R was the best predictor (among the other language abilities) of summarization performances. Although the analyses in Chapter 7 have already touched upon which summarization performance (English or Chinese) were better able to predict TOEFL-R, the question has not been addressed *directly*. This section reports the findings from the stepwise regressions on TOEFL-R (dependent) and English and Chinese summarization performances (4 pairs of independents: EERSC/CERSC, EPRSC/CPRSC, EEHS/CEHS, EPHS/CPHS).

It was found that the Chinese summarization performances were better able to predict TOEFL-R as demonstrated in three of the four stepwise regressions. Only in the pair of EPRSC and CPRSC as independents, did EPRSC have a better chance of being retained in the regression model. In all cases, however, only a very small amount of variance in TOEFL-R could be predicted significantly (Table 8.5).

	Independent variables							
	EERSC	CERSC*	EPRSC*	CPRSC	EEHS	CEHS*	EPHS	CPHS*
R^2	.066		.086		.056		.047	
F	F (1, 98)=6.930, Sig.<.0105		F (1, 98)=9.256, sig.<.0035		F (1, 154)=9.16, sig.<..0035		F(1, 98)=4.815, sig.<.0315	
TOEFL=	33.236+ 0.0808*CERSC		30.26+ 0.121*EPRSC		31.778+ 0.474*CEHS		31.997+ 0.459*CPHS	

Note: * the independent variable kept in the regression models

Table 8.5 Stepwise regressions on TOEFL-R and the four pairs of English and Chinese summarization performances

8.2 Students' perceptions of the use of different language and language order for the summarization tasks

Both the post-summarization questionnaire (Q8a, 8b, Q9-19, see Appendix 4) and interviews addressed the same issues of the use of two languages and language orders

for the summarization tasks. They focused on (i) students' familiarity with the summarization tasks in two languages, (ii) students' reasons for the preference (or lack of preference) for a particular language and language order, and (iii) self-evaluation of the dependence of their summarization performances on language abilities such as reading, writing and translation. In particular, Q8a, 8b, 9, 14, 15, 17, 18, and 19) investigated (i) and (ii), and Q10, 11, 12, 13, and16 for (iii). The basic descriptive statistics of PSQ can be found in Appendix 23.

8.2.1 Familiarity with the summarization tasks in two languages

The students were more familiar with English than Chinese summarization tasks as indicated by the Wilcoxon signed ranks test (Z=5.455, sig.<0.0005, based on positive ranks), although there were 77 ties among the 152 valid cases (i.e. familiarity with English summarization tasks=Chinese summarization tasks in ranks). Most of the students had had such summarization tasks in their university studies (Q9: experience in such summarization tasks in university) with 26.3% (No) and 73.7% (Yes). Based on this, the initial concern about students' familiarity with the summarization tasks was allayed.

	Familiarity with English summarization tasks (N=154) %	Familiarity with Chinese summarization tasks (N=152) %
Not familiar at all	1.9	6.6
Not too familiar	34.4	52.6
Of average familiarity	35.7	32.2
Somewhat familiar	24.7	5.9
Very familiar	3.2	2.6

Table 8.6 Familiarity with English and Chinese summarization tasks

8.2.2 Preference (or lack of preference) for a particular language and language order for the summarization tasks

1) Preference for a particular language

Over half of the students (n=84, 54.2%) stated a preference for using English, 27.1% (n=42) rather than Chinese and 18.7% (n=29) did not mind which language they used (i.e. lack of preference) to summarize the English source texts (for the reasons for the preferences based on qualitative data in response to Q15, see below).

Taking into account the research design, which asked half the students to summarize in *English and then Chinese* and the other half in *Chinese and then*

English, there is a question as to whether the differences in the preferences for a particular language were influenced psychologically by the particular research design itself. The Mann-Whitney U test indicated that this could well be one of the reasons for the differences in language preference (Mann-Whitney U=2424.5, Wilcoxon W=5199.5, Z=-2.273, sig.<.0235, N=155). Those who used English first were more likely to express a preference for using Chinese first, given the choice, and vice versa. This could be further confirmed by the significantly negative correlation between the summarization language preferences and actual summarization language order (r=-0.259, sig.<.0035, N=126, i.e. excluding those who did not mind which language they used. If all the students were included, then r=-0.183, sig.<.0235, N=155).

However, I would also argue that the students' preference for a particular language may reflect their real experience rather than simply psychological effects, because (a) the questionnaire and interviews were conducted after the summarization tasks, and therefore the answers were based on their comparisons of the actual use of both languages, rather than on imagination, and (b) the counter-balanced research design (see Table 4.6) had already taken into account such potential psychological effects on the students' preferences as demonstrated in PSQ and PSI.

Subsequently, the students were asked why they preferred using a particular language in an open-ended question (Q15) in PSQ, as well as in the interviews. Since the data on this question from the PSQ and PSI were strikingly similar (see below *Preference to a particular language order*), the following report is based on the answers to Q15 on the PSQ so that views of all students instead of the 24 interviewees can be included. The reasons provided by the students for their particular language preferences shed light not only on the promises but also the problems of the use of a particular language for the summarization tasks.

a) Preference for English summarization tasks

Eighty-two of the 84 students who expressed a preference for using English answered Q15. Among the reasons given (see Appendix 24), the most frequently occurring seemed to be (*a*) the obvious "benefits" of English summarization tasks – direct/straight copying from or referring to the source texts without necessarily fully understanding the copied information or the whole text, (*b*) the additional processing

(e.g. translation) and high demands of language abilities such as translation and Chinese writing in the Chinese summarization tasks which require full understanding of the source texts and also involve serious planning to produce a "polished" summary, and (*c*) the questionable translatability between the English and the Chinese languages.

The significant benefits of direct copying from and referring to the source texts in the English summarization tasks were considered not only "convenient" and "time-saving" but also a safer route to a "better" finished product – English summaries. Because of the facility of direct copying, some students thought that it was much less likely that they would "make ambiguous statements" or "go in a wrong direction". As one student commented:

> This (English summarization) is just like squeezed juice, you can see it in its original form. It is not easy to go wrong, go in a wrong direction.

The proximity of the English summaries to the English source texts also made some students think that direct/straight copying would help them to "imitate the syntax" of the authoritative source and "guarantee correct use of grammar" in their English summaries.

The other side of the coin is that students' preference for the English summarization tasks was also significantly related to the additional load of translation in the Chinese summarization tasks, because of the necessary switching between languages from English source texts to Chinese summaries. This additional processing load made the Chinese summarization tasks more time consuming and challenging, described by one student as driving around a corner:

> In Chinese summarization, you need to think very carefully to choose the most appropriate Chinese words to express your meaning you got from an English text. It is just the same as you drive around a corner; it really takes too much time to do that.

This additional requirement also raised issues concerning translatability between the two languages and faithfulness to the source texts. Is faithful translation from English to Chinese achievable? Some students questioned this:

> It is not easy, if not impossible, to replace some English words with proper Chinese equivalents.
>
> Some proper names are simply not translatable.
>
> Translation means distortion.

In addition, the students seemed more concerned about their language in their Chinese than English summaries probably because Chinese is their first language and such a status would make them have to choose more carefully their use of Chinese words and sentence structures in fear that they might be laughed at if they did not produce a decent summary in *Chinese*.

Besides the issues of the additional translation abilities required and the translatability from English to Chinese, some students in this group were also "concerned" that the Chinese summarization tasks would require their full, rather than partial understanding of the source texts. Without the camouflage and facility of direct copying and pasting as in the English summarization tasks, there was no place in the Chinese summaries for students to hide any lack of "understanding" of the source texts (see below).

b) Preference for Chinese summarization tasks

All the 42 students who preferred the Chinese summarization tasks provided their reasons for the preference (see Appendix 25). The most frequently occurring reasons seemed to be that: (*a*) they were more familiar with the Chinese language and the Chinese summarization tasks. Therefore they felt more confident as Chinese was their mother tongue and it was at their immediate disposal; they did not have to worry about grammatical and syntactical mistakes as in the English summarization tasks because of their low English proficiency compared to Chinese; (*b*) they were able to be more concise in their Chinese summaries than in their English ones, for the same reason of facility of the mother tongue. Thirty-six of the 42 students (85.7%) gave the reasons elaborated in (*a*) and (*b*).

Some students thought their understanding of the source texts by any means involved summarization in Chinese, which made Chinese summarization a "natural" ingredient of the whole process of comprehension, and therefore easier and more straightforward or direct than the English summarization tasks (Major Theme C, see Appendix 25). Interestingly, those who preferred the English summarization tasks also valued significantly the straightforwardness of their preferred tasks, but from a quite different perspective – the direct/straight copying from and referring to the *source* texts. It seemed that the students who expressed a preference for the English

summarization tasks were focusing on the straightforwardness from a surface and tangible level, while the Chinese preference group were operating on a deeper and less tangible level.

c) No particular preference

Twenty-one of the 29 students who did not mind which language they used answered Q15 (see Appendix 26). The most frequently occurring reason seemed to be that understanding was considered the primary prerequisite for successful summarization, be it in English or in Chinese. Language was only a means of conveying the summarizers' comprehension of the main ideas of a source text. In addition, some students were also well aware of the disadvantages and advantages of either language. A balanced view was established among these students. One student also pointed out that she had not developed any preference yet, because it was the first time that she had done both English and Chinese summarization tasks at the same time in a formal test context.

2) Preference for a particular language order

Q17 served only to double check the actual *language order* in which the students did the summarization tasks (see Chapter 4). As shown in Q18, the majority of the students (n=83, 54.2% for E/C; n=36, 23.5% for C/E, n=34, 22.2% for "not mind") stated that they would like to summarize first in *English then Chinese* (χ^2=30.157, df=2, sig.<0.0005). The language order in which the students actually summarized the source texts did not affect their preference (Mann-Whitney U=1656, Wilcoxon W=3801, Z=-0.664, n.s., N=119 excluding those who did not mind). This finding was further confirmed in the correlation between the preferred and the actual summarization language order (Spearman rho=0.061, n.s., N=119).

The preferences for a particular language and language order were significantly correlated, though only with a small magnitude (Spearman rho=0.215, sig.<0.0085, N=152). As can be seen in Table 8.7 below, the majority of those who preferred to use English (59/83) had a strong preference for summarizing the source texts in the order of English then Chinese (59/83), while those who preferred to use Chinese were far less concerned about the *language order* (16/41 for *English then Chinese*, 17/41 for *Chinese then English*).

summarization language preference * language order preference Crosstabulation

Count

		summarization language order preferred			
		i don't mind	English then Chinese	Chinese then English	Total
summarization language preferred	i don't mind	13	8	7	28
	English	12	59	12	83
	Chinese	8	16	17	41
Total		33	83	36	152

Table 8.7 Cross-tabulation of preferences for language and language order

Q19 aimed to understand the reasons why the students preferred a particular language order (i.e., *English then Chinese*, *Chinese then English*, and *Do not mind*). Most of the reasons given were strikingly similar to those for the *Preference for a particular language*. In fact, the majority (55.6%) of the students (N=153 excluding 4 missing cases) simply said that the reasons were the same as in Q15. The following analyses are based on the additional answers that the remaining students (n=68) provided for Q19 (see Appendix 27).

The natural order of the summarization process of an English source text was from English to Chinese, as some students claimed (see Major Theme A in Appendix 27.a)[3]. It was also in the same order (from English to Chinese) that this group of students considered it easier to translate the summaries between the two languages, as they held that Chinese summarization ultimately involved translation either directly from the English summaries written immediately earlier (not physically available to them, but already stored in their mind) or from the key original information from the source texts (see Major Theme B in Appendix 27.a). Summarization in *English then Chinese* was considered not only a direct and natural processing order but also a friendly, facilitative and step-by-step approach to the subsequent Chinese summarization tasks.

> A text of 6 pages long is difficult to understand, we need to first of all list the key or important English sentences from the source text and then summarize them in English and then in Chinese.

[3] See also the counter-argument some students gave that the natural order of summarizing an English source text had already inherently involved comprehending and summarizing it in Chinese while they were processing the source text (see *Preference for Chinese summarization tasks*, Major Theme C of Appendix 24). This to some extent also reflects the role of individual factors in the summarization process (see **2.5.4**).

> If I had summarized it first in Chinese, the English summary would be confined by the Chinese summary and would not look like English.

The other side of the coin is that those who preferred the order of *Chinese then English* reasoned using the same kind of logic but from a different perspective. They argued that: (*a*) Chinese students were more or less thinking in "a Chinese way", as one student recalled "*very often, unconsciously, I translate an English text into Chinese when I am reading*", and (*b*) Chinese summarization not only promoted better understanding of the English source texts, but also made the subsequent English summarization tasks much easier, for example, by providing "*a helpful structure for English summarization later*", and being able to find "*the sentences needed for the English summary in the source text, quickly and easily*". This to some extent also implied that both the pro-E/C and pro-C/E students agreed that the Chinese summarization tasks were more challenging than the English ones.

Those who *did not mind* thought understanding the source text was far more important than the choice of language order or language. No matter which order, "the first summarization must be helpful for the second summarization task!" as two students asserted (see Appendix 27 (c) *Reasons for "do not mind"*).

8.2.3 Evaluation of the relationship between summarization performances and other language abilities

Questions 10, 11-13 and 16 asked the students to evaluate (**1**) whether and to what extent their summarization performances depended on their language abilities such as English reading and writing, Chinese writing, and translation from English to Chinese, and (**2**) which summarization task (English or Chinese) would provide a better measure of their English reading comprehension abilities.

1) Dependence of summarization performances on other language abilities

a) English summarization performances

Two questions in the PSQ (see Appendix 4) examined perceptions of whether and to what extent English summarization performance depended on students' *English reading* (Q10a) and *English writing* abilities (Q10b).

Wilcoxon signed ranks test (Z=-6.06, sig.<.0005, based on the positive ranks) indicated significantly different distributions of students' answers to these two questions. English reading abilities were considered to play a more crucial role than English writing abilities for a successful English summarization. Around 28% of them thought English summarization was "highly dependent" and 52% "fairly dependent" on English reading abilities; while only 14% of them though it was "highly dependent" and 37% "fairly dependent" on English writing abilities (Table 8.8)

English summarization performance is	10a: English reading abilities (%)	10b: English writing abilities (%)
Highly independent	0.0	1 (0.6)
Fairly independent	4 (2.6)	11 (7.1)
Moderately (in)dependent	27 (17.3)	65 (41.7)
Fairly dependent	81 (51.9)	57 (36.5)
Highly dependent	44 (28.2)	22 (14.1)
Total	156 (100)	156 (100)

Table 8.8 Dependence of English summarization performances on English reading and writing abilities

This finding was further confirmed by Q11 which asked the students to make a further distinction on the contribution of their English reading and writing abilities to their English summarization performances. Only 15% of the students thought successful English summarization depended most on their English writing abilities, and 46.4% thought it was their English reading abilities that were more important for a successful English summarization, and 38.6% thought English reading and writing abilities were equally important (or not important) for a successful English summarization. Chi-square test on the distributions of the answers to Q11 demonstrated these differences were statistically significant (χ^2=24.471, df=2, sig.<0.0005, N=153).

b) Chinese summarization performances

Four questions were designed to seek students' views about the dependence of Chinese summarization performances on their abilities in English reading (Q12a), Chinese writing (12b), and English to Chinese translation (12c), and their further distinction on which ability contributed most to the Chinese summarization performances (Q13).

Kendall's W test indicated that there were significant differences in the distributions of students' answers to these three questions (mean rank=2.26 for Q12a,

1.78 for Q12b, and 1.96 for Q12c; Kendall's W, i.e. Kendall's coefficient of concordance=0.108; χ^2=33.207, df=2, sig.<0.0005, N=154). Chinese summarization performances were considered to be more dependent on *English reading* than *translation* and *Chinese writing* abilities. This was considered "highly dependent" on English reading by 31.6% of the students, on Chinese writing by 15.4% and on translation by 20.1% of them (Table 8.9).

	Q12a: English reading (%)	Q12b: Chinese writing (%)	Q12c: E/C translation (%)
Highly independent	2 (1.3)	2 (1.3)	5 (3.2)
Fairly independent	3 (1.9)	7 (4.5)	6 (3.9)
Moderately (in)dependent	26 (16.8)	60 (38.5)	44 (28.6)
Fairly dependent	75 (48.4)	63 (40.4)	68 (44.2)
Highly dependent	49 (31.6)	24 (15.4)	31 (20.1)
Total	155 (100)	156 (100)	154 (100)

Table 8.9 Dependence of Chinese summarization performances on abilities in English reading, Chinese writing and English to Chinese translation

This statistically significant difference was further attested to in the data from Q13. More students (39.6%) thought a successful Chinese summarization depended most on English reading abilities than the other two abilities. However, 39% of the students thought Chinese summarization depended most on their translation abilities (Table 8.10).

Chinese summarization performance depends most on:		
	Frequency	Percentage
English reading	61	39.6
Chinese writing	33	21.4
Translation (English to Chinese)	60	39.0
Total	154	100.0

Table 8.10 Dependence of Chinese summarization performance on language abilities – a further distinction

The Chi-square test on data from Q13 demonstrated that there was statistically significant difference in students' self evaluation of the contribution of the three language abilities to their Chinese summarization performances (χ^2=9.831, df=2, sig.<0.0075).

2) Which task provides a better measure of English reading comprehension abilities

Q16 elicited students' views on which summarization task would provide a better measure of their reading comprehension abilities. The Chinese summarization task was favoured by 46.4% of the students (N=153), 26.8% opting for the English

summarization tasks. The other 26.8% thought the two tasks measured their English reading comprehension equally well. A Chi-square test indicated a statistically significant difference between these three categories (χ^2=11.765, df=2, sig.<0.0035), with the Chinese summarization task considered a better measure of their reading comprehension abilities. Students' answers to Q15 and Q19 (see Appendices 24-27) also indirectly demonstrated that the Chinese summarization tasks were considered better able to measure students' reading comprehension abilities.

8.3 Summary of findings relating to RQ3

This research question aimed to understand (**a**) the effects of language and language order on summarization performances and (**b**) which activity, English or Chinese summarization, better reflected students' English reading comprehension abilities, through analysing the data of their actual performances and perceptions on such effects.

With regard to (**a**), it was found that the English summaries received significantly higher scores (RSC and HS) than the Chinese summaries, although the Chinese summaries were substantially longer than the English summaries. Besides these significant main effects on the differences between the English and Chinese summaries, LANG was also found to have exerted significant interactive effects (see 8.1.2):

- with LANGORD on RSC, regardless of scoring templates;
- with TXT on RSC of popular templates and lengths of summaries;
- with PRESMODE on RSC of expert templates and lengths of summaries;
- with TXT*PRESMODE on RSC and HS of expert templates;
- with TXT*LANGORD on HS of popular templates and lengths of summaries.

As anticipated (see Table 4.6 in Chapter 4), apart from the interactive effects mentioned above, LANGORD did not have any other significant effects on students' summarization performances as demonstrated in the analyses in this chapter (but see Chapter 9 for the significant interaction effects of LANGORD with *computer familiarity* on EERSC and EEHS).

The majority of the students preferred the English summarization task because of

(i) its obvious "benefits" of direct copying from the source without necessarily fully understanding the copied information, (ii) the extra processing load of the Chinese summarization which was considered to involve full understanding of the source and (iii) the questionable translatability from English to Chinese. The pro-Chinese students felt they were more confident in using Chinese and could better produce concise summaries because of the facility of the mother tongue. They also thought it was a natural process because understanding the source texts already involved Chinese summarization while reading the texts. The non-preference group emphasized that language was only a means. It was the understanding of the source texts that was the prerequisite for successful summarization, be it in English or in Chinese. The preferences for a particular language order were reasoned using quite similar logic.

With regard to **(b)**, the stepwise regressions demonstrated that Chinese summarization performance was better able to predict TOEFL-R than English summarization, though only a small amount (see also **7.3**). This finding was in line with the students' view that their English reading ability was the most influential predictor for both English and Chinese summarization performances. However, students also seemed to suggest that the significant relationships between summarization performances and reading abilities were much stronger than the regression analyses demonstrated in this and the previous chapter.

Further discussion of these findings is reported in **11.2.2.3**.

CHAPTER NINE

Text Presentation Modes and Computer Familiarity

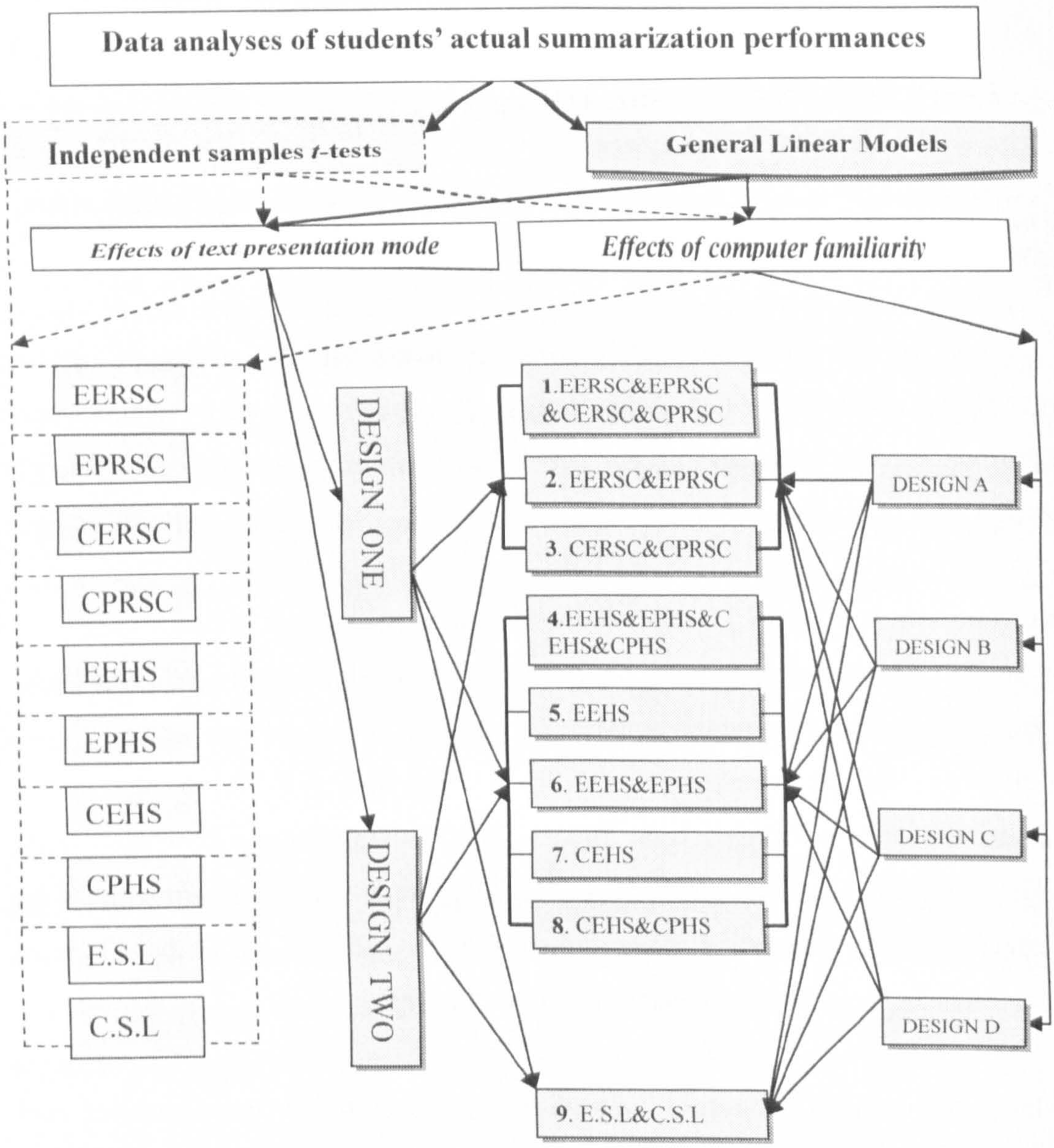
What are the effects of text presentation mode and students' computer familiarity on their summarization performances?

This research question (RQ4) aimed to investigate the effects of (*a*) text presentation mode and *b*) students' computer familiarity on their summarization performances, through analysing not only their *actual* summarization performances but also their *perceptions* of such effects.

As can be seen in Figure 9.1, the *performance* data were analysed in two phases:

- a series of independent samples *t*-tests on each individual quality indicator (RSC, HS, and Lengths) of summaries and
- univariate and multivariate general linear modelling (GLM) on the quality indicators, incorporating several factors such as *text type* and *language order*. There were six GLM designs in all: (**i**) Designs One & Two were to investigate the effects of text presentation mode and (**ii**) Designs A, B, C, & D the effects of computer familiarity on summarization performances.

Students' perceptions, collected through post-summarization questionnaires and interviews, were analysed in both quantitative and qualitative methods.



GLM Designs:
Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT
Design Two: Intercept+PRESMODE+LANGORD+PRESMODE*LANGORD
Design A: Intercept+TXT+COMPFAM LEVEL+TXT*COMPFAM LEVEL
Design B: Intercept+LANGORD+COMPFAM LEVEL+LANGORD*COMPFAM LEVEL
Design C: Intercept+TXT+COMPFAM SCORE
Design D: Intercept+LANGORD+COMPFAM SCORE
Colour scheme:
Independent samples t-tests
General Linear Models
Dash style:
- - - : Independent samples t-tests, _____: General Linear Models

Figure 9.1 Plan for the statistical analyses on the effects of text presentation mode and students' computer familiarity on their summarisation performances

9.1 Students' actual summarization performances

9.1.1 Independent samples t-tests

The independent samples *t*-tests found that none of the differences in RSC or HS between computer and paper presentation mode was significant (see Appendix 28). The only significant mean difference existed in the lengths of the Chinese summaries. The Chinese summaries of computer presented texts were significantly longer than those of paper presented texts (mean difference=47.76, t=2.014, df=155, sig.<0.0465).

Similarly, the effects of computer familiarity[1] were also analysed through independent samples *t*-tests (see Appendix 29). It was found that *low computer familiarity* students had significantly higher CERSC, CPRSC and CPHS than *high computer familiarity* students (Table 9.1). There was no other significant difference in summarization performances between low and high computer familiarity students.

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
CERSC	2.099	.154	2.587	50	.013	9.302	3.5959	2.0793	16.5244
CPRSC	1.526	.222	2.180	50	.034	7.617	3.4940	.5988	14.6347
CPHS	.013	.911	2.293	50	.026	1.363	.5941	.1692	2.5558

Note: There was no significant difference in CEHS between low and high computer familiarity students if data from all the three source texts were included, however, there was a significant difference (F=3.348, df=50, sig.<.0025) if only the summaries of textA and textC were included (see 9.1.2 and Chapter 10 for the effects of text type on summarization performances).

Table 9.1 Independent samples t-tests on the effects of computer familiarity

In summary, the independent samples *t*-tests indicated that the *Chinese* summaries of computer presented texts were significantly longer than those of paper presented texts, and that the significant means differences between Low and High computer familiarity students were only in CERSC, CPRSC and CPHS of *Chinese* summaries too. No other difference was significant.

---

[1] Only including about half of the students, i.e. those who summarized the computer-mediated source texts. Due to the small sample size, the students were only categorized in two groups of computer familiarity (High, Low).

9.1.2 Univariate and multivariate GLM

Although the *t*-tests above are quick and straightforward, the results obtained may not present a full picture of the effects under investigation. They did not take into account simultaneously factors such as different *text type* and *language order* and the interaction between *text presentation mode* and *computer familiarity*. A series of separate independent samples *t*-tests also incur an increased risk of type I error (see Stevens 2002: 174-175 for some conceptual and statistical reasons why using multivariate analyses are desirable and preferred when comparing treatments). The univariate[2] and multivariate analyses (see Figure 9.1) in this section rectify some deficits of the separate independent samples *t*-tests to draw a fuller picture of such effects.

A series of essential assumptions of multivariate analyses were checked, following the advice of Stevens (2002: 256-284) and Tabachnick and Fidell (2003) in relation to:

- sample size,
- univariate and multivariate normality and outliers, by conducting the Kolmogorov-Smirnov test and evaluating Mahalanobis distance values against the critical values[3]
- linearity (i.e., the linearity between each pair of the dependent variables, using a scatterplot),
- multicollinearity and singularity (by using correlation, condition index, and collinearity diagnostics statistic), and
- homogeneity of variance-covariance matrices (by evaluating Box's M test of equality of covariance matrices).

No serious violations were noted in most of the models (otherwise the results are not reported, see Appendix 30).

1) Effects of text presentation mode

The effects of *text presentation mode* on summarization performances were examined through Designs One and Two, with the nine groups of dependent variables (see Figure 9.1). It was found that *text presentation mode* did not have significant

---

[2] **a)** Apart from multivariate analyses with other scores of a summary, EEHS and CEHS were also subjected to univariate GLM, so that data from textB summaries which did not have RSC, EPHS or CPHS could also be used to provide more insights on effects of text presentation mode and computer familiarity.
b) These two univariate GLMs (EEHS, CEHS) incorporated both two-way and one-way covariate designs, and therefore were essentially different from the independent samples *t*-tests in the previous section.

[3] If the number of dependent variables is 2, the critical value is 13.82, if 3, then 16.27, if 4, then 18.47 [Source: Tabachnick and Fidell (2003); originally from Pearson, E. S. and Hartley, H. O. (eds) (1958). *Biometrika tables for statisticians* (vol.1, 2nd edition). New York: Cambridge University Press].

main effects on the RSC or HS of summaries (see Appendix 30 for the full report of the statistics); however it was also noticed that *presentation mode* had some significant interactive effects with *text type* on:

- **CERSC** (F=5.32, sig.<0.0235, partial η^2=0.053) when CERSC and CPRSC were dependent variables (i.e. No.3 in Figure 9.1). Summaries of textA when presented on computer screen received slightly lower CERSC than summaries of the same source text when presented on paper. However, for textC, summaries of computer-presented source texts received much higher CERSC than summaries of paper presented mode (Figure 9.2).

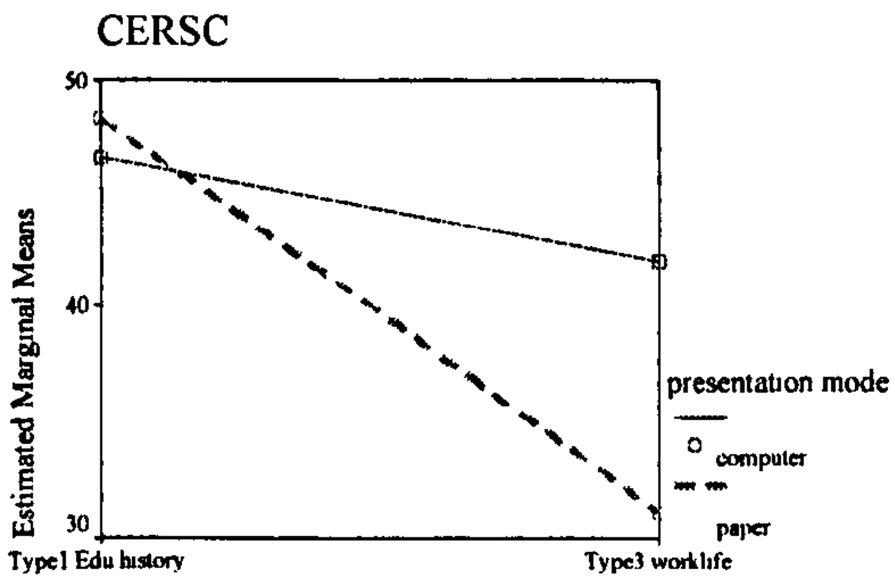


Figure 9.2 Interactive effects on CERSC of text presentation mode and text type

- **CEHS** (F=3.37, sig.<.0375, partial η^2=.043) when it was the dependent variable (No. 5 in Figure 9.1). Inspections of the estimated means found that the difference between the two presentation modes was larger in textC than in the other two source texts (textC>textA>textB), as shown in Figure 9.3.

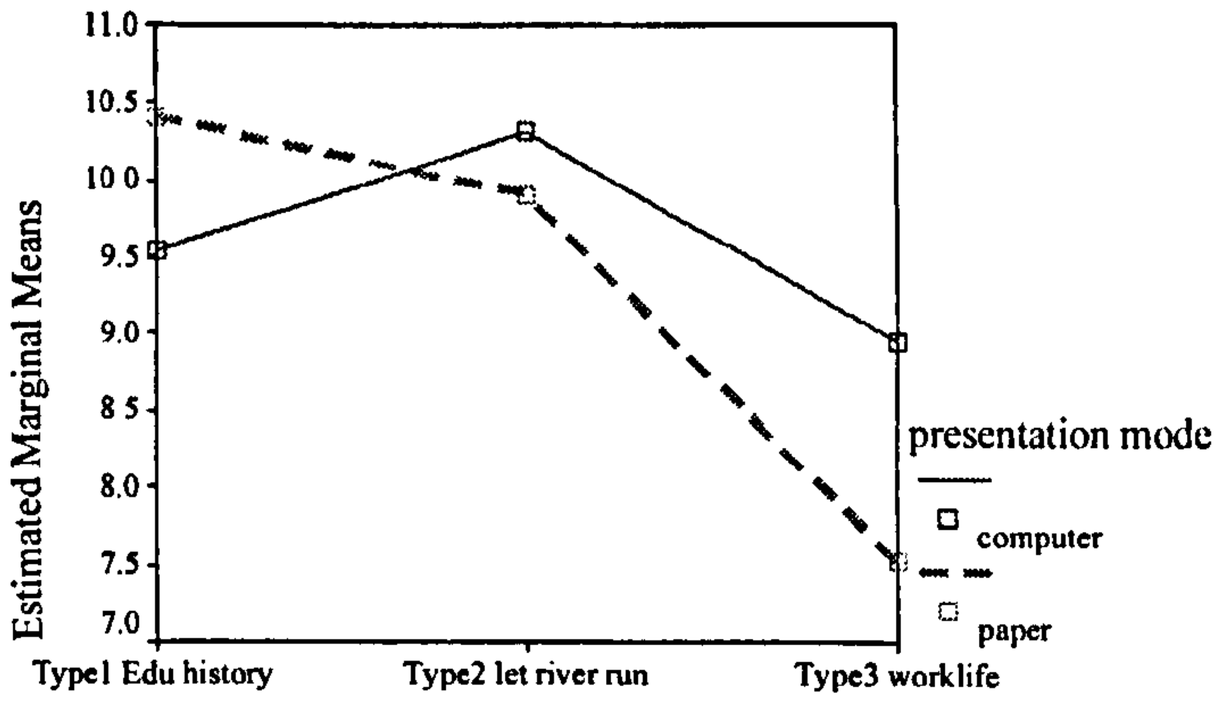


Figure 9.3 Interactive effects on CEHS of text presentation mode and text type

Although *text presentation mode* did not have significant main effects on RSC or HS, its main effects on the lengths of summaries were significant in both designs (Design One: F=5.684, sig.<.0045, partial η^2=.073; Design Two: F=4.863, sig.<.0095,

partial η^2=.062). When the results for the English and Chinese summary lengths were considered separately, it was found that the significant main effect of *presentation mode* was on the lengths of the *Chinese* summaries only (**Design One**: F=11.447, sig.<0.0015, partial η^2=0.073; **Design Two**: F=9.779, sig.<0.0025, partial η^2=0.062). The estimated means differences indicated that the Chinese summaries of computer presented texts were significantly longer (**Design One**: mean difference=67.411, std. error=19.924, sig.<0.0015; **Design Two**: mean difference=65.149, std. error=20.833, sig.<0.0025). Such effects were not evident on the lengths of the English summaries.

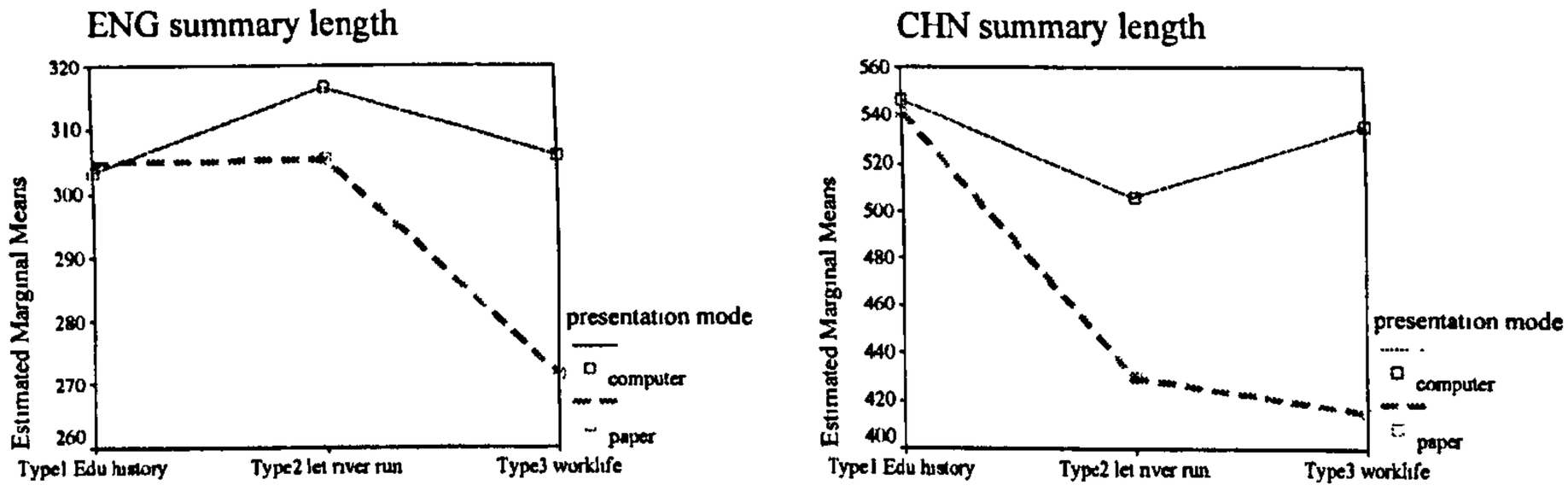


Figure 9.4 Effects of text presentation mode on the lengths of the English and Chinese summaries

In summary, the multivariate analyses further confirmed the findings from the independent samples *t*-tests (**9.1.1**), namely that the only significant main effect of *text presentation mode* was on the lengths of the Chinese summaries. The Chinese summaries of computer presented source texts were significantly longer than those of paper presented texts (see also Chapter 8). In addition, the multivariate analyses provided fuller picture of such effects. It was found that *text presentation mode* also had significant interactive effects with *text type* on CERSC and CEHS. In both scores, the differences between the two presentation modes were notably larger for textC summaries.

2) Effects of computer familiarity

In order to draw a fuller picture of the effects of computer familiarity on summarization performances, several multivariate analyses using Designs A, B, C, D were conducted (see Figure 9.1). It was found that computer familiarity did not have any significant effect on the lengths of summaries, but had some significant effects on RSC and HS to varying degrees of magnitude (see Appendix 31 for the full report of the statistics). The following section reports only such significant effects.

a) RSC

i) RSC of Chinese summarization performances

The only significant main effect of computer familiarity was in **Design B** (i.e., with *language order* as another between-subjects factor) where CERSC and CPRSC were the dependent variables. The multivariate statistics indicated that *computer familiarity level* had a statistically significant main effect (F=3.971, sig.<.0255, partial η^2=.145). When the results were considered separately for CERSC and CPRSC, it was found that both were significantly affected by computer familiarity, though with a different magnitude (F=7.373, sig.<.0095, partial η^2=.133 for CERSC; and F=5.125, sig.<.0285, partial η^2=.096 for CPRSC). *Low computer familiarity* students had significantly higher CERSC and CPRSC than their *high computer familiarity* counterparts (mean difference=9.535 for CERSC, and 7.872 for CPRSC). The interaction effect of *language order* and *computer familiarity level* (Figure 9.5) was also significant on CERSC (F=4.325, sig.<0.0435, partial η^2=0.083), though not on CPRSC (F=2.074, n.s).

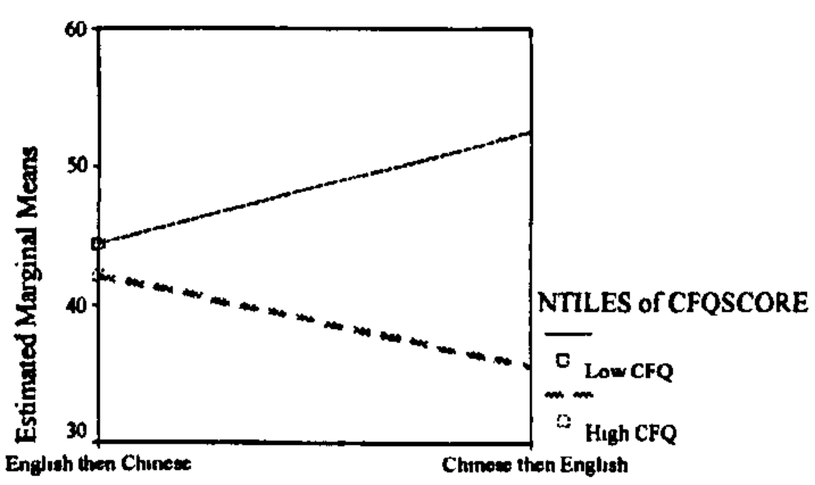


Figure 9.5 Interactive effects on CERSC of computer familiarity and language order

In **Design D**, the multivariate statistics indicated that the effect of computer familiarity (raw score as a covariate) was approaching significance level (F=2.867, sig.<0.0675). When the results were considered separately for the two dependent variables CERSC and CPRSC, it was found that the computer familiarity score also had significant effects on both of them (F=4.050, sig.<0.0505, partial η^2=0.076 for CERSC; F=5.233, sig.<0.0275, partial η^2=0.096 for CPRSC).

In **Design A**, although the multivariate statistics indicated no significant main or interactive effects on the composite of CERSC and CPRSC, the univariate statistics indicated that CERSC was significantly affected by the students' computer familiarity

(F=5.013, sig.<.0305, partial η^2=.095). *High computer familiarity* students had significantly lower CERSC (mean difference=-8.566) than *low computer familiarity* students.

ii) RSC of English summarization performances

Only in **Design B** and when EERSC and EPRSC were dependent variables did the multivariate statistics demonstrate that there was a significant interaction effect between *computer familiarity* and *language order* on the composite of EERSC and EPRSC (F=3.853, sig.<.0285, partial η^2=.141). Inspection of the results indicated that this significant interaction effect was only on EERSC (F=6.114, sig.<.0175, partial η^2=.113), rather than on EPRSC (F=0.178, n.s.). The English summaries written in the order of *English then Chinese* by low computer familiarity students had lower EERSC than those by high computer familiarity students. On the other hand, the English summaries written in the order of *Chinese then English* by low computer familiarity students had higher EERSC than those by high computer familiarity students (Figure 9.6).

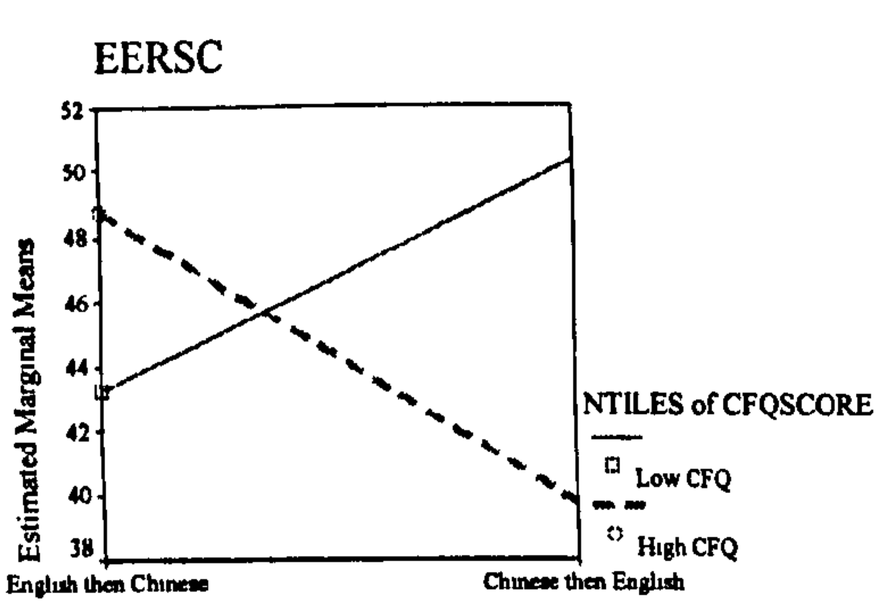


Figure 9.6 Interactive effects on EERSC of computer familiarity and language order

In summary, students' Chinese summarization performances, in terms of their RSC scores, were more likely to be affected by their computer familiarity than English summarization performances. Low computer familiarity students had significantly higher CERSC and CPRSC (i.e. regardless of which scoring template was used). These findings were in line with the results from the independent samples *t*-tests (see **9.1.1**). Furthermore, the multivariate analyses also indicated that there were significant interaction effects between computer familiarity and language order on CERSC and EERSC (the only significant effect of computer familiarity on RSC of the English summaries).

b) HS

The multivariate statistics indicated that computer familiarity had significant main effects on the composite of the four HS scores in both Designs A and B (F=2.842, sig.<.0355, partial η^2=.202 in Design A; F=3.174, sig.<.0225, partial η^2=.22 in Design B). Low computer familiarity students had significantly higher HS as a composite.

When EEHS and EPHS (i.e. HS of English summaries) were the dependent variables, there was no significant main effect of computer familiarity. However, it had a significant interaction effect with *language order* (F=3.603, sig.<0.035, partial η^2=.133) in Design B; this interaction effect was largely due to the significant difference in EEHS (F=4.182, sig.<0.0465, partial η^2=0.08), not in EPHS (F=0.156, n.s.) between the low and high computer familiarity students (Figure 9.7). Low computer familiarity students had lower EEHS when summarizing in the order of *English then Chinese*, but higher EEHS when in the order of *Chinese then English*.

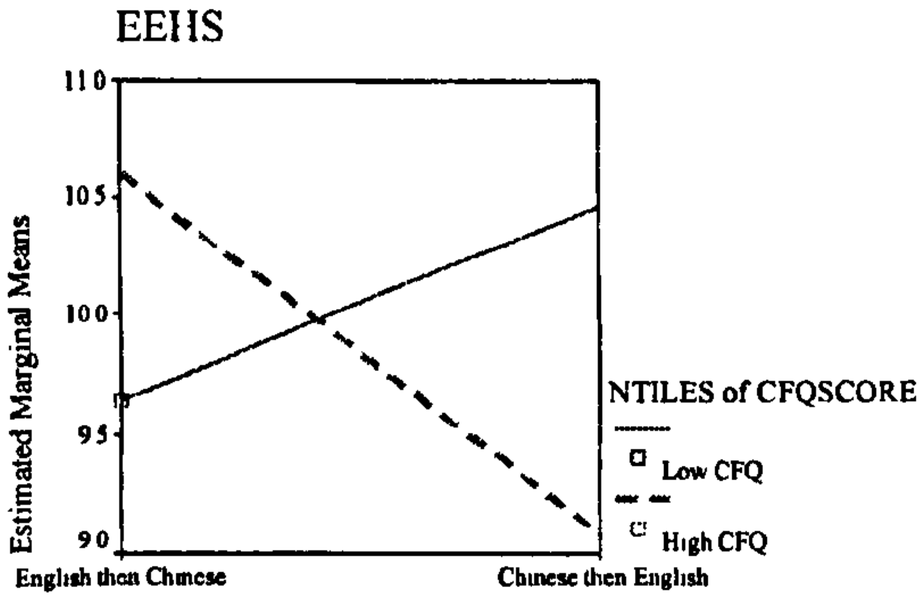


Figure 9.7 Interactive effects on EEHS of computer familiarity and language order

Nor was any significant main effect of computer familiarity found when CEHS was the dependent variable. However, a significant interaction effect was noted in Design A between *text type* and *computer familiarity* (F=3.922, sig.<0.0245, partial η^2=0.094). Low computer familiarity students had higher CEHS if they summarized textA or textC, but lower CEHS if they summarized textB, than their high computer familiarity counterparts (Figure 9.8). However, the difference in CEHS between low and high computer familiarity students of textB was quite small, compared to the differences among the students of textA and textC.

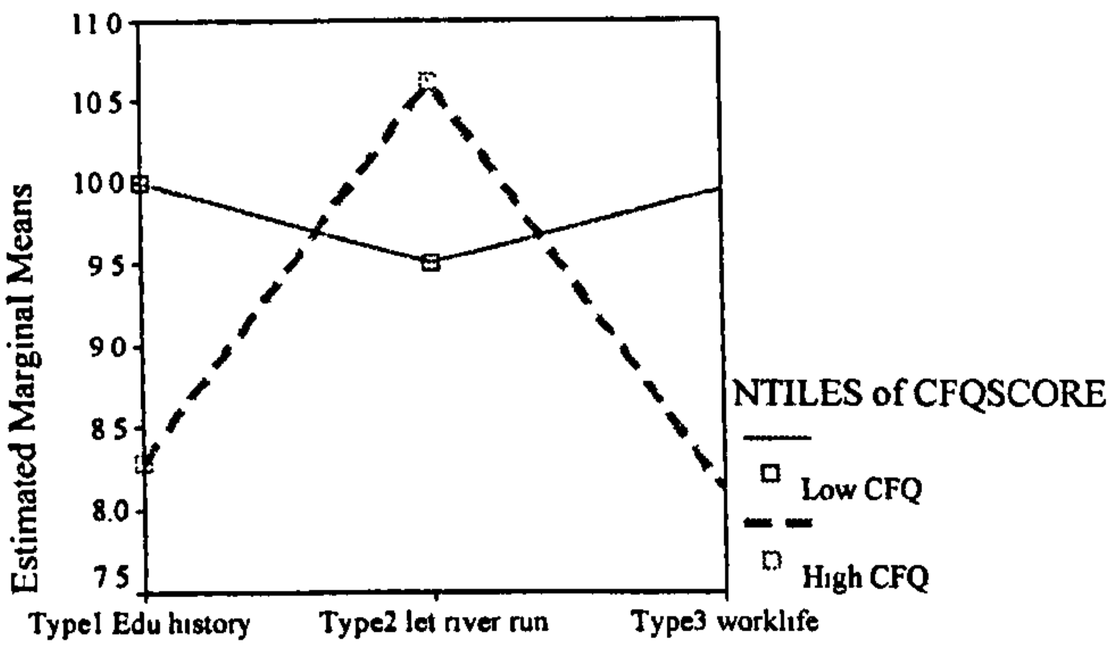


Figure 9.8 Interactive effects on CEHS of computer familiarity and text type

When CEHS and CPHS were the dependent variables (i.e. HS of Chinese summaries), computer familiarity seemed to have much more pronounced effects.

- In **Design A**, the multivariate statistics indicated a significant main effect of computer familiarity ($F=4.774$, sig.$<.0135$, partial $\eta^2=.169$). It was also found, in separate inspections of the results for CEHS and CPHS, that the main effect of computer familiarity was much larger on CEHS (mean difference=1.783, $F=9.456$, sig.<0.0035, partial $\eta^2=0.165$) than on CPHS (mean difference=1.246, $F=3.884$, sig.<0.0555, partial $\eta^2=0.075$), as shown in Figure 9.9. In fact, the effect on CPHS was only at the borderline of the pre-defined significance level. In both cases, low computer familiarity students had higher CEHS and CPHS than high computer familiarity students.

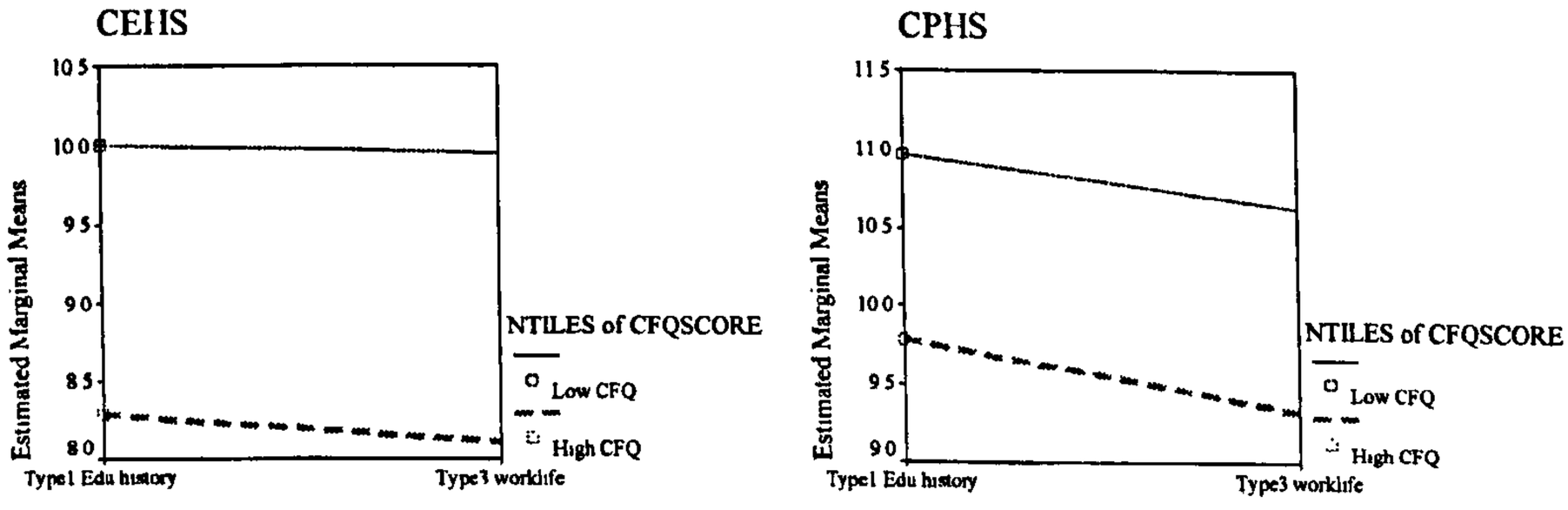


Figure 9.9 Main effects of computer familiarity on CEHS and CPHS (Design A)

- In **Design B**, the multivariate statistics indicated that computer familiarity had significant main effects ($F=5.811$, sig.<0.0065, partial $\eta^2=0.198$). The effects of computer familiarity were significant for both CEHS ($F=11.361$, sig.<0.0015, partial $\eta^2=0.191$) and CPHS ($F=5.091$, sig.<0.0295, partial $\eta^2=0.096$), though with a quite different magnitude of effect size. The low computer familiarity group had

significantly higher CEHS (mean difference=1.85, std. error=0.549) and CPHS (mean difference=1.367, std. error=0.606) than the high computer familiarity group in both language orders (Figure 9.10)

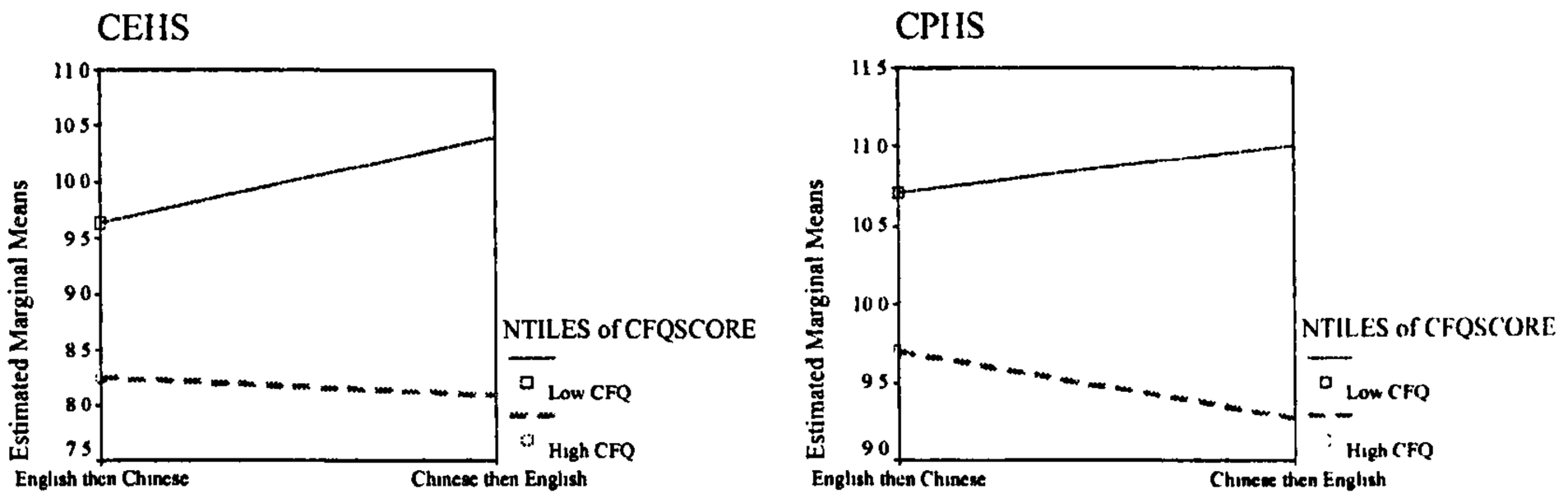


Figure 9.10 Main effects of computer familiarity on CEHS and CPHS (Design B)

- In **Design C** the effect of computer familiarity found was only approaching significance level (F=2.784, sig.<0.0725). When the results for the two dependent variables were considered separately, it was found that the effect of computer familiarity was significant on both CEHS (F=4.48, sig.<0.0395, partial η^2=0.084) and CPHS (F=4.034, sig.<0.0505, partial η^2=0.076), too.

- In **Design D**, the effect of computer familiarity (as a covariate) was also significant (F=3.702, sig.<0.0325, partial η^2=0.134). This significant effect was evident in both CEHS (F=5.909, sig.<0.0195, partial η^2=0.108) and CPHS (F=5.429, sig.<0.0245, partial η^2=0.1) when they were considered separately.

In summary, computer familiarity had significant main effects on the composite of the four HS scores. Although there was no such significant main effect when EEHS and EPHS were the dependent variables, significant interaction effects with *language order* were evident in EEHS (see Figure 9.7). Compared to high computer familiarity students, low computer familiarity students had lower EEHS when their summaries were produced in the order of *English then Chinese*, but higher EEHS when produced in the order of *Chinese then English.* When CEHS was the dependent variable, computer familiarity was also found to have a significant interaction effect with *text type* (see Figure 9.8). Compared to the high computer familiarity students, low computer familiarity students summarizing textA or textC had higher CEHS. However, as for textB summaries, high computer familiarity students had higher CEHS than their low computer familiarity

counterparts. When CEHS and CPHS were the dependent variables, computer familiarity had significant main effects on both CEHS and CPHS, but with larger effect sizes on CEHS. Computer familiarity again seemed to have more pronounced effects, be it main or interaction, on Chinese than English summarization performances.

9.1.3 Summary of main findings from summarization performances

Effects of text presentation mode

The only significant main effect of *text presentation mode* was on the lengths of the Chinese summaries. The Chinese summaries of computer presented texts were significantly longer than those of paper presented texts. It was also found that presentation mode had significant interaction effects with *text type* on CERSC and CEHS. The differences in these two scores between computer and paper presentation modes were notably larger in the Chinese summaries of textC. The effects of presentation mode on English summarization performances were far less prominent.

Effects of computer familiarity

Similarly, the effects of students' computer familiarity were more pronounced in Chinese than English summarization performances. Low computer familiarity students had statistically significantly higher CERSC, CPRSC, CEHS and CPHS than their high computer familiarity counterparts. In addition, computer familiarity was also found to have exerted significant interaction effects with *language order* on CERSC (Figure 9.5), EERSC (Figure 9.6) and EEHS (Figure 9.7) and with *text type* on CEHS (Figure 9.8). The lengths of summaries were not affected by the students' computer familiarity.

9.2 Students' perceptions of such effects

Data from the post-summarization questionnaire and interviews were analysed to examine how students thought their summarization performance might be influenced by the computer presentation mode and how their computer (lack of) familiarity might help or hinder their summarization performances.

9.2.1 Post-summarization questionnaire

Students who did the summarization tasks on computer were asked three questions in the PSQ (Appendix 4):

> **21.** How helpful was your level of computer familiarity for you to
> **(a)** "read to understand",
> **(b)** "read to summarize" the source text?
>
> **22.** And to which activity do you think your computer familiarity was more helpful, or equally (not) helpful?

Answers to questions 21a and 21b were moderately correlated (Spearman rho=0.577, sig.<0.0005). Being familiar with using computers was considered slightly more helpful in *reading to understand* than in *reading to summarize* the source text, though the helpfulness for both activities was not high (Table 9.2). Only 5-7% of these students considered it very useful for the summarization tasks.

	Not helpful at all		Not too helpful		Of average help		Somewhat helpful		Very helpful	
	Freq.	Valid %	Freq.	Valid %	Freq.	Valid %	Freq.	Valid %	Freq	Valid %
21a	4	5.1	27	34.6	7	9.0	34	43.6	6	7.7
21b	7	9.2	25	32.9	15	19.7	25	32.9	4	5.3

Table 9.2 Helpfulness of computer familiarity with reading to understand and reading to summarize

When asked to make a further distinction in the helpfulness of computer familiarity (Q22), students' answers were almost equally distributed (χ^2=1.896, df=2, n.s., see also Table 9.3 below).

Comparison of the helpfulness of computer familiarity towards:		
	Frequency	Percentage
reading to understand the text	29	37.7
reading to summarize the text	20	26.0
equally (not) helpful	28	36.4
Total	77	100.0

Note: excluding the 5 missing values

Table 9.3 Frequency of the helpfulness of computer familiarity

9.2.2 Post-summarization interviews

Data from the post-summarization interviews were analysed using winMAX. Three major themes emerged from the analyses: **(1)** acknowledged *physical* and *psychological* differences between doing the summarization tasks on paper and on

screen such as *historical friendliness, and tangibility and security and visibility* of reading on paper and the possibility of smart exploitation of computer facilitative functions, (2) minimal computer manipulation skills required for the summarization tasks, and (3) some but not significant effects expected for computer familiarity on summarization performances.

1) Acknowledged physical and psychological differences

Although the interviewees acknowledged that there was not much difference between the computer and paper presentation modes, they articulated some minute differences – both *physical* and *psychological* – between doing the summarization tasks on paper and on screen, such as (a) historical friendliness, (b) tangibility, security and visibility of reading on paper, (c) possible psychological shock of reading on screen, and (d) provision and use of some facilitative functions of MS Word.

a) Historical friendliness of reading on paper

Because of years and years of reading from print, some students thought they were highly accustomed to this "default" media for delivery, and therefore found reading on paper more "friendly", "convenient" and "tangible" and felt much more "secure". On the other hand, reading on screen was quite "awkward" and "tedious", compared to the default paper reading. Although they also acknowledged that they frequently read on computer screen, compared to paper reading, screen reading was nonetheless considered relatively new. This kind of timeline seemed to be important for them to identify themselves as a friend or a stranger to screen reading.

> We are used to reading on paper, where we can make notes or comments, write a gist of a paragraph, underline something, etc, but it would be quite awkward to do these on computers. To be frank, even though we are now at the fourth year, we started to use computers only after we entered the university, when we had courses on computers here. I had computer courses at senior secondary school, though. We, I believe, are able to do efficiently word processing, such as enlarging the font size and changing colours of a sentence; however, reading on computer, I think, we are still not very used to it. It is quite awkward. It is not as friendly as reading on paper, which we are familiar with...
>
> **Extract 9.1** Alice Zhang

> Basically I think there is not much difference. However, I find it is more convenient to read something you can hold in hand than reading on computers because our computer familiarity has not reached the level that we find it easier to read on computers, it is still the paper that we are more familiar with, although we started to use computers extensively since Year

> One at the university. After all, we have been reading on paper for so many years, compared to reading on computer screens... We read on computer screens, too... If it is not for a test, I prefer to read on computers, you can have hotlinks to loads of relevant information; if it is for a test, I would prefer to reading on paper.
>
> **Extract 9.2** Ben Zhang

> ... If I were to choose a presentation mode, I think I would choose to read a text on paper, because we have been so used to reading on paper for many years. I read on computer screen only when needed, and sometimes I also print out the materials so that I could read on paper.
>
> **Extract 9.3** Peter Zhang

b) Better tangibility, security and visibility of reading on paper

Apart from the *historical friendliness* of reading on paper, students also pointed out another "concrete" advantage of reading on paper over screen reading. As both Ben Zhang (Extract 9.2) and Grace Zhang commented (Extract 9.4), reading on paper was "tangible" and was something they could hold onto. Grace Zhang added that this would make her feel more "secure", too. When reading on screen, she felt she would have less control over the computer which might freeze.

> If I could choose, I would like to read the text on paper, because I think it is tangible, it is something you can hold in hand and you feel securer. Sometimes, computers can freeze, and I also love my eyes. However, there is not much difference between reading on paper and on computer screen...
>
> **Extract 9.4** Grace Zhang

The unfriendliness of reading on screen may be compounded to some extent by the length of the texts, as Alice Zhang mentioned that clicking on the answers in multiple choice tests was quite different from reading extended texts on screen:

> I think paperless tests are appropriate for multiple choices, we can just click the answers, but for reading a long text like the one I read in the summarization test, it is not as appropriate as for multiple choice questions.
>
> **Extract 9.5** Alice Zhang

Furthermore, it seemed that extended texts strengthened students' wishes to hold on to them rather than having to scroll up and down the pages.

> For whatever reasons, I would prefer to do the task on paper. For so many years, we are used to reading on paper, even though now we also use computers very often. Every week, I use computers for over six or seven hours, playing games, reading short online newspapers articles, sending emails, and hunting jobs, etc. However, for a very long text such as the one we had, it is on paper that you can feel you are holding onto something concrete. What's worse, it is tiring to read a long text on computer screen.
>
> **Extract 9.6** Katie Zhang

The length of source texts also seemed to make them less visible on computer screen because:

> ... Computer screen has a limited space, it is impossible to see the whole text within one screen. However, on paper, I can see the whole document, which is particularly useful for re-visiting pages already read. On computer, it is not easy to find a particular page. Especially, when I was writing the summary, it was extremely important to have the whole document at hand and to see the document as a whole.
>
> **Extract 9.7** Elyn Zhang

The full visibility, as well as tangibility, of the source texts was considered critical, especially for the summarization tasks. Seeing the source text as whole helped to write a summary of it, because, as Yvonne Zhang commented, it was "*easier to organize ideas on paper*" (see also Extract 9.1).

c) Psychological impacts of screen reading

However, not only did text presentation mode seem to have physical effects (e.g. fatigue, not being the same as doing multiple choice tests on computer), it also had some emotional or psychological impact. Quentin Zhang felt:

> It was a very long text ... and this feeling may be because the text was presented on computer. When you read on computer screen, the eyes were tired...
>
> **Extract 9.8** Quentin Zhang

Victoria Zhang (Extract 9.9), however, commented on the combined effects of text length and presentation mode from a very different perspective.

> I have my own computer at home; I am very familiar with using computers. I prefer to reading on computer screen. As you know, the text was six pages long. If it were presented to me on paper, I could have been shocked. Six pages in English! I can't bear reading it. But when it was presented on computer, you didn't know exactly how many pages the text had. Psychologically, because you don't know how long it is, you won't be so scared; you don't care.
>
> **Extract 9.9** Victoria Zhang

In fact, a very experienced user of MS Word would easily identify how many pages a document had as this information is presented at the bottom of the window. The reasons were not clear why these two students, who had computers at home and used them frequently, had such psychological reactions against the computer presentation mode.

Other students also expressed similar psychological impacts of reading on screen, but to them these impacts were short-lived and were quite easily overcome:

> ... Computer familiarity may have some effects on our emotions, for example, if we have never read on computers, the summarization test on computers may be a shock to us; but if we are very familiar with using computers, with reading on computers, we can immediately feel at home. Though I seldom read on computers, I found it was not difficult to make myself home in reading the text.
>
> **Extract 9.10** Ian Zhang

> ... Before reading on computer screen, I did have a feeling that I would rather read on paper; but when I started reading the text, that feeling was gone. No difference, I think.
>
> **Extract 9.11** Quentin Zhang

d) Provision and use of facilitative functions of MS Word

Three of the 11 interviewees who read the texts on computer and 4 of the 12 interviewees who read the texts on paper pointed out the physical disadvantages of reading on computer screen: "radiation", "fatigue" and "tediousness". However, these disadvantages could well have been compensated for by the facilitative functions of computers which some interviewees claimed to be "convenient" and time-saving. The advantages of using computers, in particular the facilitative functions of MS Word such as copying, pasting, deleting, underlining, and highlighting, could easily be exploited by most of the students who had to pass national or provincial examinations which had special emphasis on skills in using Word.

> ... I, first of all, changed the colours of the sentences I think important while reading, and then could copy them later. That's it, no other computer skills used.
>
> **Extract 9.12** Grace Zhang

> I opened a new Word window for a new file so that I could copy and edit the important sentences from the original text to the new window. It saved a lot of time. If I were to do a similar task, I would still prefer to reading on computer; and if I could write the summary on computer, it would be better. Typing is quicker than writing in hand... I am familiar with using computers; I use computers quite often. And I am also able to write and edit some programmes.
>
> **Extract 9.13** Michael Zhang

> If you were not familiar with using computers, especially Word, you might not know how to underline an important sentence, and no mark was left on the sentence. You may well forget the important sentences. However, if you are familiar with using Word, you could underline the important sentences, and when you finish reading the whole text, you can read and edit these sentences ***only*** and write a good summary... I think if you are familiar with computers, it is helpful and you can finish the summarization task faster. I

> am not that familiar with using computers, but as for Word, I know quite well; we had to pass examinations on how to use Word in the university…
>
> **Extract 9.14** Peter Zhang

> While you are reading on computer, you only need to use the mouse to scroll to the next pages; but when you are reading on paper, you have to turn the pages… I use computers often. When I was doing the summarization tasks, I used some Word functions such as copy, paste and edit, which made life much easier.
>
> **Extract 9.15** Tom Zhang

> … However, there are also some advantages of doing the summarization tasks in the computer room, because you can change colours of the sentences and copy, paste and edit them in Word. I think most of us can do these basic functions in Word, you don't have to be very familiar with using computers in general.
>
> **Extract 9.16** Ulysses Zhang

> You can use some functions of the computers to find the meanings of new words. It is helpful in understanding the text… I find it much more convenient to do the summarization tasks on computer screens, you can delete the unimportant sentences such as those examples and statistics and adjectives. I read the text twice. At the second time, I started to delete those unimportant sentences till there were only around 700-800 words left, and then I re-organized those sentences to produce a summary of the original text. It also saved time.
>
> **Extract 9.17** Wendy Zhang

These advantages of the computer presentation mode were not only expressed but also exploited by the students. It seemed that students in the different experimental conditions (computer or paper) used quite different strategies to do the summarization tasks. The smart use of Word functions such as copying, pasting and colouring important sentences was prevalent and helpful for the summarization tasks.

2) Minimal requirements of computer manipulation skills

The smart use of Word functions were further facilitated by the fact that (i) only some basic computer manipulation skills were required for the summarization tasks, and (ii) most of the students were quite familiar with MS Word (see Appendix 14). Four interviewees of the 11 who did the summarization tasks on computer commented that only very basic computer skills were required to read and summarize the source texts. Comments from Ian Zhang (Extract 9.18) and Ben Zhang (Extract 9.19) were quite representative.

> As long as you can use mouse to scroll the pages, that's it. No other computer skills were required in the summarization test. …
>
> **Extract 9.18** Ian Zhang

> ... In the test, only very basic computer skills were required. ... It is more like a question of whether you are familiar or not, it is not a question of whether you can do it or not. ...
>
> **Extract 9.19** Ben Zhang

3) Some but not significant effects expected and experienced relating to computer familiarity

Although some physical and psychological differences were acknowledged and expected, they were not necessarily experienced by the students. Several of them (e.g. Alice, Extract 9.20) commented that there might be some but not substantial effects of computer familiarity on summarization performances. This could be due to the fact that only minimal computer manipulation skills were required for successful completion of the summarization tasks in this project (see 2 above).

> OK, I think there might be some effects of my computer familiarity, but I don't think it would have a close relationship between my computer familiarity level and summarization activities.
>
> **Extract 9.20** Alice Zhang

Computer presentation of the source texts was only a different means of delivery; it was not a barrier to understanding of the texts and could not take the place of understanding the source text as the prerequisite of successful summarization.

> Computer familiarity could have some effects, more or less; but these effects are not large. We read on computer screens, too. I don't think computer familiarity would be a barrier in understanding the text....
>
> **Extract 9.21** Ben Zhang

> To me, I think it is only a matter of different means. The most important thing is to understand the text. ... I don't think it would affect my summarization performance if I were not familiar with using computers.
>
> **Extract 9.22** Daniel Zhang

Even low computer familiarity was not considered to have hindered summarization performance to any great extent.

> That's it, no other computer skills used. I don't use computers very often, but sometimes, I also use computers over three hours. I don't think my low computer familiarity would affect my summarization performances.
>
> **Extract 9.23** Grace Zhang

However, computer familiarity was likely to exert some psychological impact on the two extreme groups, as Ian Zhang commented (Extract 9.10). For students who had never read texts on computers before, undertaking the summarization task on computer would have been a "shock" to them; for those who were very familiar with

reading on computers, they would "immediately feel at home". However, even as a rare user of computers, Ian Zhang reported that "it was not difficult to make myself at home in reading the text" (on computer screen).

On the other hand, for an efficient and frequent user of Word, it might be "helpful" enough to enable the student to finish the summarization task faster than low computer familiarity students.

> ... I think if you are familiar with computers, it is helpful and you can finish the summarization task faster. I am not familiar with using computers, but as for Word, I know quite well; ...
>
> **Extract 9.24** Peter Zhang

Interestingly, almost all students who did the summarization tasks on paper, when asked whether they would like to do the tasks on computer, had a very strong preference to sticking to their original experimental condition, except for one interviewee who had very high self-evaluation of computer familiarity and would like to try reading on screen, but simply for a change.

> I would try to have a go at doing the summarization task on computer screen. I even suggested to friend from another class if he would be willing to change our experimental conditions. I like computers, and also use computers very often.
>
> **Extract 9.25** Helen Zhang

However, those students who did the summarization tasks on computer were rather ambivalent in their preference for the text presentation mode. They seemed less concerned as to which presentation mode they would be assigned. This may well reflect that the impact of text presentation mode on students' summarization performances were not experienced, although somewhat expected by those who did not do the summarization tasks on computer.

9.3 Summary of findings relating to RQ4

This research question examined the actual and the perceived effects of (a) text presentation mode and (b) computer familiarity on summarization performances.

The statistical analyses on the three key quality indicators (RSC, HS and Lengths) of summaries demonstrated some significant effects of both text presentation mode

and computer familiarity on actual summarization performances. The effects on Chinese summarization were more pronounced than on English summarization. In particular, text presentation mode had a significant main effect on the lengths of the Chinese summaries – the only significant main effect of presentation mode on summarization performances. The Chinese summaries of computer presented texts were significantly longer than those of paper presented texts. Text presentation mode was also found to have exerted a significant interaction effect with *text type* on CERSC and CEHS (see also **10.1.2**).

Computer familiarity had significant main effects on all the four quality indicators of Chinese summaries (i.e. CERSC, CPRSC, CEHS and CPHS). It was found that the Chinese summaries of low computer familiarity students were of a statistically significant higher quality than their high computer familiarity counterparts. Computer familiarity also had significant interaction effects with *text type* on CEHS, and with *language order* on CERSC, EERSC and EEHS. Students' computer familiarity did not affect the lengths of their summaries.

However, these statistical findings from the performance data were only partly supported by the student perception data. The students did not think text presentation mode and computer familiarity would affect their summarization performances to a great extent. Those who were interviewed acknowledged that there was not much difference between the two presentation modes, but they also articulated some minute physical and psychological differences between the two and that computer presentation mode might have affected their summarization. Nevertheless, it was also agreed that (lack of) familiarity with using computers did not help or hinder their summarization because basic computer manipulation skills were sufficient for the use of the facility of MS Word – the programme they were most familiar with (see Appendix 14). The interview data also indicated that the effects of computer familiarity might be more expected than experienced.

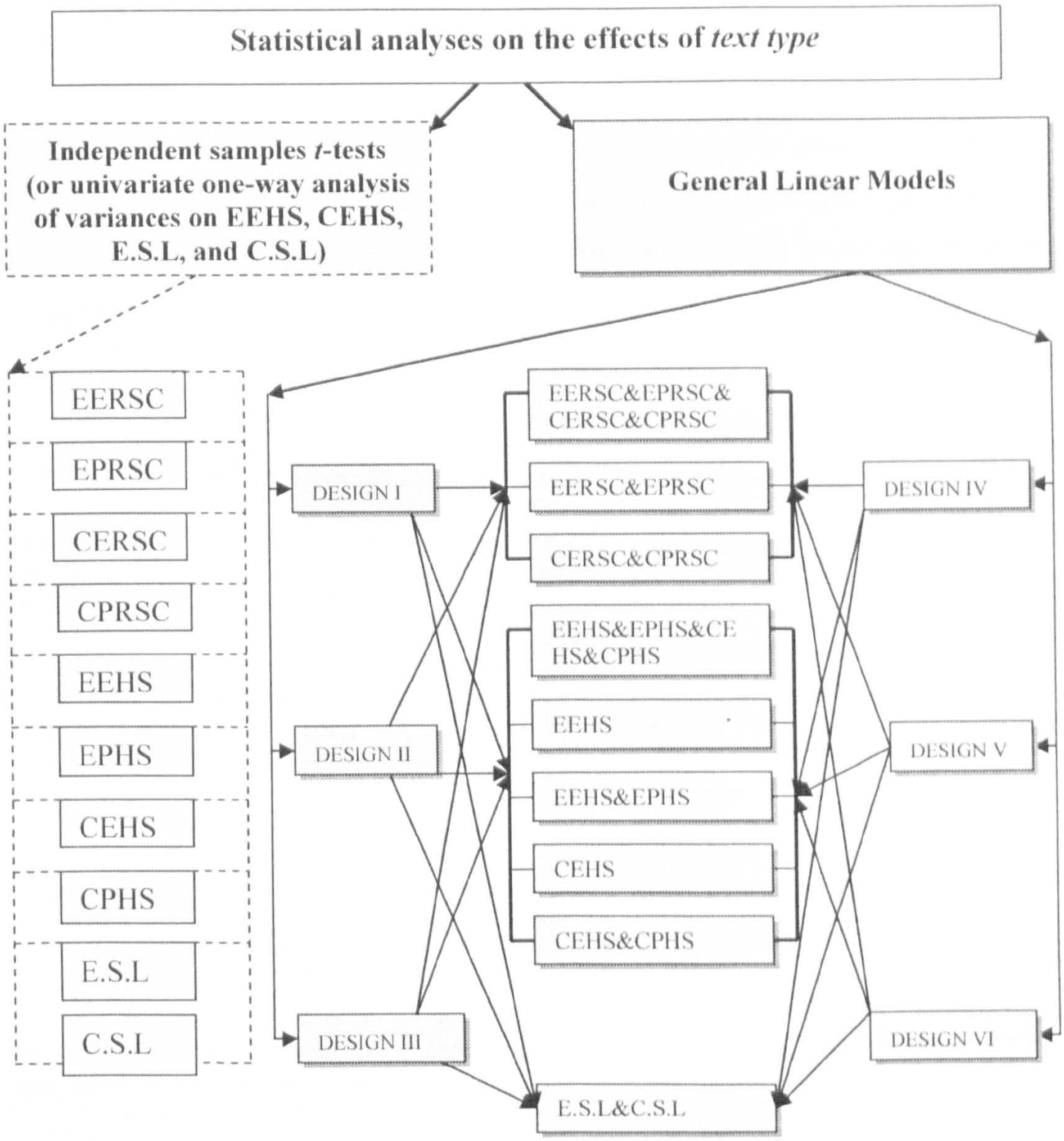
Further discussion of these findings is reported in **11.2.2.1**.

CHAPTER TEN

Effects of Text Type

What are the effects of text type on students' summarization performances?

In the previous chapter, I analysed the effects of text presentation mode on summarization performances. This current chapter will focus on the actual and perceived effects of another important feature of INPUT (see Figure 2.1) – *text type* on summarization performances (RQ5). As can be seen in Figure 10.1, the students' summarization performance data were analysed first through a series of independent samples *t*-tests (or univariate one-way ANOVA, where appropriate) of the effects of *text type* on each individual quality indicator (RSC, HS and Length) of summaries. Six GLM designs were then used to examine further the effects of *text type*, taking into account the effects of other experimental factors such as *text presentation mode*, *language order* and students' reading ability as measured by TOEFL. Students' perceptions of such effects, if any, were discussed in the light of the various indicators of the summarizability of source texts (see also the interaction effects of *text type* in relation to the key factors under investigation in Chapters 6-9).



GLM Designs♠:

Design I: Intercept+TXT

Design II: Intercept+TXT+PRESMODE+TXT*PRESMODE (see also Figures 8.1 and 9.1)

Design III: Intercept+TXT+LANGORD+TXT*LANGORD (see also Figure 8.1)

Design IV: Intercept+TXT+TOEFL

Design V: Intercept+TXT+PRESMODE+TOEFL+TXT*PRESMODE

Design VI: Intercept+TXT+LANGORD+TOEFL+TXT*LANGORD

Colour scheme:

Independent samples *t*-tests (or univariate analysis of variances on EEHS and CEHS)

General Linear Models

Dash style:

- - - : Independent samples t-tests, _____: General Linear Models

Note♠: Design I was not applied to EEHS or CEHS because these univariate one-way analyses of variances had already been conducted in the *independent samples t-tests*. Designs II was also analysed in Chapters 8 and 9, and Design III in Chapter 8.

Figure 10.1 Plan for the statistical analyses on the effects of text type on summarization performances

10.1 Students' actual summarization performances

10.1.1 Independent samples t-tests by text type

Several independent samples *t*-tests (and univariate one-way analyses of variances in the case of EEHS, CEHS, E.S.L and C.S.L) were first conducted to examine the effects of *text type* on summarization performances, without taking into account at the same time the possible influences of *language order* and *presentation mode* (Table 10.1).

Scoring criteria	Text	Mean	Standard deviation	t/F*	sig.<	η^2
EERSC	A	49.604	12.2605	3.089	0.0035	0.0887
	C	42.500	10.5191			
EPRSC	A	54.821	11.3269	1.788	n.s.	0.0316
	C	50.819	10.9883			
CERSC	A	47.443	12.3692	3.680	0.0005	0.1214
	C	37.149	15.5675			
CPRSC	A	51.302	10.8731	4.405	0.0005	0.1653
	C	41.830	12.5457			
EEHS	A	10.557	2.2007	$F_{2,154}$=4.189	0.0175	0.0516
	B	10.956	1.9600			
	C	9.809	1.8984			
EPHS	A	11.604	1.8354	1.552	n.s.	0.0240
	C	10.989	2.1226			
CEHS	A	9.981	2.1794	$F_{2,154}$=9.815	0.0005	0.1131
	B	10.123	2.0029			
	C	8.319	2.5990			
CPHS	A	10.745	1.8979	2.662	0.0095	0.0674
	C	9.606	2.3750			
E.S.L**	A	304.14	69.220	$F_{2,149}$=1.645	n.s.	0.0216
	B	313.82	66.898			
	C	290.51	57.876			
C.S.L**	A	542.15	143.236	$F_{2,152}$=3.989	0.0205	0.0499
	B	476.68	121.381			
	C	481.45	131.261			

Note: * Independent samples *t*-tests or one-way ANOVAs were applied to data where appropriate. df=98 for the independent samples *t*-tests. ** The univariate outliers (ID: 4102, 4107, 4215, 3205, 4118 for E.S.L; ID: 4102, 3118 for C.S.L) were excluded.

Table 10.1 Independent samples *t*-test on the effects of text type on summarization performances

The effects of *text type* on the actual summarization performances are:

- Text type had significant effects on all quality indicators of summarization performances with a medium to large effect size (η^2=0.0499~0.1653), except for E.S.L, EPRSC and EPHS (i.e. when an English summary was judged according to the popular scoring template, see Chapter 6).

- As shown in the independent samples *t*-tests, summaries of textA always had significantly higher EERSC, CERSC, CPRSC and CPHS than summaries of textC.
- As shown in the one-way ANOVAs, *text type* had significant effects on EEHS, CEHS and C.S.L. The post hoc Scheffe tests indicated that (i) summaries of textB had significantly higher EEHS than those of textC (mean difference=1.148, sig.<0.0185); (ii) summaries of textC had significantly lower CEHS than those of textA (mean difference=-1.622, sig.<0.0025), and also textB (mean difference=-1.804, sig.<0.0005); and (iii) the Chinese summaries of textA were significantly longer than those of textB (mean difference=65.48, sig.<0.0395). No other significant differences in these three quality indicators were significant between summaries of the three texts.
- In all cases, summaries of textA received significantly higher scores than those of textC (except for EPRSC and EPHS).

10.1.2 General linear models

As discussed in the previous chapters, the independent samples *t*-tests and univariate analyses of variances were not able to incorporate simultaneously other potential contributing factors to summarization performances, besides *the* particular variable under investigation. Therefore, several multivariate general linear models were applied to the data of summarization performances to examine the effects of text type in association with such factors as *language order*, *presentation mode* and the summarizer's reading abilities measured by *TOEFL* (see Figure 10.1). See **9.1.2** for the procedures of checking the multivariate assumptions.

1) RSC

The full statistics of the multivariate general linear modelling on the three combinations of RSC (see Figure 10.1) are presented in Appendix 32. The key findings of the multivariate tests are summarized below:

- *Text type* had significant main effects on all the four RSC scores as a composite when they were the dependent variables in each of the six models (I-VI), with very large effect size using partial η^2 (ranging from .135 to .190).

- However, when the dependent variables included only EERSC and EPRSC, it was found that only in the first three models (I-III) did *text type* have significant effects on these two scores as a composite. In the remaining three models (IV-VI) where TOEFL-R was a covariate, the effects of text type were diminished and were either borderline in terms of significance or non-significant (see also Chapter 7 for the additional effects of *text type* in the sequential regression analyses).
- When the dependent variables included only CERSC and CPRSC, it was found that *text type* had significant main effects in all six models, with a very large effect size (partial η^2 ranging from .12 to .179).
- The effect sizes of *text type* were much larger on RSC of the Chinese summaries (i.e. CERSC and CPRSC) than on RCS of the English summaries (see also Chapter 9 for the larger effect of *text presentation mode* on Chinese summarization performances).
- The multivariate statistics also showed no significant interaction effect of *text type* with other factors in the models.

Separate inspections of the effects of *text type* for each dependent variable indicated that *text type* had significant effects on all the RSC scores but EPRSC, with effect size, partial η^2, ranging from 0.053 to 0.156. Summaries of textA received significantly higher EERSC, CERSC and CPRSC. It was also found that:

- *text type* and *presentation mode* had significant interactive effects on CERSC (Model II: F=5.320, sig.<0.0235, partial η^2=0.053; see also Chapter 9);
- *text type* and *language order* had significant interactive effects on CPRSC (Model III: F=4.206, sig.<0.0435, partial η^2=0.042; Model VI: F=4.056, sig.<.0475, partial η^2=.041; see also Chapter 8);

2) HS of textA and textC summaries

Similarly, three combinations of HS scores (see Figure 10.1) were also subjected to the multivariate analyses. The full statistics are presented in Appendix 33. The key findings of the multivariate tests are summarized as follows:

- *Text type* had significant main effects on the four HS scores as a composite in all the models but IV and VI where TOEFL-R was a covariate and when the effects of

text *presentation mode* were not taken into account. The effect size was large, ranging from .101 to .138.

- *Text type* did not have statistically significant main effects in any of the six models when EEHS and EPHS (i.e. HS of English summaries) were dependent variables.
- *Text type* had significant main effects on CEHS and CPHS (i.e. HS of Chinese summaries) in all the six models, with quite a large effect size, partial η^2 ranging from .08 to .131.
- Apart from the main effects mentioned above, the multivariate statistics also showed that text type had significant interaction effects with language order (Model III: F=3.647, sig.<.0305, partial η^2=.071; Model VI: F=3.553, sig.<.0335, partial η^2=.07).

The statistics of the univariate tests demonstrated that the significant multivariate main effects of *text type* were largely due to its effects on CEHS and CPHS (i.e. HS scores of Chinese summaries). Two significant interactive effects were also noted:

- *Text type* and *presentation mode* had significant interactive effects on CEHS (Model II: F=6.037, sig.<0.0165, partial η^2=0.059; Model V: F=4.188, sig.<0.0435, partial η^2=0.042; see also Chapter 9).
- *Text type* and summarization *language order* had significant interactive effects on CPHS (Model III: F=6.878, sig.<0.0105, partial η^2=0.067; Model VI: F=6.696, sig.<.0115, partial η^2=.066; see also Chapter 8).

3) EEHS, CEHS of all summaries

The three combinations of the HS scores in the multivariate analyses in the section above excluded the use of the data from textB summaries which were assigned only EEHS and CEHS (see Table 4.8). Therefore, these two scores were analysed separately in five univariate analyses (see Figure 10.1). The full statistics of these univariate analyses are presented in Appendix 34.

Text type was found to have statistically significant main effects on both EEHS and CEHS. The pairwise comparisons indicated that the statistically significant main effects of *text type* were largely due to the following significant differences:

- in EEHS between textB and textC summaries, and
- in CEHS between textA and textC, textB and textC summaries.

Significant interactive effects on CEHS between *text type* and *presentation mode* were also noted in Model II ($F=3.37$, sig.<0.0375, partial $\eta^2=.043$).

4) Lengths of summaries: E.S.L and C.S.L

The multivariate tests indicated that *text type* had significant main effects on the lengths of both the English and Chinese summaries as a composite score, with an effect size using partial η^2 of about .05. It did not have significant interaction effects with the other factors in the models. According to the univariate statistics, the main effect of *text type* was largely due to its significant effect on the lengths of the Chinese summaries rather than the English ones. There were also significant interaction effects of *text type* and *language order* on the lengths of the Chinese summaries (see Appendix 35 for the full report of the statistics).

The pairwise comparisons indicated that the Chinese summaries of textA were significantly longer than those of textB and textC. The differences between textB and textC were however not significant. It was always textA summaries that were longer than the other two (i.e. textA>textC>textB), no matter whether the difference was significant or not. This seemed to be at the same order as the difficulty level of the texts according to student evaluation (see **10.2.1** below). Students tended to produce longer summaries for easier texts than the more difficult ones.

In summary, *text type* was found to have statistically significant effects on all the three quality indicators of summarization performances to varying degrees. It significantly affected (i) all the four RSC scores but EPRSC, (ii) HS of Chinese summarization, and (iii) the lengths of both the English and Chinese summaries. In addition, it was found that *text type* had much larger effects on the three quality indicators of Chinese summarization performances than on English. However, when the summarizers' reading comprehension abilities as measured by TOEFL-R were taken into account, the effects of *text type* were reduced. Text type also exerted significant interactive effects with *presentation mode* on CERSC and CEHS, with *language order* on CPRSC, CPHS and the lengths of the Chinese summaries. In all the contexts above, only the performance data of textA and textC

were examined. Students who summarized textA seemed to have produced better summaries than those who did textC. When the quality of textB summaries was included in the univariate analyses (dependent variable: EEHS, CEHS respectively), similar findings were observed, such as larger effect sizes on Chinese (i.e. CEHS) than English summarization performances. However, it was also evident that *text type* had significant main effects on EEHS. In both EEHS and CEHS, textB summaries were judged to be of slightly better quality than textA summaries, but significantly better than textC summaries.

These significant main and interactive effects suggest that text type must have played an important role in the students' summarization performances. However, aside from knowing that there were such statistical effects, there is no evidence of what caused these effects and to what extent students actually experienced them. The analyses on the post-summarization questionnaire and interviews were intended to shed some light on these issues.

10.2 Students' perceptions of the effects of text type on summarization performances

Students' views of the possible effects of text type on their summarization performance were obtained from the post-summarization questionnaire (Appendix 4) and interviews (Appendix 6).

10.2.1 Post-summarization questionnaire

The first 10 questions of the PSQ asked the students to evaluate difficulty levels in understanding and in summarizing the texts, and their familiarity with the topics of the texts and whether and to what extent (lack of) familiarity impeded/helped with their process of understanding and summarizing the source texts.

1) Difficulty level in understanding and summarizing source texts

Text difficulty was evaluated from two perspectives: *understanding* and *summarizing*. According to the Kruskal-Wallis Tests, there were significant differences between the three texts in the perceived difficulty of *understanding* (χ^2=27.776, sig.<0.0005), though not of *summarizing* (Table 10.2). TextC was

considered significantly more difficult to understand than the other two texts (C>B>A). TextC was also considered more difficult to summarize than textA and textB, though this difference was not statistically significant (C>B>A).

		Mean Ranks	
		Difficulty in understanding	Difficulty in summarizing
TEXT	A	54.52	69.59
	B	87.00	78.48
	C	94.72	83.31
χ^2		27.776	2.914
Sig.		0.0005	0.2335 n.s.

Table 10.2 Kruskal-Wallis Tests of the differences in difficulty in understanding and summarizing the three source texts

Overall, the students thought understanding and summarizing the source texts were both moderately difficult/easy (Table 10.3). Furthermore, the Wilcoxon Signed Ranks Test indicated that summarizing was considered significantly more difficult than understanding the source texts (Z=-6.751, sig.<0.0005), although there were around 50% ties (68/153).

	Understanding (%)	Summarizing (%)
Easy	1.9	0.0
Somewhat easy	20.5	8.5
Moderately easy/difficult	54.5	38.6
Somewhat difficult	22.4	47.1
Very difficult	0.6	5.9
Wilcoxon Signed Ranks Test▲	Z=-6.751, sig.<0.0005	

Note: ▲ Overall difficulty in summarizing – overall difficulty in understanding; based on negative ranks.

Table 10.3: Overall difficulty in understanding and summarizing the texts and Wilcoxon signed ranks test

The data from the post-summarization interviews provide further insights into the difficulty or summarizability of each source text. The interviews also indicated that textA was easier to summarize (see **10.2.2** for details).

2) Topic familiarity and its helpfulness for understanding and summarizing source texts

The majority of the students were not too familiar with the general or the specific topics of the texts (Table 10.4). As anticipated, the students were more familiar with the general than the specific topics, according to the Wilcoxon signed ranks test (Z=-7.114, sig.<0.0005).

	General topic (%)	Specific topic (%)
Not familiar at all	19.9	51.3
Not too familiar	51.9	37.5
Of average familiarity	21.2	8.6
Somewhat familiar	7.1	2.0
Very familiar	0.0	0.7
Wilcoxon Signed Ranks Test	Z=-7.114, sig.<0.0005	

Table 10.4 Familiarity with the general and the specific topics

The Kruskal-Wallis tests indicated that there was no significant difference in students' familiarity with the *general* topics between the three source texts (χ^2=1.019, n.s). However, significant difference in familiarity with the *specific* topics was found between the three texts (χ^2=12.851, sig.<0.0025). In both contexts, textA was considered more familiar than textC and textB.

According to the Wilcoxon Signed Ranks tests on the data from Questions 3a and 3b (Appendix 4), both the general and the specific topic familiarity were considered more helpful for understanding than for summarizing the source texts (Z=-4.905, sig.<0.0005 for helpfulness of general topic familiarity; Z=-3.355, sig.<0.0015 for helpfulness of specific topic familiarity). The differences in the helpfulness of the general topic familiarity for understanding (χ^2=1.191, n.s.) and for summarizing the texts (χ^2=1.124, n.s.) between the three texts were not statistically significant, neither was there any difference in the helpfulness of the specific topic familiarity (χ^2=3.329, n.s. for understanding; χ^2=1.714, n.s. for summarizing). This finding was further supported in the data from Questions 4 and 7; these questions asked the students directly which activity, understanding or summarizing, benefited more from their familiarity with the general topics (Q4) and the specific topics (Q7). Chi-square tests indicated that understanding of the source texts was considered to have benefited more from the general topic familiarity (χ^2=52.71, sig.<0.0005) and the specific topic familiarity (χ^2=15.925, sig.<0.0005) than summarizing the texts. Again, Kruskal-Wallis tests found no significant differences in views on this comparison of helpfulness between the three text groups of students (χ^2=0.653, n.s. for the comparison of helpfulness of general topic familiarity; χ^2=1.774, n.s. for the comparison of helpfulness of specific topic familiarity).

Although according to the statistical analyses it seemed that topic familiarity was beneficial only to understanding the source texts, this finding may be interpreted as also benefiting summarization because successful summarization was first of all based

on proper understanding of the source texts (see **2.2**).

Overall, it should be borne in mind that the difference in the helpfulness for understanding and summarizing was situated in the large context that both general and specific topic familiarity were considered only moderately helpful (Table 10.5).

	Helpfulness of general topic familiarity for		Helpfulness of specific topic familiarity for	
	Understanding %	Summarizing %	Understanding %	Summarizing %
Not helpful at all	0.8	2.4	1.2	1.3
Not too helpful	19.2	27.6	14.8	23.8
Of average help	27.2	36.6	33.3	32.5
Somewhat helpful	44.0	27.6	32.1	35.0
Very helpful	8.8	5.7	18.5	7.5

Table 10.5 Percentage of helpfulness of topic familiarity for understanding and summarizing

3) Topic familiarity and judgments of text difficulty

Students' judgement of the difficulty level of the texts was significantly correlated with their familiarity with the general and the specific topics of the texts, although the magnitude was small. As anticipated, the more familiar they were with the topics, the easier they considered the source texts were (Table 10.6).

Familiarity with	Overall difficulty in	
	Understanding	summarizing
General topic	-0.179, sig.<0.0255	-0.184, sig.<0.0235
Specific topic	-0.140, n.s.	-0.172, sig.<0.0365

Note: The correlation statistics are Spearman rho.

Table 10.6: Correlations between text difficulty and topic familiarity

4) Relationships between summarization performances and judgements of text difficulty and topic familiarity

The analyses above (1, 2, and 3) were based on data from the post-summarization questionnaire only. It was not clear to what extent these results corresponded to the students' actual summarization performances. Several one-way ANOVA and independent samples *t*-tests, where appropriate, were conducted to test the means differences in *RSC*, *HS*, and *Length* between low, (medium), and high groups of *text difficulty* (in understanding and summarizing) and (general and specific) *topic familiarity*.

A series of one-way ANOVA were conducted to examine the differences in RSC,

HS and Length between these three[1] groups of *text difficulty in understanding*. As anticipated, those who considered the text more difficult to understand also had lower RSC and HS scores, and they also tended to write longer summaries, particularly, in Chinese. However, it was only in EERSC, CERSC, CPRSC, and CEHS that the difference was significant (see Table 10.7 for the ANOVA statistics and Appendix 36 for the multiple comparisons). It is interesting to note the significant differences were mainly in RSC scores.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
EERSC	Between Groups	1210.580	2	605.290	4 548	013
	Within Groups	12776.213	96	133.086		
	Total	13986.793	98			
EPRSC	Between Groups	403.007	2	201.503	1.585	210
	Within Groups	12204.983	96	127.135		
	Total	12607.990	98			
CERSC	Between Groups	1832.053	2	916 027	4 649	.012
	Within Groups	18915.992	96	197.042		
	Total	20748.045	98			
CPRSC	Between Groups	1600.700	2	800.350	5 613	005
	Within Groups	13689.300	96	142 597		
	Total	15290.000	98			
EEHS	Between Groups	7.756	2	3.878	.909	.405
	Within Groups	653.013	153	4.268		
	Total	660.769	155			
EPHS	Between Groups	14.276	2	7.138	1 817	.168
	Within Groups	377.077	96	3.928		
	Total	391.354	98			
CEHS	Between Groups	64.609	2	32.304	6.198	.003
	Within Groups	797.385	153	5.212		
	Total	861.994	155			
CPHS	Between Groups	24.727	2	12.364	2 698	.072
	Within Groups	439.853	96	4 582		
	Total	464.581	98			
E.S.L	Between Groups	8871.145	2	4435.573	.763	468
	Within Groups	889119.335	153	5811.237		
	Total	897990.481	155			
C.S.L	Between Groups	103210.371	2	51605.185	2.342	100
	Within Groups	3371949.527	153	22038.886		
	Total	3475159 897	155			

Table 10.7 ANOVA statistics of summarization performances by three groups of text difficulty in understanding

[1] The original 5 groups of *text difficulty in understanding* were re-categorized post hoc into 3 optimal groups [Group A (35)= *easy* (3) + *somewhat easy* (32); Group B(85)=*moderately easy/difficult* (85); Group C(36)=*somewhat difficult* (35) + *very difficult* (1)].

No significant means difference was found in any of the quality indicators between the two[2] optimal groups on *text difficulty in summarizing*, between the two[3] groups on *specific topic familiarity*, or between the three[4] groups on *general topic familiarity*.

10.2.2 Post-summarization interviews

From the data of the post-summarization interviews, in which the students were asked to talk about the most striking features of a text that they thought made it easy or challenging in the summarization tasks[5], four major themes emerged: (i) structure/organisation, (ii) vocabulary, (iii) length of the source texts and (iv) summarizers' knowledge of the topics. These features played interactively to varying degrees for different summarizers and source texts, resulting in their different perceptions of the summarizability of the source texts. The clear structure or organisation of textA was considered extremely helpful and made it friendly to summarization; the key obstacle in textB seemed to be its high frequency of new/unknown words; the major problem with textC seemed to be the interactive effects of its length and loosely organised and the diffuse views on work life balance expressed by different parties.

1) Text structure or organisation

TextA was rated the easiest to understand and summarize, according to the PSQ data. This finding was further confirmed in the interviews. Seven out of the eight interviewees from the textA group mentioned that the clear structure or organisation of the text, from the general to the specific content, was very helpful for both the

---

[2] The original 5 groups of *text difficulty in summarizing* were re-categorized post hoc into 2 optimal groups [Group A (72) =*somewhat easy* (13) + *moderately easy/difficult* (59); Group B (81) =*somewhat difficult* (72) + *very difficult* (9)].

[3] The original groups of *specific topic familiarity* were re-categorized post hoc into 2 optimal groups [Group A=not familiar at all (78); Group B (74)=not too familiar (57) + of average familiarity (13) + somewhat familiar (3) + very familiar (1)].

[4] he original 5 groups of *general topic familiarity* were re-categorized post hoc into 3 optimal groups [Group A=*not familiar at all* (31); Group B=*not too familiar* (81); Group C=*of average familiarity* (33) + *somewhat familiar* (11)].

[5] Another possible approach to collect such data is to ask the interviewees to select from a pre-defined list of text features that they think can affect their summarization performance. The researcher then analyzes the frequency of the text features chosen by the students. I think this could mislead the students as multiple choice questions do, and confine the students' views within the pre-defined list. For the purpose of this project, I decided to have the interviewees comment rather than me cueing the possible choices.

English and Chinese summarization tasks. In particular, the first introductory paragraph, which was considered already to be a very good summary of the text, was much valued by the students.

> The text was not difficult, I think, it was acceptable... It was not because there were no new words (in fact, there were words I did not know); rather it was because of the organization of the text: It had a general introduction at the very beginning and then details for each individual country, and also had very clear time and space order. From past to present, and from one country to another. These are all very clear, and you could easily adjust yourself to that....
>
> **Extract 10.1** Ian Zhang

> This text started with a general introduction and then provided some specific supporting details for the claims at the introduction. It was from "general" to "specific" ... Such a structure was friendly to summarization. Without such a clear structure, it could have been more difficult to do the tasks.
>
> **Extract 10.2** Jack Zhang

> This text had no new words, no difficult sentence structures, I think, the text was not difficult... The text had a very clear structure. At the beginning, it had a general introduction which served as a kind of summary for the whole text, though a bit longer than what a required summary should be. The text then described these features, which were mentioned in the general introduction, country by country.
>
> **Extract 10.3** Katie Zhang

> This text was structured from "general" to "specific"... The first paragraph gave you a very good overview, which was like a cue, a guideline for the whole text.
>
> **Extract 10.4** Louis Zhang

> The text had a very clear organization or tread, so it was easy to analyse. Actually, after finishing reading the first two paragraphs, you had already had an overview of the whole text... The clear organization of the text was particularly helpful to summarization, because I can literally copy the first couple of paragraphs and use them as part of English summary. Actually, I was almost there by copying these paragraphs. And for the Chinese summary, I can translate these English paragraphs in the original text into Chinese and then they become part of my Chinese summary.
>
> **Extract 10.5** Peter Zhang

> ... The text had a very clear structure, starting from a general introduction paragraph and then writing about the details of each country within the framework of the introductory paragraph. The first few paragraphs could well serve as a summary... There were actually, it seemed to me, some kind of repetitions of the same content in the detailed descriptions of each country in the second part of the article...
>
> **Extract 10.6** Simon Zhang

> Yes, the text had a very clear structure. The histories of these eight countries are more or less the same... I think the structure of the text was extremely important to guide my summaries. I wrote the summaries at exactly the same structure as the original text and reduced the amount of details in order to meet the requirements of the length of a summary... If there were no such a structure, I would think the summarization task would be more difficult. And without the clear structure, the text might not be a text at all!
>
> **Extract 10.7** Wendy Zhang

However, students who summarized the other two texts seemed less fortunate in this respect. Only two interviewees from the textB group mentioned that its narrative order helpful for them to comprehend and summarize the source text. However, compared to the scaffolding provided by the textA structure, its helpfulness was far less evident and useful.

> This text is written in plain English, in a normal narrative order, and I think these to some extent make the text not very difficult to understand.
>
> **Extract 10.8** Alice Zhang

> The text was not too difficult to understand. Firstly, there were not many new words. Secondly, the text was not too theoretical or abstract. It had many supporting details and examples. Thirdly, it was a kind of news report, a kind of narrative with a clear thread ... I wrote the summaries in time order.
>
> **Extract 10.9** Ulysses Zhang

The students from the textC group seemed far less fortunate because the very different views on work life balance were diffuse and loosely organised, as one interviewee lamented:

> The text was somewhat difficult, because it was too long and involved too many different views on the issue of work life balance... The difficulty level of its vocabulary was acceptable, but it was hard to generalize or summarize those different views which were quite diffusing. And the structure of the text was not very clear, either. It was loosely organized.
>
> **Extract 10.10** Quentin Zhang

2) Vocabulary

Irrespective of whether the interviewees loved or hated the texts, or were proud of their vocabulary knowledge or felt the lack of it, vocabulary was brought up frequently, either as a scapegoat, an excuse for not being able to perform well, or a real source of difficulty. It seemed to be a natural tendency for students to attribute the difficulty or ease of a text to the vocabulary, though to varying degrees. TextA students generally valued the low frequency of new or unknown words in the source text; textB students thought the technical terms in the text somewhat affected their complete understanding, but did not necessarily have detrimental effects on their summarization performances because they could quite strategically avoid using these new technical terms in their summaries. This probably explains why textB students still had significantly higher EEHS and CEHS (see **10.1.2**). For textC students, vocabulary seemed less of an issue, compared to their comments on the length and topic knowledge of the text (see section **3** and **4** below).

a) TextA

Only two interviewees attributed the relatively high summarizability of a text to the fact that there were not many new words. They thought the low frequency of new words might to some extent have helped to make the text not too difficult to summarize, in addition to the main reason – the clear and friendly structure (see *Text Structure* above).

> ... Furthermore, I knew almost all the words in the text; there was no problem of understanding specific words, so the text was not difficult to understand.
>
> **Extract 10.11** Jack Zhang

> This text had no new words, no difficult sentence structure, I think, the text was not difficult.
>
> **Extract 10.12** Katie Zhang

However, one interviewee argued that vocabulary, compared to the clear and friendly structure of the text, was far less important in making the text not that difficult (see the first three lines of Extract 10.1).

Vocabulary and topic knowledge, however, were interwoven. Knowing a certain word might indicate that the students had some knowledge of the topic; on the other hand, having some topic knowledge of the text might help the students to gain some sense of the unknown words.

> I had never read educational history of South East Asian countries, but I know the geography of South East Asia. Say, I did not know the English word "Thailand", but I guessed what it was by using my knowledge of geography and the pronunciation of the word.
>
> **Extract 10.13** Wendy Zhang

b) TextB

Vocabulary seemed to be the most striking problem in textB. All the eight interviewees highlighted the issue of vocabulary, but with surprisingly contrasting views. For those (n=2) who considered the text difficult, they blamed the high frequency of new words to some extent, together with other features such as their lack of topic knowledge; however, one of them also suggested she could strategically "avoid this deficit when writing a summary." (Extract 10.15). For those (n=6) who did not consider it difficult, they thought it was partly because "the vocabulary was not difficult"; or even if there were new challenging technical terms, these new words

"did not have detrimental effects on [their] understanding of the *whole* text" (Extract 10.17); or they could also strategically omit those unknown technical terms in their summaries.

i) Two interviewees who considered the text challenging

Nancy Zhang, who considered the text difficult, commented that:

> The text was quite difficult to me because I was not familiar with such a topic and also because the vocabulary was challenging – there were lots of technical words I did not know... I should have read more such scientific articles...
>
> **Extract 10.14** Nancy Zhang

However, the other student who also considered the text difficult thought she could avoid using the unknown words in her summary.

> ... There are lots of new words I didn't know... The text was difficult to understand. I was not familiar with the topics... and I think this affected my understanding to some extent, but it was mainly because of the new words that made the text difficult to understand. ... However, I could avoid this *deficit* when writing a summary. Even without really understanding the new words, it was still possible to have a general understanding of the whole text and write a good summary.
>
> **Extract 10.15** Daniel Zhang

ii) Six interviewees who did not consider the text difficult

Similarly, one interviewee who did not consider the text difficult took a strategic approach to unknown technical terms, as did Daniel. She thought it was not necessary, though ideal, to "really know" these words to produce a short summary.

> It was true that the text was long and we may not really know some of the technical terms such as the names of the native plants and fish at the river system (Colorado River)... At first, I thought I must understand these technical terms to write a good summary, but then I realized it was not necessary to know these words to write a summary of around 300 words. You can actually omit these terms in your summaries, although it would be ideal if you understood them and include them as an example or two to support your generalization in the summaries. If you can't understand the terms, I do not think it would be a serious problem in understanding and summarizing the text as long as you can have a general understanding of the text. Some details can be omitted... If I have to mention what difficulty the text has, I think it is the technical terms that might cause some problem in understanding. But as I said earlier, we can omit these technical terms in our summaries...
>
> **Extract 10.16** Alice Zhang

Another three interviewees (Rachael, Ollie and Ben) acknowledged that there were some unknown words they did not fully understand, but the text was not difficult for

them. No specific reason was given by the first interviewee (Rachael). The second interviewee thought it was because the technical terms did not have detrimental effects on his understanding of the *whole* text.

> This text should not be considered difficult because, in fact, we also read more difficult articles in our study... It was easy to have a general understanding of the text, although there were some words that I did not know, such as those native fish and plants in the river system... These new terms did not have detrimental effects on my understanding of the *whole* text, though.
>
> **Extract 10.17** Ollie Zhang

However, the third interviewee thought that because enough time was given for the summarization tasks, the text was at the medium level of difficulty.

> I think the text is at the medium level of difficulty, but it does have some technical terms and the topic is not among what we usually study. We had enough time to do the tasks, so I think it is not difficult to understand...
>
> **Extract 10.18** Ben Zhang

The rest of the interviewees (n=2) claimed it was the fact that there were not many new words that rendered the text not too challenging.

> The text was at a medium to easy level of difficulty... The vocabulary was not difficult; I know the words without having to use dictionaries.
>
> **Extract 10.19** Michael Zhang

> The text was not too difficult to understand. Firstly, there were not many new words.
>
> **Extract 10.20** Ulysses Zhang

c) TextC

Compared to textB interviewees, textC interviewees were far less concerned with the vocabulary issue than the length of the source text (see 3 below), though three of them very briefly mentioned that there were words they did not know. Another interviewee was quite happy that there were not many new words, which she thought made the text not too difficult to understand.

3) Length

The lengths of source texts *per se* and also of the summaries were reviewed by the interviewees of textB and textC. No one mentioned the issue of the length of textA, whereas two interviewees (Alice and Ulysses) very briefly commented on this issue in relation to textB. One of them said that textB was "long and daunting". The other

interviewee thought it was not only the problem of the length of the source text, but also it was "very hard to write a summary within a certain word limit", giving an example:

> At first, I wrote a summary of about 100 words. When I realized the summaries had to be around 350 words, I added and added, and finally, the summaries exceeded well over 350 words, they were too long.
>
> **Extract 10.21** Ulysses Zhang

Strikingly, seven out of eight interviewees who had undertaken textC complained quite extensively that it was mainly the length of the source text *per se* that had made it very difficult, or at least seemingly challenging at first glance.

> At first glance, the text seemed too long, it gave a feeling that the text can also be very difficult to understand and summarize. ... However, when I read it, it was not as difficult as it seemed to be at first glance. ... The length of the text, however, did affect how I read it.
>
> **Extract 10.22** Cindy Zhang

Some thought the length of the source text exerted its effects only when they were half way through reading the text. They got tired, and it "seemed hard to keep on reading" (Grace Zhang). However, these effects were mainly psychological or emotional, rather than practical.

> The text was difficult mainly because it was too long. This was largely a psychological or emotional effect. At first, it was OK, but when you read on, you felt tired. All in all, it was because the text was too long, though the text *per se* was not that difficult to understand...
>
> **Extract 10.23** Elyn Zhang

The last straw on the lengthy text came from two sources – the loosely organised and diffuse views, as mentioned above, and unknown vocabulary.

> The text was somewhat difficult, because it was too long and involved too many different views on the issue of work life balance...
>
> **Extract 10.24** Quentin Zhang

> The text was difficult to understand or summarize. It was too long, and there were some words I did not know. I sort of got impatient...
>
> **Extract 10.25** Yvonne Zhang

Some interviewees also tried to find the reasons from their own experience in doing reading comprehension tests. As they commented, they "used to have very short texts in reading comprehension tests" (Fred Zhang), and it was probably the first time they had been asked to read extended texts like textC for test purposes. Comparatively speaking, they found the summarization tasks in this project quite challenging.

> We usually have short texts for reading comprehension tests; and very often you could find the answers directly from the source texts. Quite easy.
>
> **Extract 10.26** Grace Zhang

4) Topic knowledge

Although topic knowledge seemed to have affected students' judgements of the difficulty levels of the source texts, it did not have statistically significant effects on their actual summarization performances (see **10.2.1**). The interview data generally supported the findings from the statistical analyses. Topic knowledge seemed much less instrumental to their summarization performances than text structure, vocabulary and length for textA, textB and textC respectively. Furthermore, the differences in the views on topic familiarity and summarization performances between students from different source texts were far less clear-cut than the other three features of texts' summarizability (structure, vocabulary and length) discussed above. The interviewees tended to focus more on the relationship between familiarity and *understanding*, than on *summarizing*.

a) Potential helpfulness of actual or assumed topic familiarity

Only four interviewees thought that their actual or assumed familiarity with the topics might help with their understanding of the texts. Taking Louis for example, he commented that:

> It was probably because I was familiar with such a topic, the text was not difficult to understand… After all, we know more or less about educational history…
>
> **Extract 10.27** Louis Zhang

The second interviewee thought her familiarity with the topic (textB) was "probably" essential, as she commented:

> I once read a similar article about flooding the Nile. At first, I thought it was the same article, but then I realized, no, it was not. However, the principles of both flooding were quite the same… This kind of familiarity with the topic was helpful for me to understand the text (let the river run); otherwise, I would probably never have been able to understand it at all.
>
> **Extract 10.28** Rachael Zhang

The third interviewee assumed that if she had been familiar with the topics (work-life balance) it would have been helpful.

> I was not too familiar with the topic, but I heard of the topic… If I had already known the topic, it would be helpful in understanding the text. I

> could use what I know about the topic in my summarization of the text.
>
> **Extract 10.29** Tom Zhang

However, this kind of assumed helpfulness may be only psychological rather than practical, as the fourth interviewee commented: "*if I had known the topic, it would have made me more confident and comfortable in producing the summaries*". He also acknowledged that:

> Although I was not familiar with the topic of the text [educational history], I don't think it would affect my understanding to a great extent, as long as you could understand the text at first instance.
>
> **Extract 10.30** Simon Zhang

b) Unfamiliarity and additional processing time

Only two interviewees complained that their lack of familiarity with the topics of the source texts may have had some effects on their understanding of the texts. As Ollie commented, due to his lack of familiarity with the topics he might have required additional time to adjust himself to the new topics, which would lead to additional processing time for the summarization tasks.

> I was not familiar with the topic of the text – controlled flooding and this to some extent may have affected my understanding of the text. After all, it was the first time for me to read such a topic, and this *strangeness* may have required additional time for me to adjust myself to the topic, and may have had some effects...
>
> **Extract 10.31** Ollie Zhang

The other interviewee linked the issue of familiarity to vocabulary (see also Wendy's view in Extract 10.13), as she lamented:

> I felt the text was quite challenging. I do not know much about this topic; I do not read texts of such topics. When you are writing the Chinese summary, you have to know the Chinese names of those South East Asian countries.
>
> **Extract 10.32** Helen Zhang

c) (Lack of) familiarity and its dispensable role in comprehension and summarization

For most of the interviewees, whether they were familiar with the topics or not did not seem that important. Topic familiarity was not considered indispensable for understanding and summarizing the source texts (see also Extract 10.30).

> I was more familiar with the topic – educational history – than scientific topics, but I had never read articles on educational history of South East Asian countries. However, I don't think this unfamiliarity with the specific topic – educational history of South East Asian countries – would affect my understanding of the text…
>
> **Extract 10.33** Peter Zhang

> Whether you are familiar with the topic [work life balance] or not does not really matter, it is your understanding of the text that would affect your summarization.
>
> **Extract 10.34** Elyn Zhang

According to the PSQ data, there was not much difference in students' judgment of topic familiarity between the text groups (see **10.2.1**), but the interview data seemed to suggest that they were more in tune with the topics of textA (educational history) and textC (work-life balance) than that of textB (controlled flooding).

> I was not familiar with the topic in the text, but such a topic was not too strange to me even if I had never read such topics. You know, in Chinese, we also talk about the balance between work and life, things like work and play, "*lao yi jie he*" (a Chinese phrase which means you can't work or play all day or all the time and there should some balance between work and rest).
>
> **Extract 10.35** Yvonne Zhang

In summary, all the three texts were considered moderately difficult to understand and summarize, but significantly more challenging to summarize than understand. The overall difficulty level in *understanding* between the three texts was statistically significant, but not in *summarizing*. TextC was rated the most challenging; and textA the easiest and the most familiar (in the specific topics) to understand. In relation to the judgment of the difficulty levels of the source texts, students also held that their familiarity with the general and specific topics of the texts may have been moderately facilitative in both understanding and summarizing, but significantly more helpful for understanding than for summarizing the texts, a view endorsed by all the three groups of students. TextA, which had the lowest vocabulary density (see Table 4.5), was also considered notably friendlier to summarization because of its clear structure and organisation. Apart from this most instrumental feature for summarization performance (i.e. text structure and organisation), text length and the frequency of new words, and topic familiarity all seemed to have played some roles in students' judgements of the

difficulty in understanding and summarizing the source texts, as demonstrated in the interview data. All interviewees also thought the length of the texts made summarization particularly challenging. Although new words might present some challenges to understanding, it was very likely that the students could manage to avoid using the particular new or difficult words in their summaries, as they could substitute other simpler words or simply not use them at all. Their familiarity with the topics of the source texts was considered helpful, but not essential in producing a good summary. The interview data seemed to suggest that no single feature of a text could account for the dynamics of its summarizability, nor in a conglomeration did the same features work for all the three texts. It was rather the most striking feature of a text that could probably significantly affect the text's summarizability. However, the statistical analyses on the relationships between the judgments of text difficulty and topic familiarity (two key elements of summarizability) and students' actual summarization performances did not seem to fully support such a potential link as perceived by the students.

10.3 Summary of findings relating to RQ5

This research question examined the possible effects of text type on summarization performances from two perspectives: students' actual performances and their perceptions of such effects. It was found that text type had significant main effects on the three quality indicators of summarization performances and that the effect sizes were much bigger on Chinese than English summarization performances. Furthermore, text type had significant interaction effects on the quality of Chinese summarization performances with other factors such as language order (see also Chapter 8) and presentation mode (see also Chapter 9). Although these kinds of interaction effects was not evident in the post-summarization questionnaire or interviews, due to the nature of such data elicitation methods, the findings from the statistical analyses on the performance data were *generally* supported by the perception data which also provided further nuances of the dynamics of summarizability between the source texts and how they might have contributed to the significant statistical effects on summarization performances of *text type*. The students held that text structure, frequency of new words, topic familiarity and

length of source texts were among the most influential elements of summarizability. However, the perceived helpfulness of topic familiarity was more of a common-sense-driven understanding; this lacked substantive corroboration from the results of the statistical analyses on the means differences in summarization performances between low and high topic familiarity students.

Further discussion of these findings is reported in **11.2.2.1**.

PART V

CHAPTER ELEVEN

Discussion and Implications

The previous five chapters (6-10) in Part IV presented the main findings in the order of the five research questions. The current part consists of two chapters (11 and 12). Chapter 11 first of all presents an overview of the project and discusses the key research findings in the IFOE framework of using summarization as a measure of reading comprehension and the implications for language testing and language teaching. Chapter 12 lists several limitations of the project and puts forward a series of suggestions for exploring and developing the IFOE framework (Figure 2.1).

11.1 Overview of research

The main purpose of this project was to explore the use of summary writing as a measure of reading comprehension, within the proposed IFOE framework (see Figure 2.1). Five research questions addressed each component of the framework: *input*, *filter plant*, *output* and *evaluation* (see Table 4.1).

RQ1: *What are the differences in score variances and students' attitudes between using expert and popular templates to evaluate written summaries?*

RQ2: *Are students' summarization performances affected by their other linguistic abilities and if so, to what extent?*

RQ3: *What impact does the use of a different language and language order have on summarization performances and measurement of reading comprehension abilities?*

RQ4: *What are the effects of text presentation mode and students' computer familiarity on their summarization performances?*

RQ5: *What are the effects of text type on students' summarization performances?*

The main experiments – the summarization tasks – employed a factorial design of 3 text types x 2 text presentation modes (*computer* versus *print*) x 2 language orders (*English then Chinese* versus *Chinese then English*). The students produced both English and Chinese summaries of extended English texts (around 2,200 words each) within a specified time and word limit. Immediately following the summarization tasks, the post-summarization questionnaire (see Appendix 4) was administered to all the students to elicit their perceptions of the summarization tasks (such as text difficulty, topic familiarity, preferences to language and language order) and the relationship between their summarization performance and other language abilities. Twenty-four students randomly selected, two from each of the twelve summarization conditions, were further interviewed individually to gain greater understanding of their perceptions of the summarization tasks.

Besides the main experiments, baseline data on the students' (N=168) language abilities were collected through (i) analyzing their performance in the previous national test - *Test for English Majors*, and (ii) administering several other measures including the reading paper/section of FCE and TOEFL, English and Chinese writing tasks, and passage translation (English to Chinese). In addition, the students' evaluations of their computer familiarity were collected through the CFQ (see Appendix 1).

Five well-educated English native speakers were invited to write summaries of the three texts and their summaries were used to generate the expert scoring templates. The students' summaries were evaluated according to both "expert" and "popular" scoring templates (the latter generated from all the students summaries *per se*). *RSC*, *HS* and *Length* were the major quality indicators of summarization performance.

Therefore two parallel datasets – performance and perception – were collected. Data analyses involved (i) statistically modelling students' *actual* performance in

relation to the factors of the IFOE framework, using exploratory factor analyses, general linear models, and non-parametric tests where appropriate and (ii) enriching the statistical models with further insights from the students' own *perceptions* of the summarization tasks, using a qualitative data analysis computer programme – winMAX (Kuckartz 1998) – to code and develop categories.

11.2 Summaries of key findings and discussion

At the end of Chapters 6-10, summaries of the main findings for each research question were presented. In the following section, I will highlight the key themes from the "summaries" of the main findings, followed by discussion of the findings and the emerging issues in using summarization tasks as a measure of reading comprehension (in the order of RQ1 → RQ4 & 5 → RQ2 → RQ3) within the IFOE framework (see Figure 2.1).

11.2.1 Evaluation of students' summarization performances

1) Summary of key findings

RQ1: *What are the differences in score variances and students' attitudes between using expert and popular templates to evaluate the students' summaries?*

There were significant main effects of scoring templates on the two main quality indicators (*RSC*, *HS*) of both English and Chinese summarisations, in the absence of any interactive effects between the scoring templates and other factors such as *text type* and levels of students' reading comprehension abilities. A summary was assigned significantly higher RSC and HS when judged according to the popular template than when evaluated according to the expert template. The effect size was much larger on RSC of the English summaries than of the Chinese summaries. However, there was not much difference in effect sizes for HS between English and Chinese summaries. Furthermore, the effect sizes on RSC were much greater than those on HS of the

English summaries; for the Chinese summaries, the effect sizes on RSC and HS were approximately at the same level. It was found that scores derived from the expert template could better predict students' performances in standardized reading comprehension tests. In terms of students' views on the use of the two scoring templates, the majority of the interviewees strongly preferred the use of expert over popular templates because of the long established supremacy of experts over novice students and the tradition of using experts to judge students' performances. It seemed that the involvement of students in generating assessment criteria was not as welcome as I had expected. To some test takers, the use of the popular templates appeared to be a form of imposed democracy rather than empowerment.

In the following section, I will discuss the issues of scoring reliability and students' voices in the development of assessment criteria.

2) Issues in scoring reliability

The inevitable subjectivity in marking summaries has been perhaps the thorniest issue in terms of using summarization tasks in language testing research and practice, as commented on by Weir (1993) and Alderson (Alderson 1996: 225; Alderson *et al.* 1995), albeit without much empirical evidence to support their views. Contrary to their claims, this project has achieved respectably high scoring reliability (see **5.5.1**) through using an augmentation method, an encouraging aspect of this study. However, some subjectivity remained and several other issues also emerged.

Generating the key statements from a number of experts' and students' summaries was probably as time-consuming as analysing the three source texts using the summarization models (Bernhardt 1991: 202-203). Although the features of winMAX helped to a degree, it was still a laborious process involving subjective judgement on my part and that of the assistant who helped with the coding of all the statements in the summaries. It is hoped that automatic summarization programmes in

the future may be a way forward to generate the "ideal" summary as a scoring template (see **2.3.2**).

Within the overall picture of high inter-rater reliability, it was noticed that the rating performances varied across (*a*) the three raters, (*b*) the three types of source texts and (*c*) the scoring criteria (RSC, HS). Such variances were very likely to have been influenced by the quality of the summaries *per se* which were affected by the summarizability of the source texts (see Chapter 10). The inter-rater reliability was higher in the case of higher quality summaries of easier texts (e.g. textA, see Appendix 16 for the detailed report of the reliability analyses). In addition, the inter-rater reliability in RSC was slightly higher than that of HS. This may be due to the fact that raters had more "rigid" guidelines to follow when assigning RSC than when assigning HS (see Appendices 12 & 13).

Although the inter-rater reliability of the English and Chinese summaries was at the same level (see **5.5.1**), findings from the analyses on the effects of language and language order on summarization performances (see Chapter 8) indicated that the rating performances may have been interactively affected by the language factor. The English summaries were consistently assigned higher RSC and HS scores. Apart from the significant language effects on summarization performance, some other explanations may be plausible (see also **11.2.2**). The Chinese raters may have taken different approaches to the English and Chinese summaries. It was probably easier for the Chinese raters to detect any misunderstanding by the students in the Chinese summaries than in the English summaries. In contrast, in the English summaries, the Chinese raters could be "fooled" to some extent by the seemingly correct responses that might have been simply lifted from the English source texts. At the same time, the Chinese raters might be stricter in marking the Chinese summaries than the English, probably because they would assume the summarizers should have no problem expressing themselves in Chinese – the summarizers' first language. As foreign language learners of English themselves, the three Chinese raters may have been more

sympathetic and lenient when marking the English summaries. However, further research is needed to investigate these assumptions (see also Chapter 12), as Cohen (1994: 203) rightly pointed out:

> Another issue to explore, in the case of cross-language summaries, is whether or not the rater is a native speaker of the language in which the summary is written. For example, to what extent might raters who are non-native users of the language of the summaries focus on the reading comprehension side rather than on the writing because they do not consider themselves able to judge the merits of a summary written in what is to them a second language? Would the focus of native-language readers be different?

It would be of interest to investigate the rating performances of markers of different language background and proficiency (see **12.2**).

These complexities in judging the quality of summarization performance resonate with Weir's (1993: 154) concern regarding the subjectivity of marking written summaries (see **2.5.1**). However, this project also demonstrated that high scoring reliability was not unachievable.

3) Students' voices and the development of assessment criteria

Three main sources in the research literature that stimulated the comparative study on the use of the two scoring templates are: (i) the postmodernism ontology of difference and individualization and its implications for reading comprehension tests; (ii) summarizers' individual characteristics that contribute to their summarization performances; (iii) the unequal status of native speaker experts and test takers, and the potential value added by the involvement of test takers in development of assessment criteria.

I appreciate that postmodernist interpretations of a text should invite different voices to be heard (see **3.2.1**) and that test-takers' involvement in the development of a

language test would be desirable (see **2.5.1**). However, the use of a popular scoring template in this project unfortunately appeared to constitute a form of imposed "democracy" on the students who were accustomed to the common practice of using experts as *de facto* authoritative assessment criteria, and were more than willing to maintain the current practices. Although, as expected, there were qualitative differences in the summaries produced by the experts and the students (e.g. vocabulary density, see **5.5.2**), the similarities between them cannot be ignored: around 50% of the key statements from the expert and the popular templates overlapped (see Appendix 12). On the other hand, the differences, e.g. in vocabulary density and content coverage of the summaries written by the native speaker experts were substantial. Their agreement on which ideas were essential to the construction of a meaningful summary was lower than that among the students. This was probably due to the fact that only a limited number of experts were invited to participate (n=5), in contrast to the number of student summaries generated (over 150). However, this challenges the use of expert judgments alone in the development of assessment criteria. As an incidental finding of this project, the substantial difference between expert summaries echoes Cohen's (1993:137) observation that "even the experts did not fully agree on which ideas were essential to the construction of a meaningful summary".

As demonstrated in this research, the assessment criteria derived from the summaries of native speaker experts and non-native speaker test takers had significant differential effects on the scores that the non-native speakers' written performance could be assigned. This finding however may be sample-specific, because both the expert and the popular templates were empirically derived from a limited number of written samples. Turner and Upshur (2002) noticed that "using different samples of performance in scale development yielded different rating systems" (p.65), although this did not make significant differences to the scores that *other* students' written performances would be assigned. Turner and Upshur's study is fundamentally different from the current project in terms of the object to be evaluated. In their study,

the samples which were used to empirically develop the assessment criterion was not evaluated by that criterion; while in the current project the samples (all the students' summaries) were later judged according to the criteria which were developed from the same samples.

One of the motives for using a popular scoring template was to invite the active involvement of test-takers, so that their legitimate voices or understandings could be heard in the development of assessment criteria in order to improve test validity and ethical impact as hypothesized by Bachman and Palmer (1996: 32). However, it seemed that the majority of the student participants in this research strongly preferred the use of expert templates, because of (i) their perceived inferior experience and English language and summarization abilities, compared to English native-speaker experts, and (ii) stereotypical status and the common practice of using experts to create "*biao zhun da an*" (the standard answer) in educational assessment. It seemed that the majority of the students interviewed were so used to and ready to accept the common practice of using experts' high and authoritative standards to evaluate their performances that they doubted their abilities to contribute to the assessment criteria. Even the pro-popular students were shocked to learn that their summaries were to be evaluated according to the criteria empirically derived from their own samples. These findings are much in line with Davies's (2003) argument that native speaker membership is determined by the non-native speaker's willingness to assume confidence and identity. This power relationship is enhanced prominently by the students' willingness to identify themselves as foreign language learners of English and novice summarizers of comparatively far less expertise and knowledge. This kind of identity formation in a process of identification and self-identification (Wenger 1998) is probably rooted in the students educational and assessment cultures or common practices. As lower-status entities in the hierarchical structure of the power relations in educational assessment, the students did not yet seem ready to appreciate and take up the opportunity and indeed responsibility to contribute to the development of an empirically derived assessment criterion, although there was some evidence of

awareness of potential benefits on the part of some students.

The empirical development of assessment criteria in the current research also raised the issue of the practicality of involving test takers, as proposed by Bachman and Palmer (1996). The intended positive effects on teaching and learning may not be easily achieved if there is resistance to such involvement from test takers in the first place. It seems imperative to address the unequal power relationship between test constructors and test takers before developing empirically derived assessment criteria. However, the findings of this research are probably bounded to the specific educational assessment culture and practice where the connotation of the term "expert" (*zhuan jia*) endows it with supremacy over "student". If the intended outcomes such as increased validity and ethicality in language testing are to be achieved in this educational assessment culture, there seems to be a need to develop the proper "soil" for implementing the democracy of language testing and for realizing the hidden values of test takers in the development of assessment criteria (see **2.5.1**).

Further research is needed to investigate whether acceptance of (or resistance to) such democracy is affected by the students' English language proficiencies, whether students from different first language and cultural backgrounds have different views on the conceptualization of democracy in language test development, and whether and to what extent "indigenous assessment criteria" (Douglas 2001; Jacoby & McNamara 1999) could be developed by test takers themselves in different assessment contexts (but see Douglas & Myers 2000) and extended to students' self-assessment for learning.

11.2.2 Dynamics of summarization

The previous section discussed the evaluation of summarization performances. In this section, the focus will be on the dynamics of summarization performances *per se*,

which were influenced to various extents by multiple components such as *text input* (RQ4 & RQ5), *filter plant* (RQ2) and language and language order of *output* (RQ3).

1) Text input

The effects of two key features of *text input* – text presentation mode and text type – were studied in RQ4 and RQ5 respectively. In close relation to these two features of *text input*, students' computer familiarity (RQ4) and topic familiarity (RQ5) – essential components of *filter plant* – were also discussed in this section. For the effects of other *filter plant*, please see **2**) below.

a) Text presentation mode and computer familiarity

i) Summary of key findings

RQ4: *What are the effects of text presentation mode and students' computer familiarity on summarization performances?*

Text presentation mode significantly affected the lengths of the Chinese summaries. The Chinese summaries were substantially longer if the source texts were presented through computers. Text presentation mode also had significant interaction effects with *text type* on two quality indicators of Chinese summaries (CERSC and CEHS). The difference in textC summaries was larger than the other; while the difference in textA summaries was negligible. No other effect of text presentation mode was significant. It seemed that the effects of text presentation mode were more pronounced in Chinese than English summarization. These findings therefore only partly supported the first part of the research hypothesis (RH4, see **4.1.2**) that text presentation mode would make differential effects on summarization performances.

Similarly, the impacts of computer familiarity were also more pronounced on Chinese than English summarization performances. Students' computer familiarity had significant main effects on the four quality indicators of Chinese summarization (CERSC, CPRSC, CEHS and CPHS), while there was no such significant main effect

on English summarization performances. High computer familiarity students had significantly poorer Chinese summarization performances than their low computer familiarity counterparts. In addition, computer familiarity was also found to have significant interaction effects with *text type* on CEHS, and with *language order* on CERSC, EERSC and EEHS. No other effect of computer familiarity on summarization performances was statistically significant. Generally speaking, these findings did not confirm the second part of the research hypothesis that high computer familiarity students would find it easier to summarize computer presented texts than their low computer familiarity counterparts if their summarization abilities were held constant. In view of the fact that there were also significant interaction effects of computer familiarity with *text type* and *language order* on various quality indicators of summarization performances (see above), the effects of computer familiarity seemed much more complex than assumed (see also **4** below).

However, the findings from the statistical analyses of the actual summarization performances were not fully supported by the perception data (i.e. post-summarization questionnaire and interviews). Although the students maintained that there were several minute physical and psychological impacts of text presentation mode and computer familiarity on their summarization performances, such impacts were probably more *expected* than actually *experienced.*

ii) Further discussion

As demonstrated above, text presentation mode had significant main effects on the lengths of the Chinese summaries, but not the English summaries. The availability of the *copy* and *paste* functions in Word provided the students in the computer room with a good chance of producing longer drafts of English summaries on computer than on paper in the earlier stages of the summarization process. On the other hand, the *delete* function in Word also made it far easier for the students to keep their English summaries within the word limit, as required in the task directions (see

Appendix 3), in the final stage. Those students who did the summarization tasks in the normal classroom were very likely to be economical in producing summaries; they might have endeavoured to keep their drafts within the word limit from the very beginning of the summarization process. It is possible that the use of these facilitative functions in MS Word may have mitigated the effects of text presentation mode on the lengths of the English summaries.

However, the Chinese summaries of computer-presented texts were significantly longer than those of paper-presented texts. When producing summaries in Chinese, the students may have employed different strategies (see also **3** below: *output*). As the final summaries were all required to be presented on paper, no students opted to write the Chinese summaries on computer[1]. In fact, compared to direct copying and pasting from the source English texts, typing Chinese is onerous. Therefore, text presentation mode might have had a more direct impact on the lengths of the Chinese summaries than the English. Although it was less clear why the Chinese summaries of computer-presented source texts were significantly longer than those of paper-presented texts, it seemed evident that the use of computers, either as a presentation mode or a medium for producing the English summaries, had some differential effects on the lengths of both the English and Chinese summaries, though in a very different pattern. The use of computers in drafting the English summaries *mitigated* the effects of text presentation mode, while writing the Chinese summaries on paper may have *magnified* the effects of presentation mode on length.

Text presentation mode also had significant interaction effects with *text type* on CERSC and CEHS. The mean differences in CERSC and CEHS of summaries of textC, considered by the students to be the most challenging (see Table 10.2), were notably larger than those of the other source texts. When a text was easier to summarize, the effects of text presentation mode were less significant than when it

[1] Although this research was not designed to collect students' summarization processes, some field notes were made throughout the project, for example, during the summarization tasks.

was more challenging. The higher summarizability of the source texts counterbalanced the effects of text presentation mode to some extent. On the other hand, a more challenging source text could increase the effects of text presentation mode. This interplay between the summarizability of source texts and presentation mode may also have been coupled with the degree of the students' willingness and efforts to do their utmost in the summarization tasks, especially when they were summarizing the texts in Chinese, a task generally considered more challenging than English summarization (see **8.2**).

Apart from the significant main effects on the lengths of Chinese summaries and the two interaction effects with text type on CERSC and CEHS as discussed above, text presentation mode did not make any other substantial difference to the students' summarization performances. However, these findings should take into consideration the context of the project: (a) the student participants had a high level of computer familiarity (see **5.1.4**, c.f. O'Sullivan *et al*. 2004), (b) the summarization tasks did not require a substantial amount of computer manipulation skills (see **9.2.2**), and (c) the summaries were actually written on a piece of paper, which might have further mitigated the requirement for computer skills, such as the speed of typing Chinese.

As noted above, in considering the analyses of the effects of text presentation mode the students' familiarity with using computers needs to be taken into account. Contrary to expectation, the students with low computer familiarity produced much better Chinese summaries than their counterparts with high computer familiarity. The research design of this project was unfortunately not able to explore this issue further. One speculation may be that high computer familiarity students might have spent more time manipulating the source texts on the computer screen, which meant that they would have less time to write the Chinese summaries on paper. On the other hand, those students who were less confident in using computers might have used their basic skills as required, been less adventurous and saved more time for actually writing the Chinese summaries. In view of the fact that the time for the summarization tasks was

limited, the amount of time available to write the summaries on paper may have been crucial in determining the quality of the Chinese summaries produced.

To the best of my knowledge, no other research study in the field of language assessment has compared the differences in summarization of computer- and paper-presented source texts. Dyson and Haselgrove (2000), looking at general reading comprehension, posited that there was more speed-accuracy trade-off in understanding details than main idea comprehension when speed was an essential requirement of screen-reading. Following their arguments, it seemed that summarization tasks would be less affected by time constraints. However, the evaluation of the quality of a summary in this project was not only based on its accuracy (e.g. RSC) but also its overall quality (HS), incorporating various other quality indicators (see **4.2.4**). Another fundamental difference between Dyson and Haselgrove's study (2000) and this project was that their main idea comprehension tasks did not involve as much writing as this project. These factors together make the findings of the two projects less comparable.

The qualitative data generated from the post-summarization questionnaire and interviews further attested to the complexity of the effects on summarization performances of text presentation mode and computer familiarity. None of the interviewees considered that the potential disadvantages of screen-reading and the possible deficiency in computer familiarity would be detrimental to their summarization performances. What is more, it seemed that these effects were more likely to be expressed and anticipated than actually experienced. These may well be because of the high computer familiarity of the students in this project. However, the data also suggested that it is important to take into account test takers' previous experience of using computers when designing computer-based summarization tasks, particularly when the tasks would require a substantial amount of computer manipulation skills. Test takers' experiences of using computers and their perceptions of (*a*) the historical friendliness, tangibility and security of reading on paper and (*b*)

the physical and psychological fatigue of screen reading (see also Alderson 2000) might affect their readiness to take up such summarization tasks and potentially their performances.

b) Text type

i) Summary of key findings

RQ5: *What are the effects of text type on students' summarization performance?*

The statistical analyses of the students' summarization performance found that *text type* had significant main effects on the three quality indicators (*RSC*, *HS* and *Length*) of both English and Chinese summaries, and that such effects were greater on Chinese than English summarization performance. Text type almost had a similar amount of impact on summarization performance as the students' reading abilities measured by TOEFL. Several significant interaction effects of text type with *language order* and *text presentation mode* were also noted in the students' Chinese summarization performance. The qualitative data from the post-summarization questionnaire and interviews established further evidence on (*a*) the dynamics of the summarizability of a source text and (*b*) how a unique feature of a text might have affected the students' summarization idiosyncratically. Readability of a text and students' familiarity with its topics – the common measures of the difficulty of a text – seemed less indicative of its summarizability than a certain prominent feature of the text, for example, its vocabulary density and whether it had a clear structure and a summative introductory paragraph.

ii) Further discussion

The findings from the statistical analyses on the relationship between text type and summarization performance, however, could be an artefact of the research design *per se*. The allocation of *text type* to a participant was closely related to which *class* s/he was originally from (see Table 4.6). Therefore it was possible that the significant

difference between text types could also be a manifestation of the existing difference in summarization abilities of the students between different classes. A series of one-way analyses of variances, using *class* as the between-subjects factor, were conducted to investigate whether the significant difference in students' summarization performance was attributable to *text type* or *class*[2]. The findings of these analyses are summarized below:

- **English summaries**: **(i)** although there were significant differences in EERSC ($F_{3,96}$=3.223, sig.<0.0265) and EEHS ($F_{5,151}$=2.449, sig.<0.0365) between *classes*, the post-hoc Scheffe tests found that the classes formed a homogeneous group; **(ii)** there were no significant differences in EPRSC ($F_{3,96}$=1.172, n.s.) and EPHS ($F_{3,96}$=0.797, n.s.) between *classes*, which demonstrated that the classes also formed a homogenous group, according to these two scores; **(iii)** in terms of the lengths of the English summaries, the classes also formed a homogeneous group ($F_{5,146}$=1.224, n.s.).

- **Chinese summaries**: **(i)** there were significant differences between classes in CERSC ($F_{3,96}$=7.303, sig.<0.0005), CPRSC ($F_{3,96}$=6.77, sig.<0.0005), CEHS ($F_{5,151}$=5.579, sig.<0.0005) and CPHS ($F_{3,96}$=2.802, sig.<0.0445); however, **(ii)** in terms of CPHS, the post-hoc Scheffe test indicated that the classes formed a homogeneous group; in terms of the other three scores (CERSC, CPRSC, CEHS), two sub-groups were noted. In all the three scores, Class33 was always significantly different from Class32 and Class41 (for CERSC and CPRSC), and from Class31, Class41, and Class42 (for CEHS). These significant differences, however, were always between different *text types* rather than within a particular *text type* (see Table 4.6). Therefore, it is reasonable to argue that the significant differences in these three scores between the classes were manifestations of the significant differential effects of *text type*. **(iii)** In terms of the lengths of Chinese summaries, there was a significant difference between classes ($F_{5,149}$=5.51,

[2] These analyses are reported here rather than in Chapter 10 to avoid distracting the reader from the key focus on the analyses of the effects of text type.

sig.<0.0005). The post-hoc Scheffe test indicated three sub-groups. Within each sub-group, two classes had the same *text type*, and the other classes were from a different *text type*, which to some extent indicated that the differences between classes might also be affected by *text type*.

The analyses above allayed concerns that the artefact of the research design unduly confounded the effects of text type on summarization performance and also helped to present a fuller picture of the effects of *text type*. Although it was possible that *class* might be responsible for the differences as detected between *text types* to some extent, it was reasonable to argue that the main findings reported in Table 10.1 demonstrated mainly the effects of *text type*, rather than *class*, on students' summarization performances. The following paragraphs discuss the contribution to summarization performances of various features of source texts, such as their (i) macro-organisation, vocabulary density and readability and (ii) summarizers' topic familiarity and vocabulary knowledge, and the implications for the selection of source texts for designing summarization tasks.

The macro-organisational features (see Table 4.4), vocabulary density, readability and percentage of passivisation (see Table 4.5) of the source texts may all have affected the students' judgement of the texts' summarizability, and made differential effects on their actual summarization performances, to varying degrees. However, it seemed that the significant differences in the students' summarization performances were more likely to be affected idiosyncratically by certain prominent features of the source texts, combined or alone, as demonstrated in the post-summarization interviews. For example, the well-structured organisation and the summative introductory paragraph, clear narrative timeline and low vocabulary density of textA rendered it the easiest to summarize. Although textB was also presented along a timeline, its "narrativity" (Giora and Shen 1994) seemed less detectable and therefore probably less helpful for the summarization tasks. The challenges of textC mainly derived from the lack of clarity in the development of arguments which represented

conflicting views on the work-life balance scheme among different parties involved (see Appendix 2). The students found it difficult to follow the arguments in this extended text. Its high vocabulary density may have added to the challenges of summarizing this text. It seemed that (*a*) all these factors were idiosyncratic in terms of their effects on the summarizability of the source texts and (*b*) one single prominent element of the features of the texts may have made a huge difference, by chance, on the summarizability of the source texts. However, it was not clear which feature might account for more of the dynamics of the texts' summarizability than another, although it seemed that the macro-organisational features such as the presence of a summative introductory paragraph in a source text were most instrumental to summarization tasks. This finding is in line with research studies on the helpfulness of such paragraphs for summarization, recall and main idea comprehension (e.g. Garner & McCaleb 1985; Lorch & Lorch 1986; see **2.5.2**). In the field of language testing, Huhta and Randell (1996: 100) hypothesized that such effects of macro-organisational features presented challenges in the selection of source texts for summarization tasks:

> The selection of a text has an obvious effect on how easily a summary can be constructed. If the text has an opening paragraph that in fact summarizes the text, ... it appears to be relatively easy to select the right summary. It would probably also be quite easy to write a good summary of such a text.

The common measures of the difficulty level of a text for reading comprehension such as its readability and percentage of passivisation were less illuminating on the summarizability of the text than its vocabulary density. As noted in Table 4.5, the three source texts had quite similar readability using the F-K Grade level, but very different summarizability judged by the students. Contrary to the common assumption that a higher percentage of passive sentences in a text make it more challenging to understand (Namukwai & Williams 1988, cited in Clapham 1996: 94), textC, although containing the lowest percentage of passive sentences, was considered the most challenging. On the other hand, the gradual *increase* in vocabulary density in the three

texts (textA=84.76, textB=92.20, textC=115.90) corresponded with the *decrease* in the texts' summarizability. Common sense would dictate that the higher the vocabulary density of a text, the more information can be packed in when the number of words is held constant, and the more challenges it could present in terms of summarization tasks because summarization fundamentally involves condensing information. It seems that the readability of a text is not necessarily synonymous with its summarizability. On the other hand, vocabulary density may well be.

Taking the aspect of vocabulary, this research also found that the summarizers' insufficient grasp of certain words in the source texts did not seem to be detrimental to their summarization performance, because the employment of certain strategies, such as avoiding using unfamiliar words in the summaries, may have compensated for their insufficient grasp of the foreign language vocabulary (c.f. Cohen 1994). Furthermore, unfamiliarity with certain words may not be able to hinder the students' general understanding of the texts and their summarization (see **10.2.2**). This complex interaction between vocabulary density of source texts and students' smart use of strategies to compensate for their lack of understanding of certain words may indicate a fundamental difference between summarization tasks and multiple choice questions focusing specifically on the literal understanding of particular words.

In close relation to the innate features of a source text such as its vocabulary density, organisation and readability as discussed above, students' familiarity with the topics of the text was hypothesized to be able to affect their summarization performances significantly. However, this hypothesis was not fully supported. Familiarity with the *general* topics of the source texts was not considered facilitative either to understanding or summarizing the texts. On the other hand, familiarity with the *specific* topics of the texts was considered helpful for understanding but not summarizing the source texts. As comprehension is the prerequisite for summarization, it seemed illogical to conclude that topic familiarity would not affect summarization. Familiarity with the specific topics may have exerted indirect impacts on

summarization. This finding was in contradiction to the significant correlations between topic familiarity and summarization performance found in studies such as Afflerbach (1990), Hahn and Smith (1986), Kiewit (1997) and Kintsch & Greene (1978). At the same time, the indirect impacts of topic familiarity supported to some extent Swoope and Johnson's (1988) and Carrell's (1983) claims that prior knowledge did not have significant effects on written summarization and recall performances respectively. The differential contributions of topic familiarity to general reading comprehension (e.g. Clapham 1996) and summarization tasks (e.g. Carrell 1983; Swoope & Johnson 1988) imply that it is necessary for test designers to (*a*) take different approaches to the selection of reading passages when considering the possible effects of topic familiarity and (*b*) attach different value or importance on such effects.

In summary, the significant effects of text type on the students' summarization performance were the idiosyncratic functions of various characteristics of the source texts and the summarizers. In order to select an appropriate text for designing summarization tasks, it is imperative to undertake a detailed examination of these characteristics, although this may not necessarily be entirely satisfactory. The common approaches to measuring the difficulty level of texts for general reading comprehension, such as readability indices, percentage of passive sentences and topic familiarity, are less functional and illuminating than its macro-organisation and vocabulary density.

2) Filter plant: language abilities

In the previous two sections, the effects of three components of *filter plant* – topic and computer familiarity and knowledge of English vocabulary – were discussed as part of the effects of *text input*. The focus of this section will be specifically on the relationship between students' summarization performances and their other language abilities.

a) Summary of key findings

RQ2: *Are students' summarization performances affected by their other linguistic abilities and if so, to what extent?*

The students' reading abilities as measured by TOEFL were the only significant and the best predictor of students' summarization performances among the language abilities (FCE reading, English and Chinese writing, and translation from English to Chinese). Students of higher TOEFL-R produced better English and Chinese summaries. Though significant, TOEFL-R can only explain a very small amount of the students' summarization performance (less than 10%). What is more, it seemed that TOEFL-R was a better basis for prediction of RSC scores than of HS. Generally speaking, these findings were in line with the students' perceptions of the differential contributions of their other language abilities towards summarization performance. However, the perception data also suggested that the students considered that (*a*) there was a much stronger link between their reading abilities and summarization performance than was evidenced in the performance data (see **8.2.3**), (*b*) reading comprehension was a *sine qua non* of summarization and (*c*) summarization tasks were a better reflection of reading abilities than the commonly used multiple choice questions.

b) Further discussion

The main reason for rejecting the use of traditional summarization tasks in large-scale language tests is based on the concern that such tasks may require students' to use writing abilities, therefore muddying the measurement of reading abilities[3] (Alderson 1996: 225; 2000; Alderson *et al.* 1995; Weir 1993, 2005). Although the statistical findings from this project seem to suggest that the traditional summarization tasks did not measure reading comprehension abilities alone, it is clearly demonstrated that the students' summarization performances were not significantly affected by their English writing ability – the language ability that causes

[3] And also concerns regarding the subjectivity in marking summary protocols (see **11.2.2**).

the major concern concerning the claimed muddiedness of traditional summarization tasks in EFL tests, nor by their abilities in Chinese writing and translation (English to Chinese) in the case of the Chinese summarization tasks. The students' reading ability was the only significant predictor of their summarization performance. On the one hand, these empirical findings challenge some assumptions regarding the effects of writing abilities muddling up summarization as a measure of reading comprehension and may contribute to allaying such concerns. Taking into consideration the promise of summarization tasks in communicative language testing and teaching (see **2.3**), to reject this type of task solely on the basis of the claimed effects of English writing abilities in the case of English summarization tasks and Chinese writing and translation abilities in the case of Chinese summarization tasks seems ungrounded. On the other hand, these findings probably raise more questions than they answer. In the following four paragraphs, four such questions are discussed in relation to (i) the significant but small contribution of reading ability, (ii) the non-significant contribution of writing ability, (iii) the differential contribution of TOEFL-R and FCE-R towards summarization performance and (iv) students' perceptions of the relationship between summarization performance and other language abilities.

The small amount of variance in summarization performance explained by the students' reading abilities raises further concerns as to how the other components of the *filter plant* (see Figure 2.1) might have contributed to the students' summarization performance. Due to the small scale nature of this project, it was not possible to explore students' other characteristics such as their first language summarization skills and literacy expertise, cognitive styles and analytical skills (e.g. Mast 1988; Rickards *et al.* 1997). It is probable that their first language summarization skills and literacy expertise may be at least as illuminating as their foreign language reading ability for summarization performance, as found in Corbeil (2000) and Cumming *et al.* (1989) respectively (see **2.5.4**). Corbeil (2000) found that English learners of French who had a good command of using the macro-rules of summarization in their first language (English) also attempted to do the same in the summarization tasks in French.

Similarly, Cumming *et al.* (1989) found that the thinking processes of summarizing a challenging text in one's first language (English) seemed to be fundamentally the same as those in summarization in their second language (French). However, this transfer of summarization skills between the first and the second language might be due to (*a*) the close relationships between the English and French languages, being within the same language family, and (*b*) the possible shared understanding of the requirements of a good summary in English and French academic and cultural contexts (see also **2.5.4.4** *cultural variations in summarization*). In the context of this study, in which there is greater distance between the English and Chinese languages and also academic traditions in summary writing (Shi 2004), further research is needed to examine to what extent a similar transfer of summarization skills from Chinese to English occurs (see also effects of language distance on performance in ESL examinations, e.g., Elder & Davies 1998).

Although this research clearly demonstrated that students' writing ability was not able to make a significant contribution to their summarization performance, these findings may well be attributable to (*a*) the students' writing abilities and (*b*) the writing tasks *per se*. It was probably because the undergraduates in this project had quite similar writing skills (see **5.3**) and these writing skills were sufficient for the summarization tasks, and therefore the effect of their writing skills on summarization performance might be less pronounced statistically than that found by Karl Taylor (1986) who studied children's written summarization. On the other hand, the two independent writing tasks (see Appendix 7) in this project might not measure *the* same writing skills as required in the summarization tasks. Furthermore, the different focus of the rating criteria between the independent writing (see Appendix 8) and the summarization tasks (Appendices 12 & 13) may have increased the gap between them. In other words, the writing skills for the summarization tasks might not share the same construct with the two independent writing tasks. This supports, to some extent, my earlier argument that the underlying construct of integrated reading/writing tasks are fundamentally different from independent writing tasks and they require different

writing skills (see **1.2.1**).

Both TOEFL-R and FCE-R are reported as measuring reading abilities of EFL learners and are supposed to share some underlying construct of reading comprehension. However, the notable difference in their predictability of students' summarization performance not only confirms Bachman *et al.*'s (1995: 15) finding that the two tests represent "radically different approaches to language test development", but also raises some methodological concerns in studying correlations between summarization and reading abilities, and raises questions relating to the comparability of the research findings in the literature regarding such correlations (e.g. Head *et al.* 1989; L. Taylor 1996; Thomas & Bridge 1980). As demonstrated in this project, the use of a different reference point could apparently have some dramatic effects on research findings. The conclusions reached using a single reference point in the literature are therefore questionable. What is more, the evaluation of the correlations between written summarization and basic reading comprehension exclusively using multiple choice questions (e.g. Head *et al.* 1989; and the current project) may be over-simplistic, as demonstrated in Trites and McGroarty's (2005) research on the relationships between summarization-like tasks and basic reading comprehension measured by multiple choice questions using TOEFL specifications. In order to gain greater understanding of the correlations between traditional summarization and other language abilities, between traditional and "innovative" summarization tasks, more research studies are needed, using various assessment tools of different educational measurement traditions and formats including teacher assessment (e.g. Cohen 1994; L. Taylor 1996) and learner self-assessment and appraisal.

The much stronger correlations between summarization performance and reading abilities as assessed by the students themselves than were demonstrated in their actual performance may be encouraging to language testers. Students' positive experiences in undertaking the summarization tasks and positive perceptions of the contribution of

reading abilities towards summarization performance have the potential to exert positive washback on their development of the very much needed summarization skills in the information inflation age (see **2.3**). In addition, a test more welcomed by test takers themselves has the benefits of potentially being more humane and motivating than a cold clicking of multiple choice items.

3) Output

a) Summary of key findings

RQ3: *What impact does the use of a different language and language order have on summarization performances and measurement of reading comprehension abilities?*

The use of different languages had significant impacts on both the students' summarization processes and products. Different strategies were employed in the English and Chinese summarization processes. In terms of the products of the summarization tasks, although the students wrote significantly longer Chinese summaries with the possible facility of Chinese as their first language, Chinese summaries received consistently significantly lower scores than English ones, regardless of which scoring templates were used. However, Chinese summarization was better able to reflect the students' EFL reading abilities than English summarization tasks. Because *language order* was deliberately controlled in the research design; no significant main effects of *language order* were found in summarization performance, as anticipated. However, significant interactive effects of *language* and *language order* were noticed in the students' summarization processes and products. The initial summarisation, be it in English or Chinese, affected the following summarization processes and products. For example, the English summaries were significantly longer if produced in the order of *English then Chinese* than *Chinese then English.*

In addition, it was found that Chinese summarization was more sensitive to the effects of the other components in the IFOE framework such as *text type* and

presentation mode than English summarization (see **11.2.1**).

b) Further discussion

Contrary to Bensoussan and Kreindler's (1990) finding, the current research clearly demonstrated that use of different languages – an essential test method facet (Bachman 1990) determining the "type of summary to be produced" (Hidi & Anderson 1986: 473) – had played a significant role in the students' summarization processes and products. The obvious "advantages" of English summarization from the students' viewpoint, such as direct copying and imitating the source texts, shed light on the problems or disadvantages of target language summarization tasks as a measure of reading comprehension. The majority of the students favoured English summarization tasks over Chinese because of the possibility of verbatim copying without necessarily fully understanding source texts, as in the English summarization tasks (see also **11.2.2.1**), which to some extent "leaked" the information that English summarization may be less capable of measuring their EFL reading abilities. On the other hand, successful Chinese summarization was considered to require full understanding of source texts.

The students' reading abilities were better predicted by their Chinese, rather than their English summarization – a finding not only *evidenced* in their actual performance but also *endorsed* by the students themselves. This provides empirical supporting evidence for Lee's (1986: 208) strong preference for using first language summarization tasks (recall in his research) and Alderson's (1996: 225) implicit suggestion in his questioning of "whether the first language responses would be more suitable in this form of test [summarization]" to mitigate the confounding effects of target language writing abilities on summarization. In this respect, because Chinese summarization did not involve target language writing abilities, they seemed to be better able to tap into the students reading comprehension. Therefore, more evidence of reading comprehension was yielded in the first language summarization than in the

target language summarization. Furthermore, in the Chinese summarization tasks, there was no chance for the students to copy directly from the source texts, a feature prevalent in their English summaries (see also similar finding in Shi 2004). In the context of the computer-based English summarization tasks, the opportunity for and facility of direct copying and pasting increased significantly (see **11.2.2.1**).

From raters' point of view, it would be more difficult to discern if a statement in a summary written in the target language reflected the summarizers' true understanding of the text or simply a smart lifting from the source. This may be particularly hard to distinguish when the raters themselves are learners of that target language. However, when marking summaries written in their first language, Chinese raters may have found it easier to detect the students' misunderstandings in the Chinese summaries (see also the discussion in **11.2.1.2** on *issues in scoring reliability*). These various reasons combined may explain why the students received lower scores for their Chinese summaries than for those in English and why the first language summarization better reflected the students' reading comprehension abilities.

As discussed above, the effects of language on summarization tasks are twofold. Future research studies are needed to establish more evidence on (*a*) how the use of different language affects the summarization processes and products of students of different proficiency in the first and the target languages and, correspondingly, (*b*) how the cross-language summarization products affect the performance of raters of different first language background and target language proficiency.

4) Interactions of *text input*, *filter plant*, *output* and *evaluation system*

Apart from the main effects, several components of the IFOE framework (see Figure 2.1) also had significant interaction effects between them on some quality indicators of summarization performance. The following figure summarizes these interactions on the individual quality indicators discussed throughout the dissertation.

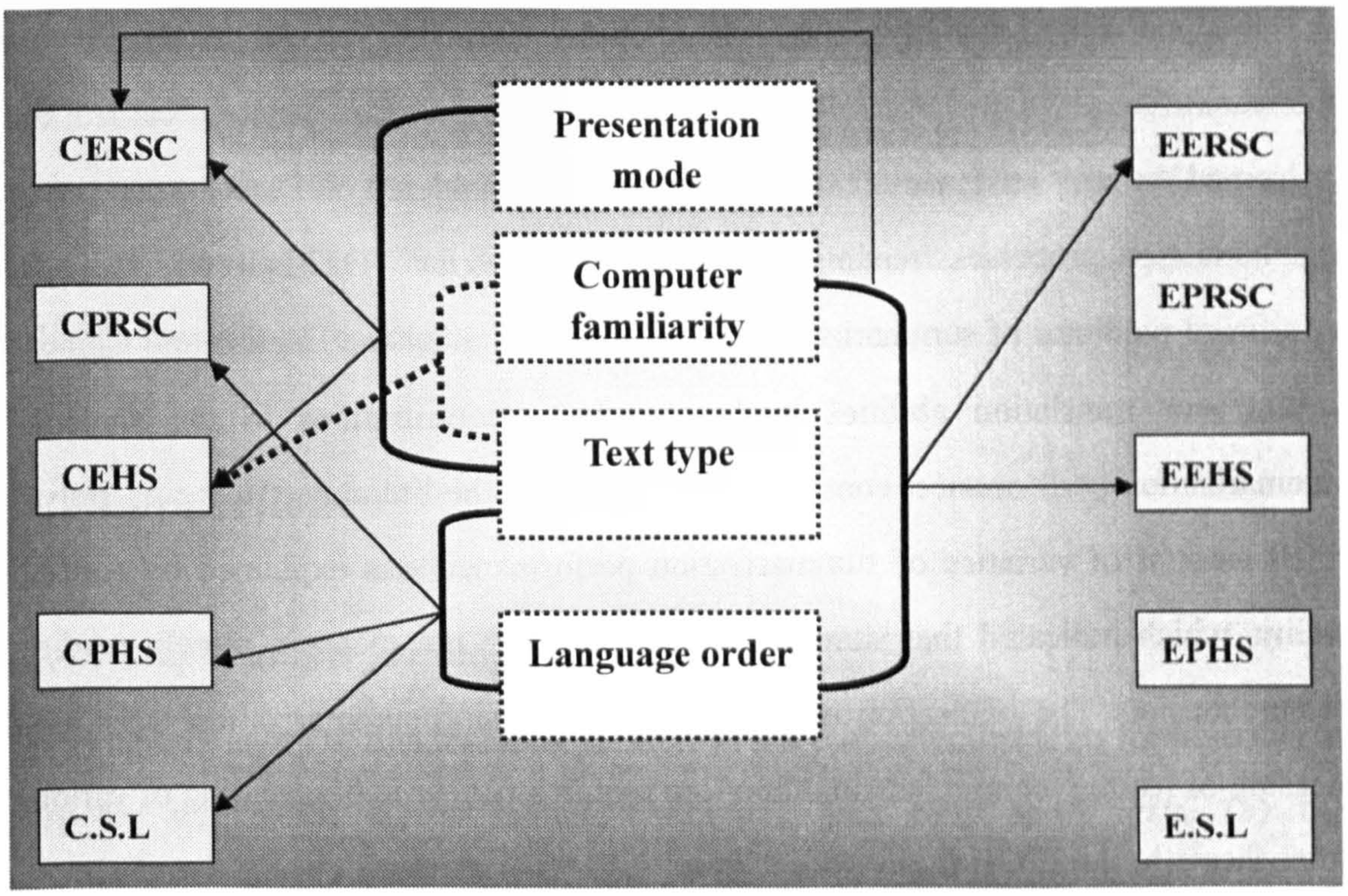


Figure 11.1 Summary of the interactions between components of IFOE framework

Besides these interactions on the individual quality indicators, there was also a series of other interactive effects on some combinations of the individual quality indicators (see Figures 8.2-8.12 in Chapter 8). These interactive effects, together with the main effects of the components of the IFOE framework, demonstrated (*a*) the complexity and dynamics of the framework *per se* on the one hand and (*b*) various facets (e.g., the use of different languages) that need to be taken into account when using summarization tasks as a measure of reading comprehension on the other. As demonstrated in Figure 11.1, the interactive effects were more pronounced on Chinese than English summarization performances.

11.3 Conclusion

The investigations of the various components of the IFOE framework (Figure 2.1) in this project – text *input* (text type and presentation mode), *filter plant* (English and Chinese writing and translation abilities), *output* (English and Chinese summarization)

and *evaluation* (RSC, HS; expert and popular templates) – demonstrated that summarization tasks are promising measures of reading comprehension, valued and welcomed by the students. Comprehension was considered the *sine qua non* of summarization processes; reading ability as evidenced by TOEFL-R was the only significant predictor of summarization performance. Neither the English and Chinese writing nor translation abilities made significant contributions to the students' summarization performance, contrary to the claims in the literature. However, only a small amount of variance of summarization performance was explained by reading ability, which indicated that summarization tasks may not measure merely reading comprehension. The realisation of the potential for using summarization tasks as a measure of reading comprehension therefore requires further understanding of various other facets in the IFOE framework (Figure 2.1). This research clearly demonstrated that summarization process is dynamic and complex. Apart from the characteristics of summarizers (e.g. language ability and computer familiarity), some features of the tasks *per se* such as *text type*, *presentation mode* and *language* had significant main and interactive effects on their summarization performance to varying degrees. Certain textual features such as vocabulary density and macro-organisation had substantial influences on a text's summarizability. Different text presentation mode also affected students' summarization process and the lengths of summaries. In addition, students of high computer familiarity produced significantly poorer Chinese summaries than students of low computer familiarity. All the effects mentioned above were much more pronounced on Chinese than English summarization. However, Chinese summarization was better able to reflect students' reading abilities than English summarization. In terms of the development and use of assessment criteria, both RSC and HS were able to discriminate the quality of summaries, and the expert templates were more welcomed by the students and better able to predict their reading abilities than the popular templates.

In summary, the IFOE framework provided a useful and dynamic methodological solution and an ecological approach to research into summarization tasks as a measure

of reading comprehension. The dynamics and complexity of summarization processes and products *per se*, as demonstrated in this project, call for language testers to take a systematic approach to the various components of the framework – *input*, *filter plant*, *output* and *evaluation* – when designing traditional summarization tasks as a valid and valuable measure of reading comprehension.

11.4 Implications

11.4.1 Implications for language testing

The originality of this research lies in (*a*) its use of extended texts for assessment purposes where the current common practice uses only short texts, (*b*) the involvement of test takers in the development of assessment criteria and the comparisons of implementation of different evaluation systems, and (*c*) the exploration of various components of the proposed IFOE framework for summarization tasks. As one reviewer commented in my recent submission to *Language Testing*, summarization tasks are a relatively "unknown construct" in the field of language testing. The IFOE framework and the findings of this project have several theoretical and methodological implications for language testing research and practice.

The dynamics and complexity of summarization processes and products mean that various facets of the framework need to be taken into account systematically when designing summarization tasks as a measure of reading comprehension.

- **Selection of source texts and effects of presentation modes**

In terms of selecting appropriate source texts, the pronounced effects of text type and its inherent features on summarization performance may present language testers with challenges that are different from those presented in designing basic reading

comprehension tests using MCQ items. The common practice of measuring the readability of a source text and readers' vocabulary knowledge and topic familiarity with the text is less indicative of its summarizability than evaluating text type (e.g., whether it has got a clear timeline of the development of ideas) and macro-organisation (e.g., whether it has an introductory summative paragraph).

Although the source texts in this project were controlled in terms of being of similar length and therefore the possible differential effects of short and extended texts on summarization performance were not researched, the length of source texts in this project seemed to be a serious issue for the students. They found it hard and uncomfortable to read extended texts, because they were already very used to reading short passages as a common practice of measuring their comprehension. If we accept that real-life reading is not simply to read texts as short as those say in FCE or TOEFL, language testers need to review the underlying philosophy of using short texts to measure reading comprehension. The use of extended and/or short texts in a language test could have the potential to engineer a certain change in the test preparation and consequently students' reading behaviours in relation to their foreign language learning. Following this logic, the use of extended texts for summarization tasks may have considerable potential.

Another consideration in designing summarization tasks is in relation to the presentation mode of the source texts – in print or on computer. Although the findings of this research suggest that computer presentation mode did not cause substantial problems in students' summarization because of their relatively high computer familiarity, it nevertheless changed their summarization processes and products. It would be imprudent to suggest computer presentation mode did not affect students' summarization performance simply based on the results from the statistical analyses of the products. When investigating effects of presentation mode on test performance, language testers need to examine not only the possible physical differences in the products but also the subtle psychological nuances in the processes of different test

delivery media.

■ **Filter plant**

Reading ability was the only significant predictor of summarization performance. High rating reliability was achieved in this project. On the one hand, these promises of using summarization tasks as a measure of reading comprehension challenge the rejection of this test method based on the commonsensical assumption of subjectivity in rating summaries and the confounding effects of writing ability on summarization performance. On the other hand, the small amount of variance in summarization performance explained by basic reading abilities measured by tests such as FCE, IELTS and TOEFL (*a*) provides the imperatives for further research to be conducted to investigate what else constitutes summarization skills and (*b*) raises questions concerning the use of separate measures of *basic* reading comprehension and *independent* essay writing abilities for the purposes of evaluating overseas students' readiness for academic study where their achievement depends to a great extent on their *integrated* reading/writing skills such as summarization (Maclellan 1997). A recent study on the subject learning experience of overseas postgraduates in a British university (Rea-Dickins *et al.* 2005) demonstrated that even the "successful" IELTS students had enormous difficulty in assignment writing which involved a significant amount of their reading-to-write skills. The narratives of these student participants also indicated discrepancies in the demands of reading abilities between IELTS, which tests basic reading comprehension[4], and subject learning contexts, which require summarization skills. Trite and McGroarty (2005) also noticed that basic reading comprehension items on the old TOEFL were a different construct from summarization-like tasks such as "reading to learn" and "reading to integrate".

---

[4] Although IELTS has summary-cloze items to test candidates' reading comprehension, the comparability of the underlying constructs of the "innovative" summary-cloze and the traditional summarization tasks is yet to be researched (see also Chapter 12).

It seems that basic reading comprehension tests are not sufficiently capable of evaluating students' linguistic readiness for academic study in English-medium higher education (Rea-Dickins *et al.* 2005). I would suggest that traditional summarization tasks be used as one of the additional measures to evaluate students' linguistic readiness. The integrated tasks involving listening, reading and writing abilities in the next generation TOEFL, although mainly for testing writing abilities, represent a positive move towards meeting the needs of the end users – not only the test candidates but also other stakeholders such as subject tutors and language support staff for overseas students.

■ Output

It has long been an established tradition in both foreign language testing and teaching that the use of first language is not as valued as the target language. This may be particularly the case if test takers are from different family of first language. However, the apparent advantages of using the first language evidenced in this research, such as the better predictability of Chinese summarization in terms of the students' English reading comprehension, the absence of confounding effects of English writing ability on Chinese summarization performance and the lack of opportunity to copy verbatim from source texts, have demonstrated that first language is clearly desirable and deployable for summarization tasks. However, in international contexts, the use of test takers' first languages would mean that test providers would need to (*a*) recruit and train raters of all test takers' first languages, (*b*) establish the comparability of summarization tasks between different first languages, and (*c*) investigate the degree of transferability of summarization skills of different first languages to the target language.

■ Evaluation

The students' attitudes towards the use of the expert and the popular evaluation

systems and the statistical differences the two systems made on the scores their summaries received demonstrated that (*a*) the involvement of test takers in the empirical development of "indigenous assessment criteria" may not be as valued and valuable as theoretically assumed if there is potential resistance from end-users – the test takers themselves – and (*b*) native speakers may still form the most practical protocol for the development of assessment criteria because of the non-native speaker test takers' willingness to accept their "non-native" identity.

In terms of the rating scales, although the scoring guidelines for RSC, WSP, SSS and HS were not impeccable, their ability to discriminate the various facets of the quality of students' written summaries was encouraging. No single criterion, independent or integrated (see **4.2.4**), was sufficient to capture the multiplicity of the quality of summarization performance. The HS scale is innovative and practical for use in wider contexts to evaluate overall quality of summaries. The use of the augmentation rating method is helpful for boosting the chance of achieving high rating reliability. By using appropriate scoring procedures and guidelines, it is not impossible to achieve high rating reliability for summarization tasks.

11.4.2 Implications for foreign language teaching and academic study

An incidental finding from this research is that the students had significant difficulty in summary writing (see Tables 5.3 & 5.5). Similar text summarization deficiencies as noted in Garner (1985) were also prevalent in these students. Further understanding of their difficulty is essential in order to facilitate students' foreign language learning in specific and academic study in general. Training in summarization strategies in students' first and second language may be an initial step in improving the chances of academic success. In addition, the incorporation of traditional summarization tasks into foreign language teaching, learning and testing may also be helpful in preparing students for academic study in English and improving their overall summarization abilities.

11.5 Summary

This chapter has discussed the key findings to the five research questions with reference to the IFOE framework (Figure 2.1) and the dynamics of summarization processes. The implications of the research findings were envisaged from the perspectives of language testing and language teaching. In the final chapter of this dissertation, I will turn to the limitations of this research and make several suggestions for future studies to explore and develop further the IFOE framework.

CHAPTER TWELVE

Limitations and Directions for Future Studies

As in many experimental studies, this project has several limitations in terms of research design and interpretation of data. In this chapter, future research studies that seek to reduce these limitations and that further explore the dynamics of the IFOE framework are suggested.

12.1 Limitations of this research

12.1.1 Reliability analyses

Although detailed rater reliability analyses were conducted in this project, there was room for further analysis of the reliability of assigning RSC and IIS scores (see Appendices 12 & 13). The raters might agree well on a score they assigned to a given summary protocol, but how they reached the score and how they marked an individual statement (for RSC) might be quite different from one another. For example, Cohen (1993: 142) found that the inter-rater agreements in marking summaries varied from 0.56 to 0.94 for a total score, and from –0.09 to 1.00 for the individual idea units. On the other hand, rater reliability is not the same as the reliability of the test *per se*. Ebel and Frisbie (1991) found that it was possible for inter-rater reliability to be high but for test reliability to be low. It must be conceded that Cronbach's Alpha is more to do with the reliability of obtaining the test results than ensuring the reliability of the test *per se*. As Brennan notes, "Reliability is a characteristic of scores, not of tests or forms of a test" (2001: 295). Although reliability estimates using Cronbach's Alpha provide some important evidence of the reliability of the test, other statistics such as standard error of measurement may be more indicative. In Cronbach's terms:

> ... the standard error of measurement ... is the most important single piece of information to report regarding an instrument, and not a coefficient. The standard error, which is a report on the uncertainty associated with each score, is easily understood not only by professional test interpreters but also by educators and other persons unschooled in statistical theory, and also to lay persons to whom scores are reported. (Cronbach & Shavelson 2004:413)[1].

Further analyses, for example, using Many-Facet Rasch (Linacre & Wright 1992), are needed to assess rater severity and reliability and to adjust examinee scores for differences in raters. It is also essential to analyse the reliability of summarization tasks *per se* through more experiments.

12.1.2 Task fatigue

The three-hour summarization tasks were demanding and indeed it might have been one of the longest tests that the students had ever taken: a number of them said in the interviews that they were tired after the summarization tasks. The fatigue from sitting for three hours doing such demanding tasks might have affected the students' performance, particularly the second part of the tasks. Although this research randomised the order of the tasks (*English then Chinese*, or *Chinese then English*, see Table 4.6) and, as anticipated, there was no significant main effect of language order, the fatigue that the students experienced was "real" and it might not be possible to remove this through the research design.

12.1.3 Predominant female student participants

In this project, the student participants were predominantly female (see Table 4.2); only 27 out of the 157 students were male. The findings of this project therefore could be limited in this sense. In future studies, it would be desirable to have a more balanced design in terms of gender.

---

[1] Shavelson provided editorial support in this paper.

12.1.4 Assigning student participants to summarization conditions

The summarization condition to which a student was assigned (see Table 4.6) was not entirely random; it was somewhat determined by which *class* s/he was originally from. Although this was the best design possible for this research context in terms of being feasible, this kind of randomization made interpretation of the findings difficult because of certain potential effects of the differences between *classes* (but see discussion of the effects of *class* on summarization performance, **11.2.2**).

The design of this research allowed detailed and simultaneous investigations of several components of the IFOE framework within one project – one of the features of originality and excellence I would claim in this research. However, I concede that this was at the expense of more in depth analyses of the data obtained. Had time allowed, I would have been able to do further analyses on the data (see also limitations in reliability analyses, **12.1.1**).

12.2 Suggestions for future studies

As raised in **12.1**, due to time constraints, not all the data obtained have been analysed in as great a depth as originally planned. There are three key areas that require attention to gain greater understanding of the promises and the problems of summarization tasks as a measure of reading comprehension, addressed in **12.2.1**. In addition, the IFOE framework (Figure 2.1) has considerable potential for further exploration of summarization tasks, as indicated in **12.2.2**.

12.2.1 Attention to data already obtained

The three key areas that need attention through further data analyses are as follows:

1) Content coverage and topographical features of English and Chinese summaries

Although the differential effects of *language* on summarization were analysed in terms of students' performance and their perceptions, it would be as interesting to examine the individual key statements reported in their English and Chinese summaries. For example, what were the similarities and differences in content between the two versions? Did the students report different statements in their English and Chinese summaries and if so, to what extent? Besides the content coverage, it would also be interesting to examine the topographical features of summaries (Sherrard 1986). In the current project, SSS was used to evaluate the summary-and-source text relationship, but further detailed analyses are needed. For example, were the positions of the key statements mentioned in both versions changed? What were the topographical relationships between the summaries and the source texts? Was there any difference in terms of topographical relations between English and Chinese summaries?

2) Discoursal features of the English summaries

In the current project, only one of the discoursal features of English summaries was analysed (i.e. vocabulary density). It would be of great interest to examine the other discoursal features such as syntactic complexity and grammatical accuracy. For instance, to what extent were the discoursal features related to the summarizers' reading comprehension abilities and to the summarizability of source texts? What were the relationships in the discoursal features between the English summaries and the short essays written in English by the same students? Investigations of these questions would shed further light on the use of summarization tasks in language testing research and practice (see also Cumming *et al.* 2005).

3) Views on the use of the two scoring templates

In the analyses on the students' attitudes towards the use of the expert and the

popular scoring templates, I did not make specific reference to the students' gender, summarization skills and general language proficiency. In order to further understand this issue, it would be necessary to distinguish to what extent students' gender and language skills were linked to their attitudes towards this type of empirical development of assessment criteria.

12.2.2 Exploring the IFOE framework

The IFOE framework (Figure 2.1) is dynamic and capable of providing language testing researchers with considerable scope for further research. Below I list several such studies that I am keen to conduct in the near future in order to refine the framework and establish a more theoretically firm and methodological friendly model for using summarization tasks as a measure of reading comprehension.

1) Input

- Multiple- or single-sourced

Test takers may be assigned several texts on the same related topic; they may also be allowed to choose at least two of the source texts for their summarization tasks. It would be interesting to examine the difference in their performance in summarizing multiple- and single-source texts.

- Extended or short source texts

Test takers may be asked to summarize both extended and short texts (not related); comparisons could then be made between their summarization performance across extended and short source texts.

- Traditional and "innovative" summarization tasks

It would be interesting to compare test takers' performance on traditional and "innovative" summarization tasks (e.g. summary-cloze) which could involve the same

and/or different source text(s).

- Integrated listening/reading/writing tasks

The pre-summarization tasks could involve listening, reading, or listening and reading from different delivery media.

2) Filter plant

- Summarization strategy instruction

An incidental finding of this project suggested that students had significant difficulty in summarization. Therefore, summarization strategy instruction prior to tests may be necessary to improve their summarization strategy and reduce the effects of strategy on performance. It would also be potentially beneficial to their academic study.

- Participants of various subject background and age

In the current project, only university undergraduates from the same department were recruited. In future studies, participants of different subject backgrounds and maturity might be invited to be involved.

- International comparisons of summarization skills

Given the increasing number of overseas students at British higher education institutions, it is important to examine the differences in summarization skills of undergraduates and postgraduates of different first language backgrounds in the globalized higher education system and how this might be linked to students' success or difficulty in academic studies.

3) Output

- Summarization on computer or in print

Due to the initial concern about the effects of computer familiarity on summarization

performance, this project asked the students to write the final draft of their summaries on paper. In future studies, participants may be asked to directly type their summaries into the computer.

- Writer- or reader-based summaries

Whether a summary is written for earning a score in an examination (i.e. reader-based) or for personal study (writer-based) may exert significant impact on how and what the students include in their summaries. In future studies, both types of summaries could be elicited.

- Written or oral summarization

Participants might be asked to produce written and/or oral summaries.

- Handwriting or word-processing

In relation to written summarization tasks, future studies may examine the effects of the quality of handwriting on rater performances.

- Extension from literal to critical summarization

This project focused only on literal summarization. Future studies may be extended to critical summarization tasks.

4) Evaluation system

- Summaries of original author(s)

If possible, original authors of source texts might be invited to write the summaries of their articles to be used as one of the scoring templates. The linguistic and topographical similarities and differences between authors' and students' summaries could then also be analysed.

- Rater performance and decision making

Investigating rater performance and decision-making processes in evaluating the quality of written or oral summaries (see **11.2.1**) might prove another fruitful area of research to inform the developments of strategies for rater training and scales for judging the quality of summarization performance (see also Cumming *et al.* 2001).

These research topics may be somewhat ambitious, but, whether alone or in combination, they demonstrate not only the dynamics of the framework – *input*, *filter plant*, *output*, and *evaluation*, but also the necessity of building a better model for using summarization tasks as a measure of reading comprehension. There is still a long way to go!

References:

Adams, C., Labouvie Vief, G., Hobart, C. J., & Dorosz, M. (1990). Adult age group differences in story recall style. *Journals of Gerontology: Psychological Sciences*, 45, P17-P27.

Aebersold, J. A., & Field, M. L. (1997). *From reader to reading teacher: Issues and strategies for second language classrooms*. Cambridge: Cambridge University Press.

Afflerbach, P. P. (1990). The influence of prior knowledge on expert readers' main idea construction strategies. *Reading Research Quarterly*, 25, 31-46.

Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language*, 6, 425-438.

— (1991). Testing reading comprehension skills (part two). *Reading in a Foreign Language*, 7, 465-503.

— (1996). The testing of reading. In C. Nuttall (Ed.), *Teaching reading skills in a foreign language*. London: Heinemann.

— (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (part i). *Language Teaching*, 34, 213-236.

— (2002). Language testing and assessment (part two). *Language Teaching*, 35, 79-113.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alexander, P. A., & Jetton, T. L. (1996). The role of importance and interest in the processing of text. *Educational Psychology Review*, 8, 89-121.

Alhaidari, A. O. (1992). *How Arab students summarize English prose and how they revise their summaries*. Unpublished PhD Dissertation, Michigan State University.

Allison, D., Berry, V., & Lewkowicz, J. (1994). Pig in the middle? Effects of mediating tasks on cognitive processing of text. In N. Bird, P. Falvey, A. B. M. Tsui, D. Allison & A. McNeill (Eds.), *Language and learning* (pp. 463-490). Hong Kong: Government Printer.

— (1995a). Processes and their products: A comparison of task sequences and outcome in EAP writing classes. *Hong Kong Papers in Linguistics and Language Teaching*, 18, 13-32.

— (1995b). Reading-writing connections in EAP classes: A content analysis of written summaries produced under three mediating conditions. *RELC Journal: A Journal of Language Teaching and Research in Southeast Asia*, 26, 25-43.

Alterman, R., & Bookman, L. A. (1990). Some computational experiments in summarization. *Discourse Processes*, 13, 143-174.

Anderson, N. J., Bachman, L. F., Perkins, K., & Cohen, A. D. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8, 41-66.

Arnold, H. F. (1942). The comparative effectiveness of certain study techniques in the field of history. *Journal of Educational Psychology*, 33, 449-457.

Atkinson, E. (2000). The promise of uncertainty: Education, postmodernism and the politics of possibility. *International Studies in Sociology of Education*, 10, 81-102.

— (2002). The responsible anarchist: Postmodernism and social change. *British Journal of Sociology of Education*, 23, 73-88.

Axelrod, J. (1975). Getting the main idea is still the main idea. *Journal of Reading*, 18, 383-387.

Ayari, S. (1998). *Using oral summarization to assess English reading comprehension of Arabic-speaking learners of English*. Unpublished PhD thesis, University of Minnesota.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

— (1991). What does language testing have to offer? *TESOL Quarterly*, 25, 671-704.

— (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17, 1-42.

Bachman, L. F., Davidson, F., Ryan, K. E., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baggaley, A. R. (1982). Deciding on the ratio of number of subjects to number of variables in factor analysis. *Multivariate Experimental Clinical Research*, 6, 81-85.

Balajthy, E., & Weisberg, R. (1990). Transfer effects of prior knowledge and use of graphic organizers on college developmental readers' summarization and comprehension of expository text. *National Reading Conference Yearbook*, 39, 339-345.

Banerjee, J. (2003). *Interpreting and using proficiency test scores*. Unpublished PhD dissertation, Lancaster University.

Banerjee, J., & Luoma, S. (1997). Qualitative approach to test validation. In C. Clapham & D. Corson (Eds.), *Language testing and assessment. Encyclopedia of language and education. Vol. 7* (pp. 275-287). Dordrecht: Kluwer.

Barati, H. (2005). *Test-taking strategies and the assessment of reading skills: An approach to construct validation*. Unpublished PhD thesis, University of Bristol.

Barnwell, D. (1989). "naïve" native speakers and judgments of oral proficiency in Spanish. *Language Testing*, 6, 152-163.

Barry, S., & Lazarte, A. (1995). Embedded clause effects on recall: Does high prior knowledge of content domain overcome syntactic complexity in students of Spanish? *Modern Language Journal*, 79, 491-504.

— (1998). Evidence for mental models: How do prior knowledge, syntactic

complexity, and reading topic affect inference generation in a recall task for nonnative readers of Spanish? *Modern Language Journal*, 82, 176-193.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, Mass.: Cambridge University Press.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society*, 16 (Series B), 296-298.

Basham, C. S. (1986). *Summary writing: A study in textual and contextual constraints*. Unpublished PhD thesis, University of Michigan.

— (1987). Summary writing as cultural artifact. Fairbanks, AK: Cross Cultural Communications, University of Alaska.

Basham, C. S., & Rounds, P. L. (1984). A discourse analysis approach to summary writing. *TESOL Quarterly*, 18, 527.

— (1986). A discourse analysis approach to summary writing. *Papers in Applied Linguistics - Michigan*, 1, 88-104.

Baumann, J. F. (1986). Teaching main idea comprehension. Newark, DE: International Reading Association.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions*. London: Sage.

Bensoussan, M. (1982). Testing the test of advanced EFL reading comprehension: To what extent does the difficulty of a multiple-choice test reflect the difficulty of the text? *System*, 10, 285-290.

— (1993). A guided summary completion test for long academic texts. *Language Testing Update*, 14, 59-61.

Bensoussan, M., Goldenblatt, L., & Kreindler, I. (1984). Changing the difficulty level of multiple-choice EFL reading comprehension questions. *Language Testing*, 1, 105-109.

Bensoussan, M., & Kreindler, I. (1990). Improving advanced reading comprehension in a foreign language: Summaries vs. Short-answer questions. *Journal of Research in Reading*, 13, 55-68.

Bereiter, C. (1994). Implications of postmodernism for science, or, science as progressive discourse. In D. C. Phillips (Ed.), *Epistemological perspectives on educational psychology* (pp. 3): Lawrence Erlbaum Associates Inc.

Bereiter, C., Scardamalia, M., Cassells, C., & Hewitt, J. (1997). Postmodernism, knowledge building, and elementary science. *Elementary School Journal*, 97, 329-340.

Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.

Bernhardt, E. B., & Deville, C. (1991). Testing in foreign language programs and testing programs in foreign language departments: Reflections and recommendations. In R. V. Teschner (Ed.), *Assessing foreign language proficiency of undergraduates*. Boston, MA: Heinle & Heinle.

Best, S., & Kellner, D. (1997). *The postmodern turn*. New York: The Guilford Press.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to

assessment. In M. Birenbaum & F. R. J. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes, and prior knowledge* (pp. 3-29). Dordrecht: Kluwer.

Blake, N. (1996). Between postmodernism and anti-modernism: The predicament of educational studies. *British Journal of Educational Studies*, 44, 42-65.

— (1997). A postmodernism worth bothering about; a rejoinder to cole, hill and rikowski. *British Journal of Educational Studies*, 45, 293-305.

Blake, N., Smeyers, P., Smith, R., & Standish, P. (1999). Thinking again: Education after postmodernism. *British Journal of Educational Studies*, 47, 407-408.

Borderia-Garcia, A., & Oskoz, A. (2001). Classical test theory in investigating the reliability of the recall protocol. Poster presented at the 23rd annual Language Testing Research Colloquium. Feb. 20-24. St. Louis, Missouri, USA.

Boyd, K., & Davies, A. (2002). Doctors' orders for language testers: The origin and purpose of ethical codes. *Language Testing*, 19, 296-322.

Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31, 675-685.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38, 295-317.

Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. *Contemporary Educational Psychology*, 4, 145-153.

Briggs, D. (1970). The influence of handwriting on assessment. *Educational Research*, 13, 50-55.

Brooks, L. W., Dansereau, D. F., Spurlin, J. E., & Holley, C. D. (1983). Effects of heading on text processing. *Journal of Educational Psychology*, 75, 292-302.

Brookshire, R. H., & Nicholas, L. E. (1984). Comprehension of directly and indirectly stated main ideas and details in discourse by brain-damaged and non-brain-damaged listeners. *Brain and Language*, 21, 21-36.

Brown, A. (2003). Legibility and the rating of second language writing: An investigation of the rating of handwritten and word-processed IELTS task two essays. In R. Tulloh (Ed.), *International language testing system research reports volume 4* (pp. 131-151). Canberra: IELTS Australia Pty Limited.

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher*, 10, 14-24.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, 22, 1-14.

Brown, A. L., Day, J. D., & Jones, R. S. (1983). The development of plans for summarizing texts. *Child Development*, 54, 968-979.

Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem of metacognitive development. *Child Development*, 48, 1-8.

— (1978). The development of strategies for studying text. *Child Development*, 49, 1076-1088.

Brown, J. D. (1991). Do English and ESL instructors rate writing samples differently? *TESOL Quarterly*, 25, 587-603.

Bueckendorf, J. M. (1992). *Comparative assessment of reading comprehension: Using multiple-choice and written summary formats with narrative and expository texts*. Unpublished EdD thesis, University of Missouri, St Louis.

Byler, C. R. (2001). *An improved method for text summarization using lexical chains*. Unpublished PhD thesis, University of Tennessee.

Byrd, M. (1985). Age differences in the ability to recall and summarize textual information. *Experimental Aging Research*, 11, 87-91.

Carrell, P. L. (1983). Three components of background knowledge in reading comprehension. *Language Learning*, 33, 183-207.

— (1992). Awareness of text structure: Effects on recall. *Language Learning*, 42, 1-20.

Cattell, R. B. (1952). *Factor analysis: An introduction and manual for the psychologist and social scientist*. New York: Harper & Brothers, Publishers.

Cavalcanti, M. C. (1987). Investigating FL reading performance through pause protocols. In C. Faerch & G. Kasper (Eds.), *Introspection in second language research* (pp. 230-250). Clevedon, England: Multilingual Matters Ltd.

Chalhoub-Deville, M. B., & Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, 19, 273-299.

Charge, N., & Taylor, L. B. (1997). Recent developments in IELTS. *ELT Journal*, 51, 374-380.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical review. *Research in the Teaching of English*, 18, 65-81.

Chase, C. I. (1968). The impact of some obvious variables on essay test scores. *Journal of Educational Measurement*, 5, 315-318.

— (1979). The impact of achievement expectations and handwriting quality on scoring essay tests. *Journal of Educational Measurement*, 16, 39-42.

— (1983). Essay test scores and reading difficulty. *Journal of Educational Measurement*, 20, 293-297.

— (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, 23, 33-41.

Child, D. (1990). *The essentials of factor analysis*. London: Cassell Educational Limited.

Clapham, C. M. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Coffman, W. E., & Kurfman, D. (1968). A comparison of two methods of reading essay examinations. *American Educational Research Journal*, 5, 99-107.

Cohen, A. D. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 132-159). Washington, DC: TESOL, Inc.

— (1994). English for academic purposes in Brazil: The use of summary tasks. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an*

international language (pp. 174-204). London: Longman.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Coniam, D. (1993). Co-text or no-text? A study of an adapted cloze technique for testing comprehension/summary skills. *Hong Kong Papers in Linguistics and Language Teaching*, 16, 1-10.

Connor, U. (1984). Recall of text: Differences between first and second language readers. *TESOL Quarterly*, 18, 239-255.

Connor, U., & McCagg, P. (1983). Cross-cultural differences and perceived quality in written paraphrases of English expository prose. *Applied Linguistics*, 4, 259-268.

Constas, M. A. (1998). The changing nature of educational research and a critique of postmodernism. *Educational Researcher*, 27, 26-37.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15-18.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.

Corbeil, G. (2000). Exploring the effects of first- and second-language proficiency on summarizing in French as a second language. *Canadian Journal of Applied Linguistics*, 3, 35-62.

Corder, S. P. (1973). Interpretive procedures: Seeing, reading and understanding discourse. In R. Anderson (Ed.), *New dimensions in second language acquisition research*. Rowley, MS: Newbury House.

Cordero-Ponce, W. L. (2000). Summarization instruction: Effects on foreign language comprehension and summarization of expository texts. *Reading Research & Instruction*, 39, 329-350.

Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology*, 88, 288-295.

Courchêne, R., & Bayliss, D. (1995). Summary cloze: What is it? What does it measure? In R. Courchêne, S. Burger, C. Cornaire, R. LeBlanc, S. Paribakht & H. Seguin (Eds.), *Twenty-five years of second language teaching at the university of Ottawa* (pp. 305-327). Ottawa: Second Language Institute, University of Ottawa.

Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 474-479.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). Generalizability analysis for educational assessment. Los Angeles: UCLA's Center for the Study of Evaluation & National Center for Research on Evaluation, Standards, and Student Testing.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Education and Psychological Measurement*, 64, 391-418.

Crowley, E. (1987). *The relationship between ability to summarize and field*

independence or dependence. Unpublished PhD thesis, Illinois State University.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, 10, 5-43.

Cumming, A., Kantor, R., & Powers, D. E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL monograph series, report no. 22). Princeton, NJ.: Educational Testing Service.

Cumming, A., Rebuffot, J., & Ledwell, M. (1989). Reading and summarizing challenging texts in first and second languages. *Reading and Writing: An Interdisciplinary Journal*, 2, 201-219.

Cunningham, J. W., & Moore, D. W. (1986). The confused world of main idea. In J. F. Baumann (Ed.), *Teaching main idea comprehension* (pp. 1-17). Neward, DE: International Reading Association.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24, 197-222.

Dallmayr, F. R. (1987). Politics of the kingdom - Pannenberg's anthropology. *Review of politics*, 49, 85-111.

Daly, J. A., & Dickerson-Markman, F. (1982). Contrasts effects in evaluating essays. *Journal of Educational Measurement*, 19, 309-316.

Davies, A. (1997). Australian immigrant gatekeeping through English language tests: How important is proficiency? In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 71-84). Tampere: Universities of Tampere and Jyväskylä.

— (2003). *The native speaker: Myth and reality*. Clevedon, Bristol: Multilingual Matters Ltd.

Davies, E., & Whitney, N. (1984). *Study skill 11: Writing summaries*. London: Heinemann.

Day, J. D. (1980). *Teaching summarization skills: A comparison of training methods*. Unpublished PhD thesis, University of Illinois at Urbana-Champaign.

— (1986). Teaching summarization skills: Influences of student ability level and strategy difficulty. *Cognition and Instruction*, 3, 193-210.

Dermody, M. M., & Speaker, R. B., Jr. (1999). Reciprocal strategy training in prediction, clarification, question generating and summarization to improve reading comprehension. *Reading Improvement*, 36, 16-23.

Deville, C., & Chalhoub-Deville, M. (1993). Modified scoring, traditional item analysis and Sato's caution index used to investigate the reading recall protocol. *Language Testing*, 10, 117-132.

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35, 1297-1326.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing*, 18, 171-185.

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary

students: Whose criteria? In A. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium* (pp. 60-81). Cambridge: University of Cambridge Press with University of Cambridge Local Examinations Syndicate.

Dyson, M. C., & Haselgrove, M. (2000). The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of Research in Reading*, 23, 210-223.

Eagleton, T. (1996). *The illusions of postmodernism*. Oxford: Blackwells Publishers Ltd.

Eamon, D. B. (1978/1979). Selection and recall of topical information in prose by better and poorer readers. *Reading Research Quarterly*, 14, 244-257.

Ebel, R., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Edmundson, H. P. (1964). Problems in automatic abstracting. *Communications of the ACM*, 7, 259-263.

Educational Testing Service (2002). *LanguEdge courseware: Handbook for scoring speaking and writing*. Princeton, NJ: Educational Testing Service.

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). Development of a scale for assessing the level of computer familiarity of TOEFL examinees. TOEFL research report 60. Princeton, NJ: Educational Testing Service.

Elder, C., & Davies, A. (1998). Performance on ESL examinations: Is there a language distance effect? *Language and Education*, 12, 1-17.

Endres-Niggemeyer, B. (2000). Simsum: An empirically founded simulation of summarizing. *Information Processing & Management*, 36, 659-682.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT.

Farr, R., & Carey, R. F. (1986). *Reading: What can be measured?* Newark, DE: International Reading Association.

Fletcher, M. B. (1990). *Effects of text structure and text difficulty on summary writing*. Unpublished EdD thesis, Harvard University.

Fløttum, K. (1985). Methodological problems in the analysis of student summaries. *Text*, 5, 291-307.

Foucault, M. (1977). *Discipline and punish: The birth of the prison*. New York: Vintage Books.

— (1980). Truth and power. In C. Gordon (Ed.), *Power/knowledge: Selected interviews and other writings, 1972-1977* (pp. 133). New York: Pantheon Books.

Friend, R. (1995). *Effects of strategy instruction and self-monitoring on summary writing of college students*. Unpublished PhD thesis, City University of New York.

— (2001). Effects of strategy instruction on summary writing of college students. *Contemporary Educational Psychology*, 26, 3-24.

— (2002). Summing it up--teaching summary writing to enhance science learning. *The Science Teacher*, 69, 40-43.

Fulcher, G. Ethics in language testing: http://taesig.8m.com/news1.html.

Gajria, M. L. (1989). *Direct instruction of a summarization strategy: Effect on text comprehension and recall in learning-disabled students*. Unpublished PhD thesis, Pennsylvania State University.

Gajria, M. L., & Salvia, J. (1992). The effects of summarization instruction on text comprehension of students with learning disabilities. *Exceptional Children*, 58, 508-516.

Garner, R. (1982). Efficient text summarization: Costs and benefits. *Journal of Educational Research*, 75, 275-279.

— (1985). Text summarization deficiencies among older students: Awareness or production ability? *American Educational Research Journal*, 22, 549-560.

Garner, R., & McCaleb, J. L. (1985). Effects of text manipulations on quality of written summaries. *Contemporary Educational Psychology*, 10, 139-149.

Gauntt, H. L. (1989). *The roles of prior knowledge of text structure and prior knowledge of content in the comprehension and recall of expository text*. Unpublished PhD thesis, University of Delaware.

Ghiselli, E., Campbell, J., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. New York: W. H. Freeman &Co.

Giora, R., & Shen, Y. (1994). Degrees of narrativity and strategies of semantic reduction. *Poetics: Journal for Empirical Research on Literature, the Media and the Arts*, 22, 447-458.

Godev, C. B., Martinez-Gibson, E. A., & Toris, C. M. (2002). Foreign language reading comprehension test: L1 versus L2 in open-ended questions. *Foreign Language Annals*, 35, 202-221.

Golden, J., Haslett, B., & Gauntt, H. (1988). Structure and content in eighth-graders' summary essays. *Discourse Processes*, 11, 139-162.

Golden, R. M., & Rumelhart, D. E. (1993). A parallel distributed processing model of story comprehension and recall. *Discourse Processes*, 16, 203-237.

Goldman, S. R., Saul, E., & Coté, N. (1995). Paragraphing, reader, and task effects on discourse comprehension. *Discourse Processes*, 20, 273-305.

Gomulicki, B. (1956). Recall as an abstractive process. *Acta Psychologica*, 12, 77-94.

Gordon, C. M., & Hanauer, D. (1995). The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly*, 29, 299-324.

Grabe, W. (2000). Reading research and its implications for reading assessment. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 226-262). Cambridge: Cambridge University Press.

Graham, S., Boyer Shick, K., & Tippets, E. (1989). The validity of the handwriting scale from the test of written language. *Journal of Educational Research*, 82, 166-171.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 26, 499-510.

Greene, M. (1993). Reflections on postmodernism and education. *Educational Policy*,

7, 206.

Grenz, S. J. (1996). *A primer on postmodernism*. Michigan: William B. Eerdmans Publishing Co.

Guido, B. A., & Colwell, C. G. (1987). A rationale for direct instruction to teach summary writing following expository text reading. *Reading Research & Instruction*, 26, 89-98.

Hadwin, A. F., Kirby, J. R., & Woodhouse, R. A. (1999). Individual differences in notetaking, summarization, and learning from lectures. *Alberta Journal of Educational Research*, 45, 1-17.

Hahn, A. L., & Smith, T. F. (1986). Topic familiarity as a factor in the summarizing process. *Reading Improvement*, 23, 94-99.

Hales, L. W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grade assigned to subsequent responses to an essay question. *Journal of Educational Measurement*, 12, 115-118.

Hall, J. W., Miskiewicz, R., & Murray, C. G. (1977). Effects of test expectancy (recall vs. Recognition) on children's recall and recognition. *Bulletin of the Psychonomic Society*, 10, 425-428.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, 17, 164-172.

Hambleton, R. K., & De Jong, J. H. A. L. (2003). Advances in translating and adapting educational and psychological test. *Language Testing*, 20, 127-134.

Hamp-Lyons, L. (1998). Ethics in language testing. In D. Corson & C. Clapham (Eds.), *Language testing and assessment. Vol. 7 of the encyclopedia of language and education* (pp. 323-333). Amsterdam: Kluwer Academic Publishers.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000: Writing: Composition, community and assessment*. Princeton, NJ: Educational Testing Service.

Hare, V. C., & Borchardt, K. M. (1984). Direct instruction of summarization skills. *Reading Research Quarterly*, 20, 62-78.

Hare, V. C., Rabinowitz, M., & Schieble, K. M. (1989). Text effects on main idea comprehension. *Reading Research Quarterly*, 24, 72-88.

Harris, T. L., & Hodges, R. E. (1981). A dictionary of reading and related terms. Newark, DE: International Reading Association.

Head, M. H. (1986). *Factors affecting summary writing and their impact on reading comprehension assessment*. Unpublished PhD thesis, The Louisiana State University and Agricultural and Mechanical College.

Head, M. H., Readence, J. E., & Buss, R. R. (1989). An examination of summary writing as a measure of reading comprehension. *Reading Research & Instruction*, 28, 1-11.

Heaton, J. B. (1990). *Writing English language tests (2nd edition)*. London: Longman.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newbury House.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands,

cognitive operations, and implications for instruction. *Review of Educational Research*, 56, 473-493.

Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment*. Tampere: Universities of Tampere and Jyväskylä.

Hoaglin, D., & Welsch, R. (1978). The hat matrix in regression and ANOVA. *American Statistician*, 32, 17-22.

Holmes, J. L. (1996). *Studying in two languages: Study summaries in the classroom.* Unpublished PhD thesis, University of Lancaster.

Holmes, J. L., & Ramos, R. G. (1993). Study summaries as an evaluation instrument: Questions of validity. *English for Specific Purposes*, 12, 83-94.

Hooper, S., Sales, G., & Rysavy, S. D. M. (1994). Generating summaries and analogies alone and in pairs. *Contemporary Educational Psychology*, 19, 53-62.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation.* Needham Heights, MA: Allyn and Bacon.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, D. E., Keeling, B., & Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement*, 17, 131-136.

— (1983). Affects of achievement and handwriting quality on scoring essays. *Journal of Educational Measurement*, 20, 65-70.

Huhta, A., & Randell, E. (1996). Multiple-choice summary: A measure of text comprehension. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 94-110). Clevedon, England: Multilingual Matters.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 41, 201-213.

Hyönä, J., Lorch, R. F., Jr., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, 94, 44-55.

ISO-214-1976 (1976). International organization for standardization: Documentation: Abstracts for publication and documentation.

Jackson, J. D., & Kemper, S. (1993). Age differences in summarizing descriptive and procedural texts. *Experimental Aging Research*, 19, 39-51.

Jacoby, S., & McNamara, T. F. (1999). Locating competence. *English for Specific Purposes*, 18, 213-241.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19, 57-84.

Jitendra, A. K., Cole, C. L., Hoppes, M. K., & Wilson, B. (1998). Effects of a direct instruction main idea summarization program and self-monitoring on reading comprehension of middle school students with learning disabilities. *Reading and Writing Quarterly: Overcoming Learning Difficulties*, 14, 379-396.

Johns, A. M. (1985). Summary protocols of "underprepared" and "adept" university students: Replications and distortions of the original. *Language Learning*, 35, 495-517.

— (1988). Reading for summarizing: An approach to text orientation and processing. *Reading in a Foreign Language*, 4, 79-90.

Johns, A. M., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics*, 11, 253-271.

Johnson, N. S. (1983). What do you do if you can't tell the whole story? The development of summarization skills. In K. E. Nelson (Ed.), *Children's language, volume 4* (pp. 315-383). Hillsdale, NJ: Lawrence Erlbaum.

Johnson, R. E. (1970). Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior*, 9, 12-20.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Newton, Massachusetts: Allyn and Bacon, Inc.

Kaiser, H. (1970). A second generation little jiffy. *Psychometrika*, 35, 401-415.

— (1974). An index of factorial simplicity. *Psychometrika*, 39, 31-36.

Katz, S., Lautenschlager, G., Blackburn, A., & Harris, F. (1990). Answering reading comprehension items without passages on the sat. *Psychological Sciences*, 1, 122-127.

Kiewit, S. F. (1997). *The relationship of prior background knowledge to the summarization skills of developmental college students*. Unpublished PhD thesis, University of Akron.

Kim, J. S. (1989). *The effects of thematic titles on recall measures of reading comprehension of Korean students of English as a second language*. Unpublished PhD thesis, Ohio State University.

Kim, S. A. (1995). Types and sources of problems in L2 reading: A qualitative analysis of the recall protocols by Korean high school EFL students. *Foreign Language Annals*, 28, 49-70.

— (2001). Characteristics of EFL readers' summary writing: A study with Korean university students. *Foreign Language Annals*, 34, 569-581.

Kintsch, E. H. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction*, 7, 161-195.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

— (1982). Text representations. In W. Otto & S. White (Eds.), *Reading expository material* (pp. 87-101). New York: Academic Press.

— (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182.

Kintsch, W., & Greene, E. (1978). The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes*, 1, 1-13.

Kintsch, W., & Kozminsky, E. (1977). Summarizing stories after reading and listening. *Journal of Educational Psychology*, 69, 491-499.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

Kirby, J. R., & Pedwell, D. (1991). Students' approaches to summarisation. *Educational Psychology: an International Journal of Experimental Educational Psychology*, 11, 297-307.

Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly*, 25, 105-121.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. R. (1998). Computer familiarity among TOEFL examinees (TOEFL research report 59). Princeton, NJ: Educational Testing Service.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10, 62-102.

— (1984). Readability. In P. D. Pearson, R. Barr, M. L. Kamil & P. B. Mosenthal (Eds.), *Handbook of reading research* (pp. 681-744). New York: Longman.

Kobayashi, M. (1995). *Effects of text organisation and test format on reading comprehension test performance*. Unpublished PhD thesis, Thames Valley University.

— (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19, 193-220.

Kuckartz, U. (1998). winMAX: Scientific text analysis for the social sciences. Berlin: VERBI Software-Consult-Research GmbH (Germany).

Lado, R. (1961). *Language testing: The construction and use of foreign language tests. A teacher's book*. London: Longman.

— (1986). Analysis of native speaker performance on a cloze test. *Language Testing*, 3, 130-146.

Lambiotte, J. G., & Dansereau, D. F. (1992). Effects of knowledge maps and prior knowledge on recall of science lecture content. *Journal of Experimental Education*, 60, 189-201.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition*, 8, 201-211.

— (1987). Comprehending the Spanish subjunctive: An information processing perspective. *Modern Language Journal*, 71, 50-57.

Lehnert, W. G. (1981). Plot units and narrative summarization. *Cognitive Science*, 5, 293-331.

— (1984). Narrative complexity based on summarization algorithms. In B. G. Bara & G. Guida (Eds.), *Computational models of natural language processing* (pp. 247-259). Amsterdam: North-Holland.

Lehnert, W. G., & Loiselle, C. (1989). An introduction to plot units. In D. Waltz (Ed.), *Semantic structures: Advances in natural language processing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewy, A. (1996). Postmodernism in the field of achievement testing. *Studies in Educational Evaluation*, 22, 223-244.

Linacre, J. M., & Wright, D. (1992). *A user's guide to FACETS Rasch measurement computer program*. Chicago: MESA.

Long, J., & Harding-Esch, E. (1978). Summary and recall of text in first and second languages: Some factors contributing to performance differences. In D. Gerver & H. W. Sinaiko (Eds.), *Language interpretation and communication* (pp.

273-287). New York: Plenum.

Lorch, R. F., Jr. (1989). Text-signaling devices and their effects on reading and memory processes. *Educational Psychology Review*, 1, 209-234.

Lorch, R. F., Jr., & Lorch, E. P. (1985). Topic structure representation and text recall. *Journal of Educational Psychology*, 77, 137-148.

— (1986). On-line processing of summary and importance signals in reading. *Discourse Processes*, 9, 489-496.

— (1995). Effects of organizational signals on text-processing strategies. *Journal of Educational Psychology*, 87, 537-544.

— (1996). Effects of headings on text recall and summarization. *Contemporary Educational Psychology*, 21, 261-278.

Lorch, R. F., Jr., Lorch, E. P., & Inman, W. E. (1993). Effects of signaling topic structure on text recall. *Journal of Educational Psychology*, 85, 281-290.

Lorch, R. F., Jr., Lorch, E. P., Ritchey, K., McGovern, L., & Coleman, D. (2001). Effects of headings on text summarization. *Contemporary Educational Psychology*, 26, 171-191.

Loyd, B. H., & Steele, J. L. (1986). Assessment of reading comprehension: A comparison of constructs. *Reading Psychology: An International Quarterly*, 7, 1-10.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2, 159-165.

Lyotard, J. F. (1984). *The postmodern condition: A report on knowledge*. Manchester: Manchester University Press.

MacCann, R., Eastment, B., & Pickering, S. (2002). Responding to free response examination questions: Computer versus pen and paper. *British Journal of Educational Technology*, 33, 173-188.

Maclellan, E. (1997). Reading to learn. *Studies in Higher Education*, 22, 277-288.

Mahoney, D., Hill, J., & Shallow, J. (1997). Storing simple stories: Narrative recall and the Chinese student. *Language, Culture and Curriculum*, 10, 66-87.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon, England: Multilingual Matters.

— (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19, 85-104.

Mani, I., & Maybury, M. T. (1999). Advances in automatic text summarization. Cambridge, Massachusetts: The MIT Press.

Markham, L. R. (1976). Influence of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 13, 277-283.

Mast, C. O. (1988). *The effects of cognitive styles on summarization of expository text*. Unpublished PhD thesis, University of North Texas.

McAnulty, S. J. (1981). Paraphrase, summary, précis: Advantages, definitions, models. *Teaching English in the Two-Year College*, 8, 47-51.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 15, 323-337.

McNamara, T. (2001a). Language assessment as social practice: Challenges for research. *Language Testing*, 18, 333-349.

McNamara, T. F. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics*, 18, 304-319.

— (2001b). Language assessment as social practice: Challenges for research. *Language Testing*, 18, 333-349.

McNamara, T. F., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, 22, 221-242.

Messick, S. (1989). Validity. In R. Linn, L (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

— (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research (6th edition)*. New York: Macmillan.

— (1996). Validity and washback in language testing. *Language Testing*, 13, 241-256.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam: North-Holland.

Meyer, B. J. F., & Freedle, R., O (1984). Effects of discourse type on recall. *American Educational Research Journal*, 21, 121-143.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 92-114). Cambridge: Cambridge University Press.

Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 335-354.

Mills, C. B., Diehl, V. A., Birkmire, D. P., & Mou, L. C. (1993). Procedural text: Predictions of importance ratings and recall by models of reading comprehension. *Discourse Processes*, 16, 279-315.

Moore, T. (1997). From text to note: Cultural variation in summarization practices. *Prospect*, 12, 54-63.

Mossenson, L., Hill, P., & Masters, G. (1987). *Tests of reading comprehension (TORCH)*: Australian Council for Educational Research.

Myers, R. (1990). *Classical and modern regression with applications*. Boston, MA: Duxbury Press.

Namukwai, V., & Williams, R. (1988). The readability of government reports: A case study from Zambia. *Reading in a Foreign Language*, 5, 205-207.

Nicholas, L. E., & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research*, 38, 145-156.

Nuttall, C. (1996). *Teaching reading skills in a foreign language*. Oxford: Macmillan Heinemann.

O'Mallan, R. P., Foley, C. L., & Lewis, C. D. (1993). Effects of the guided reading procedure on fifth graders' summary writing and comprehension of science text. *Reading Improvement*, 30, 194-201.

O'Sullivan, B., Weir, C. J., & Jin, Y. (2004). Does the computer make a difference? An investigation into the differences between writing on computer and on paper. Report on IELTS research project. London: University of Roehampton.

Olssen, M. (1999). *Michel Foucault: Materialism and education*. London: Bergin & Garvey.

Paulson, P. A. (1972). *Paragraph summarization by computer: A comparison with summaries made by pupils*. Unpublished PhD thesis, Stanford University.

Pearson, P. D. (1981). A retrospective reaction to prose comprehension. In C. M. Santa & B. L. Hayes (Eds.), *Children's prose comprehension: Research and practice* (pp. 117-132). Newark, DE: International Reading Association.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design, and analysis*. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Peeck, J., & Knippenberg, W. J. (1977). Test expectancy and test performance. *Tijdschrift voor Onderwijsresearch*, 2, 270-274.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing*, 7, 143-164.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, 4, 72-92.

Pollitt, A. P. (1993). Summary completion and assessing the ability to comprehend. EFL research report. Cambridge: UCLES.

Powell, G. H., & Isaacson, D. (1984). Effects of text structure on children's recall of science text, *Paper presented at the Annual Meeting of the National Reading and Language Arts Educators' Conference (1st, Kansas City, MO, September 26-28, 1984).*

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think of less of my handwritten essay if others words process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement*, 31, 220-233.

Pressley, M., Johnson, C. J., Symons, S., McGoldrick, J. A., & Kurita, J. A. (1989). Strategies that improve children's memory and comprehension of text. *Elementary School Journal*, 90, 3-32.

Pyrczak, F. (1974). Passage-dependence of items designed to measure the ability to identify the main ideas of paragraphs: Implications for validity. *Educational and Psychological Measurement*, 34, 343-348.

Radmacher, S. A., & Latosi-Sawin, E. (1995). Summary writing: A tool to improve student comprehension and writing in psychology. *Teaching of Psychology*, 22, 113-115.

Ratteray, O. M. T. (1985). Expanding roles for summarized information. *Written Communication*, 2, 457-472.

Rea-Dickins, P. M. (1997). So why do we need relationships with stakeholders in language testing. A view from the UK. *Language Testing*, 14, 304-314.

Rea-Dickins, P. M., Kiely, R., & Yu, G. (2005). Student identity, learning and progression (SILP): With specific reference to the affective and academic

impact of IELTS on 'successful' IELTS students (pp. 124). Bristol: Graduate School of Education, University of Bristol.

Rickards, J. P., Fajen, B. R., Sullivan, J. F., & Gillespie, G. (1997). Signaling, notetaking, and field independence-dependence in text comprehension and recall. *Journal of Educational Psychology*, 89, 508-517.

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, 13, 173-189.

Rivard, L. P. (2001). Summary writing: A multi-grade study of French-immersion and Francophone secondary students. *Language, Culture and Curriculum*, 14, 171-186.

Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton: University of Princeton Press.

Rosenau, P. M. (1992). *Post-modernism and the social sciences: Insights, inroads and intrusions*. Princeton: Princeton University Press.

Rost, M. (1990). *Listening in language learning*. New York: Longman.

Royer, J. M. (1990). The sentence verification technique: A new direction in the assessment of reading comprehension. In S. Legg & J. Algina (Eds.), *Cognitive assessment of language and math outcomes*. Norwood, NJ: Ablex.

Ruddell, R. B., & Boyle, O. F. (1989). A study of cognitive mapping as a means to improve summarization and comprehension of expository text. *Reading Research & Instruction*, 29, 12-22.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding* (pp. 211-236). New York: Academic Press.

— (1977). Understanding and summarizing brief stories. In D. L. Berge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 255-303). Hillsdale, NJ: Lawrence Erlbaum.

Russell, M. (1999). Testing writing on computers: A follow-up study comparing performance on computer and on paper. Available online: Http://epaa.Asu.Edu/epaa/v7n20/. *Education Policy Analysis Archives*, 7.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. Available online: Http://epaa.Asu.Edu/epaa/v5n3.Html. *Education Policy Analysis Archives*, 5.

— (2000). Bridging the gap between testing and technology in schools. Available online: Http://epaa.Asu.Edu/epaa/v8n19.Html. *Education Policy Analysis Archives*, 8.

Russell, P. (1994). Investigating summary typology: Considerations for classification. *Technostyle*, 11, 37-47.

Rybczynski, M. A. (1987). *The effects of peer collaboration and summarization instruction on the ability of grade six students to learn from expository prose*. Unpublished PhD thesis, University of Minnesota.

Sanchez, R. P., Lorch, E. P., & Lorch, R. F., Jr. (2001). Effects of headings on text

processing strategies. *Contemporary Educational Psychology*, 26, 418-428.

Sarig, G. (1989). Testing meaning construction: Can we do it fairly. *Language Testing*, 6, 77-94.

Savignon, S. (2003). Teaching English as communication: A global perspective. *World Englishes*, 22, 55-66.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in second language. *Language Learning and Technology*, 5, 38-59.

— (2003). *A comparison of summarization and free recall as reading comprehension tasks in web-based assessment of Japanese as a foreign language*. Unpublished PhD thesis, University of California, Los Angeles (UCLA).

Schellings, G. L. M., Van Hout Wolters, B. H. A. M., & Vermunt, J. D. (1996). Individual differences in adapting to three different tasks of selecting information from texts. *Contemporary Educational Psychology*, 21, 423-446.

Schnotz, W. (1983). On the influence of text organization on learning outcomes. In G. Rickheit & M. Bock (Eds.), *Psycholinguistic studies in language processing* (pp. 152-181). Berlin and New York: de Gruyter.

Schwarz, M. N. K., & Flammer, A. (1981). Text structure and title: Effects on comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 20, 61-66.

Scott, M. L., Stansfield, C. W., & Kenyon, D. M. (1996). Examining validity in a performance test: The listening summary translation exam (LSTE) - Spanish version. *Language Testing*, 13, 83-109.

Seidlhofer, B. (1991). *Discourse analysis of summarization*. Unpublished PhD Dissertation, Institute of Education, University of London.

— (1995). *Approaches to summarization: Discourse analysis and language education*. Tübingen: Gunter Narr.

Selinger, B. M. (1995). Summarizing text: Developmental students demonstrate a successful method. *Journal of Developmental Education*, 19, 14-16,18,20.

Shaw, S. D. (2003). Legibility and the rating of second language writing: The effect on examiners when assessing handwritten and word-processed scripts. *Research Notes*, 11, 7-10.

Shermis, M. D., & Lombard, D. (1998). Effects of computer-based test administrations on test anxiety and performance. *Computers in Human Behavior*, 14, 111-123.

Sherrard, C. (1986). Summary writing: A topographical study. *Written Communication*, 3, 324-343.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, 18, 303-325.

— (2004). Textual borrowing in second-language writing. *Written Communication*, 21, 171-200.

Shih, M. (1992). Beyond comprehension exercises in the ESL academic reading class. *TESOL Quarterly*, 26, 289-318.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1, 147-170.

— (2001a). Democratic assessment as an alternative. *Language Testing*, 18, 373-391.
— (2001b). *The power of tests: A critical perspective on the uses of language tests.* London: Pearson Education.
Smiley, S. S., Oakley, D. D., Worthen, D., Campione, J. C., & Brown, A. L. (1977). Recall of thematically relevant material by adolescent good and poor readers as a function of written versus oral presentation. *Journal of Educational Psychology*, 69, 381-387.
Smith, C. B. (1988). Does it help to write about what you're reading? *Journal of Reading*, 32, 276-285.
Stansfield, C. W., Scott, M. L., & Kenyon, D. M. (1990). Listening summary translation exam (LSTE)-Spanish. Final project report. Revised. Washington, DC: Center for Applied Linguistics. ED323786.
Stansfield, C. W., Wu, W. M., & Liu, C.-C. (1997). Listening summary translation exam (LSTE) in Taiwanese (also known as) Minnan, Southern Fukienese, Southern Min, Xiamen, Amoy. Final project report. Ed413788. Bethesda, MD: Second Language Testing, Inc.
Stansfield, C. W., Wu, W. M., & van der Heide, M. (2000). A job-relevant listening summary translation exam in Minnan. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th LTRC* (pp. 177-200). Cambridge: Cambridge University Press with University of Cambridge Local Examinations Syndicate.
Steele, J. L. (1985). *Recall and comprehension: The interactive relationship of text and reader*. Unpublished PhD thesis, University of Virginia.
Stein, B. L., & Kirby, J. R. (1992). The effects of text absent and text present conditions on summarization and recall of text. *Journal of Reading Behavior*, 24, 217-232.
Stevens, J. (2002). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ.: Lawrence Erlbaum Associates.
Stordahl, K. E., & Christensen, C. E. (1956). The effect of study techniques on comprehension and retention. *Journal of Educational Research*, 49, 561-570.
Swales, J. M., & Feak, C. B. (1994). *Academic writing for graduate students: Essential tasks and skills---- a course for nonnative speakers of English.* Ann Arbor: The University of Michigan Press.
Swoope, K. F., & Johnson, C. S. (1988). A reexamination of effects of reader- and text-based factors on priority judgments in expository prose. *Journal of Educational Research*, 82, 5-9.
Tabachnick, B. G., & Fidell, L. S. (2003). *Using multivariate statistics*. Boston: Allyn and Bacon.
Taylor, B. M. (1982). A summarizing strategy to improve middle grade students' reading and writing skills. *Reading Teacher*, 36, 202-205.
Taylor, C., Jamieson, J., & Eignor, D. (2000). Trends in computer use among international students. *TESOL Quarterly*, 34, 575-585.
Taylor, C., Jamieson, J., Eignor, D. R., & Kirsch, I. (1998). The relationship between computer familiarity and performance on computer-based TOEFL test tasks

(TOEFL research report 61). Princeton, NJ: Educational Testing Service.

Taylor, C., Kirsch, I., Eignor, D. R., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning*, 49, 219-274.

Taylor, K. K. (1986). Summary writing by young children. *Reading Research Quarterly*, 21, 193-208.

Taylor, L. B. (1996). *An investigation of text-removed summary completion as a means of assessing reading comprehension.* Unpublished PhD thesis, University of Cambridge.

Thomas, S., & Bridge, C. A. (1980). A comparison of subjects' cloze scores and their ability to employ macrostructure operations in the generations of summaries. In M. L. Kamil & A. J. Moe (Eds.), *Perspectives on reading research and instruction. Twenty-ninth yearbook of the national reading conference* (pp. 69-77). Washington, DC: National Reading Conference.

Thorndyke, P. W. (1975). *Cognitive structure in human story comprehension and memory.* Unpublished PhD thesis, Stanford University.

Trabasso, T., Secco, T., & van den Broek, P. (1984). Causal cohesion and story coherence. In H. Mandl, N. L. Stein & T. Trabasso (Eds.), *Learning and comprehension of text* (pp. 83-111). Hillsdale, NJ: Lawrence Erlbaum Associates.

Trabasso, T., & Sperry, L. (1985). Causal relatedness and importance of story events. *Journal of Memory and Language*, 24, 595-611.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language*, 24, 612-630.

Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing*, 22, 174-210.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, 36, 49-70.

Upton, T. A. (1993). *The influence of first and second language use on the comprehension and recall of written English texts by Japanese readers.* Unpublished PhD thesis, University of Minnesota.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. London: Longman.

Valencia, S., & Pearson, P. D. (1987). Reading assessment: Time for a change. *The Reading Teacher*, 40, 726-732.

Valette, R. M. (1977). *Modern language testing*. New York: Harcourt Brace Jovanovich, Inc.

van den Broek, P. (1988). The effects of causal relations and hierarchical position on the importance of story statements. *Journal of Memory and Language*, 27, 1-22.

van den Broek, P., & Trabasso, T. (1986). Causal networks versus goal hierarchies in summarizing text. *Discourse Processes*, 9, 1-15.

van Dijk, T. A. (1977). Semantic macro-structures and knowledge frames in discourse

comprehension. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 3-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

— (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

van Dijk, T. A., & Kintsch, W. (1977). Cognitive psychology and discourse: Recalling and summarizing stories. In W. U. Dressler (Ed.), *Current trends in textlinguistics* (pp. 61-80). New York: de Gruyter.

— (1983). *Strategies of discourse comprehension*. New York: Academic Press.

van Elmpt, M., & Loonen, P. (1998). Open questions: Answers in the foreign language? *Toegepaste Taalwetenschap in Artikelen*, 58, 149-154.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex Publishing Corporation.

Wade, S. E., Buxton, W. M., & Kelly, M. (1999). Using think-alouds to examine reader-text interest. *Reading Research Quarterly*, 34, 194-216.

Ward, A. M., & Xu, L. (1994). The relationship between summarizaton skills and TOEFL scores. Paper presented at the annual meeting of the teachers of English to speakers of other languages (28th, baltimore, md, march 8-12, 1994). Ed394332.

Weaver, C. A., & Kintsch, W. (1991). Expository text. In R. Barr, M. L. Kamil, P. B. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research (vol. 2)* (pp. 230-245). New York: Longman.

Wegner, M. L., Brookshire, R. H., & Nicholas, L. E. (1984). Comprehension of main ideas and details in coherent and noncoherent discourse by aphasic and nonaphasic listeners. *Brain and Language*, 21, 37-51.

Weir, C. J. (1988). *Communicative language testing with special reference to English as a foreign language*. Exeter: University of Exeter.

— (1993). *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall International (UK) Ltd.

— (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J., Yang, H., & Jin, Y. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes*. Cambridge: Cambridge University Press with University of Cambridge Local Examinations Syndicate.

Welling-Slootmaekers, M. (1999). Language examinations in Dutch secondary schools from 2000 onwards. *Levende Talen*, 542, 488-490.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge: Cambridge University Press.

Williams, J. P. (1988). Identifying main ideas: A basic aspect of reading comprehension. *Topics in Language Disorders*, 8, 1-13.

Wilson, B. H. (1984). *The relationship of field dependence-independence and prior knowledge of passage content to recognition of main ideas and details in*

illustrated and nonillustrated expository text. Unpublished PhD thesis, University of Wisconsin, Madison.

Winograd, P. N. (1982). *An examination of strategic differences in summarizing texts*. Unpublished PhD thesis, University of Illinois at Urbana-Champaign.

— (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, 19, 404-425.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal*, 77, 473-489.

Wolf, D. F., Bixby, J., Glenn, J. I., & Gardener, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education*, 17, 31-74.

Wu, W. M., & Stansfield, C. W. (2001). Towards authenticity of task in test development. *Language Testing*, 18, 187-206.

Yang, L., & Shi, L. (2003). Exploring six MBA students' summary writing by introspection. *Journal of English for Academic Purposes*, 2, 165-192.

Yu, G. (2005). *Reading-to-summarize written discourse: A bibliography database*. Bristol: University of Bristol.

Zou, S., Weir, C. J., & Green, R. (1998). *The Test for English Majors validation project*. Shanghai: Shanghai Foreign Languages Education Press.

Zuck, L. V., & Zuck, J. G. (1984). The main idea: Specialist and non-specialist judgements. In J. Ulijn & A. Pugh (Eds.), *Reading for professional purposes: Studies and practices in native and foreign languages* (pp. 130-135). London: Heinemann.

Appendix 1.A: Computer Familiarity Questionnaire (English version)

Computer Familiarity Questionnaire

Your Name: _____ Test-taker ID: _____

Gender: Male ○ Female ○

University: _____ Department_____

Year _____

This questionnaire is to help the present researcher find out about the different ways you use computers and related technologies. I would be very grateful if you could answer **independently** all the questions on the **3 pages.** All data will be kept strictly confidential to the researcher, and be protected by the UK Data Protection Act 1998. The data you provide in the questionnaire will only be used in the present PhD research regarding computer based language tests. Thank you very much for your time.

Now read the questions below and fill in **ONE** circle for each question where appropriate.

How often is there a computer available for you to use at these places?	**4 times a month or more often**	**between 2 and 3 times a month**	**less than once a month**	**never**
1. at home	○	○	○	○
2. at your university computer labs	○	○	○	○
3. outside the university (eg. at Internet Café, friend's home)	○	○	○	○

How long ago did you get your first computer at these places?	**over 3 years ago**	**between 1 and 3 years ago**	**less than 1 year ago**	**not available**
4. at home	○	○	○	○
5. in university accommodation	○	○	○	○

How familiar are you with using/doing these things?	**very familiar**	**familiar**	**a little familiar**	**not at all familiar**
6. using a computer	○	○	○	○
7. using a "mouse" (ball and touch pad)	○	○	○	○
8. word processing in English	○	○	○	○
9. word processing in Chinese	○	○	○	○
10. reading from a computer screen	○	○	○	○

11. How many examinations/tests have you taken on a computer?	**more than 5** ⬭	**3 or 4** ⬭	**1 or 2** ⬭	**none** ⬭
12. How would you rate your ability to use a computer generally?	**excellent** ⬭	**good** ⬭	**fair** ⬭	**poor** ⬭

How often do you use these?	**4 times a month or more often**	**between 2 and 3 times a month**	**less than once a month**	**never**
13. a mobile phone	⬭	⬭	⬭	⬭
14. an automatic banking machine (ATM)	⬭	⬭	⬭	⬭
15. a computer	⬭	⬭	⬭	⬭

If you answered "NEVER" to question 15, **STOP** here please. Thanks again for your time! ☺☺☺
If you use a computer, please continue➔

How often do you use or do these?	**4 times a month or more often**	**between 2 and 3 times a month**	**less than once a month**	**never**
16. the internet	⬭	⬭	⬭	⬭
17. multimedia to watch VCD/DVD program on a computer	⬭	⬭	⬭	⬭
18. send or receive electronic mail (E-mail)	⬭	⬭	⬭	⬭
19. a "mouse" (including ball and touch pad)	⬭	⬭	⬭	⬭
20. games on computer	⬭	⬭	⬭	⬭
21. word process in Chinese	⬭	⬭	⬭	⬭
22. word process in English	⬭	⬭	⬭	⬭
23. spreadsheets (e.g., Microsoft Excel®)	⬭	⬭	⬭	⬭
24. graphics	⬭	⬭	⬭	⬭
25. participate in "chat room"	⬭	⬭	⬭	⬭

How often do you do the following things if you are stuck when using a computer?	**always**	**frequently**	**occasionally**	**never**
26. play around to see if I myself can sort it out	⬭	⬭	⬭	⬭
27. use the help button in the program	⬭	⬭	⬭	⬭
28. use a manual or refer to computer magazines	⬭	⬭	⬭	⬭
29. go to the Internet for help	⬭	⬭	⬭	⬭
30. turn-off/re-set the computer and start again	⬭	⬭	⬭	⬭
31. give up	⬭	⬭	⬭	⬭

32. Have you received any computer training in your current university over the past **two** years?

a. No ⬭

b. Yes ⬭ [please indicate the contents of the trainings, in Chinese and/or English]

_____

_____

_____

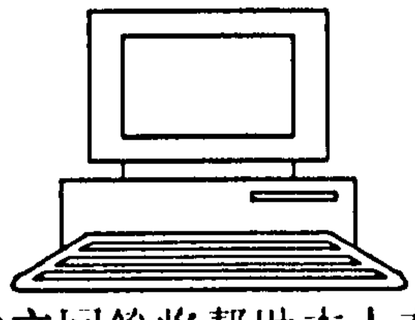
_____

33. Please use this space to record anything else you want to say about how familiar you are with computer technologies, e.g. the national or provincial computer examinations you've passed, and your attitudes (either positive or negative) towards using computers.

Many thanks again for your time!

Appendix 1.B: Computer Familiarity Questionnaire (Chinese version)

计算机熟练程度调查问卷



您的姓名: _____性别: 男○ 女○考生号:_____

所在大学_____ 所学专业_____

年级_____

此问卷, 共二页, 旨在了解您使用计算机及其相关技术的熟练程度; 您的独立回答将帮助本人在英国布利斯托大学所作的语言测试的研究. 本人将严格遵循英国数据保护法(1998)以及布利斯托大学数据安全的有关规定, 保障您在本问卷中提供的任何信息的安全特权. 您的信息只用于此项研究, 并在完成研究后两年内全部销毁. 非常感谢您的宝贵时间!

现在请您仔细阅读下列问题并在最适合您的答案上图圈(注意: 每题只选一个)

在下列地方您能使用计算机的机会是?	每月 4 次以上	每月 2 到 3 次	每月小于 1 次	没有
1. 在家里	○	○	○	○
2. 在学校计算机房	○	○	○	○
3. 在校园外 (如: 网吧, 朋友家)	○	○	○	○

您多久前, 在下列地方, 开始拥有自己(包括共同拥有)的第一台计算机?	3 年以上	1 到 3 年	1 年以下	没有计算机
4. 在家里	○	○	○	○
5. 在学校宿舍	○	○	○	○

您对下列物件的熟练程度如何?	非常熟练	熟练	有些熟练	一点不熟练
6. 使用计算机	○	○	○	○
7. 使用鼠标 (球, 触摸板)	○	○	○	○
8. 使用计算机进行英语文字处理	○	○	○	○
9. 使用计算机进行汉语文字处理	○	○	○	○
10. 在计算机屏幕上阅读文章	○	○	○	○

11. 您参加过多少次计算机化(即无纸化)考试?	5 次以上	3 或 4 次	1 或 2 次	0 次
	○	○	○	○

12. 您如何评价自己的计算机总体应用能力?	很棒	好	还可以	差
	○	○	○	○

您使用下列物件的频率是?	每月 4 次以上	每月 2 到 3 次	每月小于 1 次	没有
13. 移动电话	○	○	○	○

14. 自动柜员机(ATM)	○	○	○	○
15. 计算机	○	○	○	○

如果您第 15 题回答了“没有”，请停止. 再次感谢您的宝贵时间! ☺☺☺
如果您第 15 题没有回答“没有”，请继续➔

您使用或从事下列物件的频率是?	每月 4 次以上	每月 2 到 3 次	每月小于 1 次	没有
16. 上网	○	○	○	○
17. 在计算机上观看 VCD/DVD 节目	○	○	○	○
18. 发送或接受电子邮件(E-mail)	○	○	○	○
19. 使用鼠标 (球, 触摸板)	○	○	○	○
20. 在计算机上玩电子游戏	○	○	○	○
21. 使用计算机进行汉语文字处理	○	○	○	○
22. 使用计算机进行英语文字处理	○	○	○	○
23. 数据报表 (如: Microsoft Excel®)	○	○	○	○
24. 图案	○	○	○	○
25. 参与聊天室	○	○	○	○

使用计算机碰到问题时，您使用下列物件的频率是?	总是	通常	很少	没有
26. 自己摸索，看能否自己解决问题	○	○	○	○
27. 使用相关软件的帮助功能	○	○	○	○
28. 使用相关手册或计算机杂志	○	○	○	○
29. 上网搜索寻求帮助	○	○	○	○
30. 关机，冷启动或重置(re-set)，热启动	○	○	○	○
31. 放弃	○	○	○	○

32. 在过去两年内您是否受过计算机培训?

a. 不 ○

b. 是 ○ [请说明培训内容，可用汉语或英语]

_____

_____

33. 另请说明本问卷没有涉及，但能够表明您的计算机熟练程度的有关情况，如:
您已经获得的计算机等级证书，以及您对使用计算机的(积极的或消极的)态度等等.

谢谢您的合作!

Appendix 2.A: TextA for the Summarization Tasks

History of Education in Southeast Asian Countries

Indigenous culture, colonialism, and the post-World War II era of political independence influenced the forms of education in the nations of Southeast Asia -- Myanmar (Burma), Kampuchea (Cambodia), Indonesia, Laos, Malaysia, the Philippines, Singapore, Thailand, and Vietnam.

Before AD 1500, education throughout the region consisted chiefly of the transmission of cultural values through family and community living, supplemented by some formal teaching of each locality's dominant religion -- animism, Hinduism, Buddhism, Taoism, Confucianism, or Islam. Religious schools typically were attended by boys living in humble quarters at the residence of a pundit who guided their study of the scriptures for an indeterminate period of time.

With the advent of Western colonization after 1500, and particularly from the early 19th to mid-20th century, Western schooling with its dominantly secular curriculum, sequence of grades, examination, set calendar, and diplomas began to make strong inroads on the region's traditional educational practices. For the indigenous peoples, Western schooling had the appeal of leading to employment in the colonial government and in business and trading firms.

After World War II, as all sectors of Southeast Asia gained political independence, each newly formed nation attempted to achieve planned development -- to furnish primary schooling for everyone, extend the amount and quality of postprimary education, and shift the emphasis in secondary and tertiary education from liberal, general studies to scientific and technical education. Although indigenous culture was still learned through family living and traditional religion continued to be important in people's lives, most formal schooling throughout Southeast Asia had become predominantly of a Western, secular variety.

Schooling in all of these countries was organized in three main levels, primary, secondary, and higher. In addition, nursery schools and kindergartens, operated chiefly by private groups, were gradually gaining popularity. The typical length of primary schooling was six years. Secondary education was usually divided into two three-year levels. A wide variety of postsecondary institutions offered academic and vocational specializations. Beginning in the 1950s, nonformal education to extend literacy and

vocational skills among the adult population expanded dramatically throughout the region. Most of the nations were committed to compulsory basic education, typically for six years but up to nine years in Vietnam. However, by the close of the 1980s, the inability of governments to furnish enough schools for their growing populations prevented most from fully realizing the goal of universal basic schooling.

In each nation a central ministry of education set schooling structures and curriculum requirements, with some responsibilities for school supervision, curriculum, and finance often delegated to provincial and local educational authorities. Government-sponsored educational research and development bureaus had been established since the 1950s in an effort to make the countries more self-reliant in fashioning education to their needs. Regional cooperation in attacking educational problems was furthered by membership in such alliances as the Southeast Asian Ministers of Education Organization (SEAMEO) and the Association of Southeast Asian Nations (ASEAN).

Problems which most Southeast Asian education systems continued to face were those of reducing school dropout and grade-repeater rates, providing enough school buildings and teachers to serve rapidly expanding numbers of children, furnishing educational opportunities to rural areas, and organizing curricula and the access to education in ways that suited the cultural and geographical conditions of multiethnic populations.

Myanmar (formerly Burma). The indigenous system of education in Myanmar consisted mainly of Buddhist monastic schools of both primary and higher levels. They were based on (1) the moral code of Buddhism, (2) the divine authority of the kings, (3) the institution of *myothugyi* (township headmen), and (4) widespread male literacy. The Western system was established after the British occupation in 1886. The new system recognized women's right to formal education in public schools, and women began to play an increasingly important role as teachers. The Government College at Rangoon and the Judson College established in the 19th century were incorporated as the University of Rangoon under the University Act of 1920.

Following independence in 1948, the country experienced more than a decade of political instability until a coup d'etat in 1962 brought a strongly centralized socialist government to power. Subsequently, marked improvements in education occurred. Science was emphasized along with general academic subjects, civic education, and practical arts. Primary-school attendance for children ages five through nine became

free where available. From 1965 to 1985 enrollments increased in primary schools from two to five million, in secondary schools from 503,000 to 1.25 million, and in higher education from 21,000 to 189,000.

Malaysia and Singapore. The Malay States, Singapore, and sectors of North Borneo were British colonies until reorganized as the nation of Malaysia in 1963. Singapore left the coalition in 1965 to become an independent city-nation. As a result, while Malaysia and Singapore share common educational roots, their systems have diverged since 1965.

Under British rule, the most significant feature of education on the Malay peninsula was the structuring of primary schools in four language streams -- Malay, Chinese, English, and Tamil. Students in the English stream enjoyed favoured access to secondary and higher education as well as to employment in government and commerce. After 1963 Malaysian leaders sought to indigenize and unify their society by adopting the Malay language as the medium of instruction in schools beyond the primary level and by teaching English only as a second language. In contrast, the government of Singapore urged everyone to learn English, plus one other local tongue – Chinese, Malay, or Tamil. Thus, in both nations the learning of languages became a critical issue in people's efforts to gain access to socio-economic opportunity and in political leaders' attempts to unify their multiethnic populations.

Efforts to popularize schooling in Malaysia and Singapore were notably successful. By the early 1980s, 93 percent of all Malaysian children ages six to 11 attended primary school, with nearly 90 percent of primary-school graduates entering lower-secondary school. By 1968, all primary-age children in Singapore were in school. In both countries, secondary- and higher- education enrollments continued to increase rapidly. Both nations were well supplied with school buildings, textbooks, and trained teachers.

Indonesia. From AD 100 to 1500 the Indonesian aristocracy adopted Hindu and Buddhist teachings, while education for the common people was provided mainly informally, through daily family living. Islam, introduced into the archipelago around 1300, spread rapidly in the form of Qur'an schools, which have continued through the 20th century, though in diminishing numbers. The first few schools on Western lines were established by Portuguese and Spanish priests in the 16th century. As the Dutch colonialists gained increasing control over the islands, they set up schools patterned after those in Holland, primarily for European and Eurasian pupils. In 1848 the Dutch

East Indies government officially committed itself to providing education for the native population. However, even though the amount of education for indigenous islanders increased over the following century, Western schooling under the Dutch never reached the majority of the population.

After Indonesians gained independence from the Dutch in 1949, they sought to provide universal elementary schooling and a large measure of secondary and higher education. Progress toward this goal after 1950 was rapid, despite the challenge of an annual population growth rate of around 2.3 percent. Enrollments over the 1950-1985 period increased from five million to 30 million at the elementary level, from 230,000 to 7.5 million at the secondary level, and from 6,000 to one million at the tertiary level. Although the Indonesian population was 90 percent Muslim, three-fourths of the nation's schools were of Western secular variety. The remaining one-fourth were Islamic schools required to offer at least 70 percent secular studies and no more than 30 religious subjects. This ratio reflected the government's efforts to use the schools for preparing manpower for socio-economic modernization, as guided by a sequence of five-year national development plan.

Philippines. The pre-Spanish Philippines possessed a system of writing similar to Arabic, and it was not uncommon for adults to know how to read and write. Inculcation of reverence for the god Bathala, obedience to authority, loyalty to the family or clan, and respect for truth and righteousness were the chief aims of education. After the Spanish conquest, apart from parochial schools run by missionaries, the first educational institutions to be established on Western lines were in higher education. The Santo Tomas College, established in 1611 and raised to the status of a university in 1644-1645, served for centuries as a centre of intellectual strength to the Filipino people. Education growth, however, was slow, mainly because of lack of government support.

With the advent of American rule, the stress laid on universal primary education in the policy announced by U.S. President William McKinley on April 7, 1900, led to a rapid growth in primary education. A number of institutions of higher education were also established between 1907 and 1941, including the University of the Philippines (1908). Private institutions of higher education, however, far outnumbered the state institutions, thus indicating a trend that remains a characteristic feature of the system of higher education in the Philippines.

The new Republic of the Philippines emerging after World War II launched a

series of national development plans that included components aimed at the renovation and expansion of education to promote socio-economic modernization. Over the period 1948 to 1986, enrollments rose in primary schools from four million to nine million and in secondary schools from 424,000 to 3.3 million. By the late 1980s, 1.5 million students were in the nation's more than 1,000 higher-education institutions. More than 95 percent of primary pupils and 41 percent of secondary students attended public schools, while the remainder attended private institutions.

Thailand. The traditional system of education in Thailand was inspired by the Thai philosophy of life based on (1) dedication to Theravada Buddhism, with its emphasis on moral excellence, generosity, and moderation, (2) veneration for the king, and (3) loyalty to the family. The beginning of the present system of education can be traced to 1887, when King Chulalongkorn set up a department of education with foreign advisers, mostly English educationists. Gradually, temple schools were established. The process of westernization of education was strengthened with the establishment of a medical school in 1888, a law school in 1897, and royal pages' school in 1902 for the general education of "the sons of the nobility". It was converted into the Civil Service College in 1910.

The abolition of the absolute monarchy after the 1932 revolution stimulated the government to increase educational provisions at all levels, particularly for training specialists in higher-learning institutions. Beginning in 1962, the nation's series of five-year development plans assigned educational institutions a crucial role in manpower preparation. The government supervises all educational institutions, public and private. Financing education is primarily a government responsibility, supplemented by the private sector. Thai is the language of instruction at all levels, with English taught as a second language above grade four.

By the mid-1980s there were more than 7.3 million pupils (over 90 percent of the age group) enrolled in compulsory six-year elementary schools, 2.2 million in the six years of secondary schooling, and 715,000 in the nation's 31 registered universities and colleges.

Kampuchea (formerly Cambodia). For nearly four centuries before the advent of the French in 1863, the educational system in Cambodia grew up around Theravada Buddhism, which became the established religion toward the end of 1430 under Thai influence. In 1887 Cambodia became a part of the French Indochina Union and did not achieve complete independence until 1954. Pagoda schools,

imparting education at the primary level, were remodelled and integrated into the primary school system administered by the Ministry of Education.

Civil war throughout the 1970s disrupted education until Vietnamese forces overthrew the Khmer Rouge government in 1979. By the mid-1980s schools had reopened with a total enrollment of nearly two million throughout the four-year primary, three-year junior-secondary, and three-year senior-secondary structure. Secondary schools and the country's few higher-education colleges were in the state of rebuilding. Much of the teacher-training was in the form of short courses, and nonformal adult literacy classes multiplied at a rapid pace.

Laos. The pagoda school was the main unit of the traditional educational system in Laos. Efforts toward modernization came in the wake of the country's becoming a French protectorate in 1893 and finally after its inclusion in 1904 within the French Indochina Union. The medium of education was changed to French when the French Education Service was created.

In 1975, after 30 years of uninterrupted revolution, a socialist government was established and schooling was accorded high priority. By the mid-1980s 79 percent of all children seven to 11 years old were in the five-year primary school, 48 percent of children 12 to 14 years old were in the three-year junior-secondary school, and 23 percent of the 15- to 17-year-olds were in the three-year senior-secondary school.

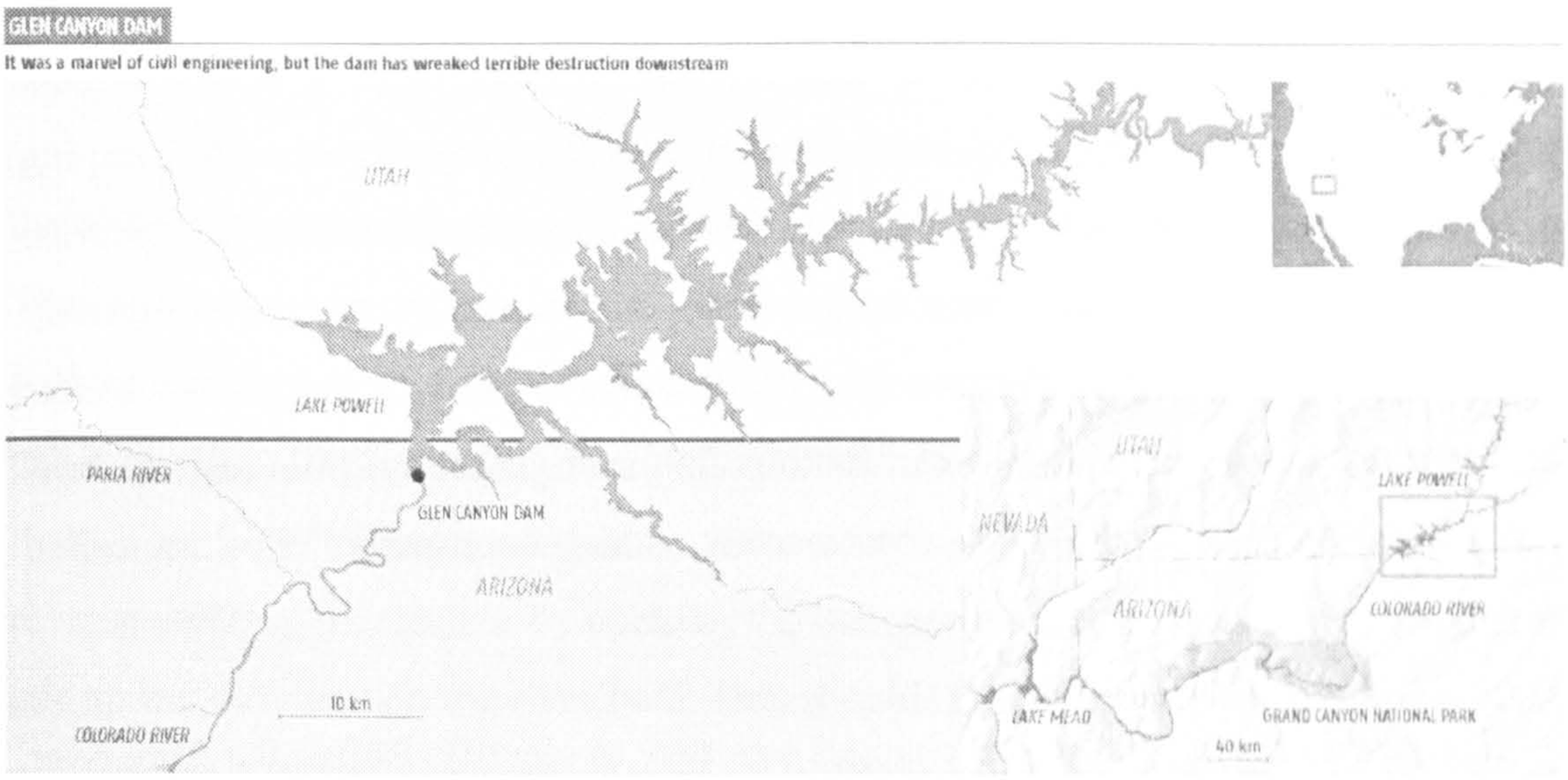
Vietnam. Long Chinese domination over the emperors of Vietnam resulted in strong Confucian and Taoist influences on the Vietnamese educational system, though it centred on Buddhism. The establishment of French rule, commencing with the occupation of Saigon (now Ho Chi Minh City) in 1895, led to the gradual growth of a pattern of education similar to that of the rest of the former Indochina Union. Vietnamese attempts to develop education were thwarted by the continued fighting from World War II onward and, after the partition of the country in 1954, by fighting between the South and the North. After the war's end in 1975, the communist government attempted to "reeducate" the conquered South and sought to establish urgently needed technical and vocational education in secondary and higher levels. By the mid-1980s there were eight million pupils in elementary schools, four million in secondary schools, and more than 115,000 in higher-education institutions.

Appendix 2.B: TextB for the Summarization Tasks

Let the River Run

The only way to undo years of severe environmental damage to the Grand Canyon is to flood it regularly, says Matt Kaplan

New Scientist vol. 175, issue 2362, 28 September 2002, page 32



In March 1996, flood waters raged through Arizona's Grand Canyon for the first time in over thirty years. Water rushed up beaches usually reserved for campsites, small trees drowned and rapids vanished from sight. For a week, the entire Colorado River was transformed into a turbulent monster.

The awe-inspiring flood was not a natural event. The water had been released from an upstream dam to reproduce the sort of flood the Grand Canyon would have experienced every year during winter and spring rains, before the river was dammed in 1963.

The reason for the controlled flood was not nostalgia. It was an attempt to undo years of environmental damage caused by the dam. Although it didn't go entirely to plan, it was mostly a success and the lessons learned will be invaluable in planning the next flood, which is scheduled for later this year. The hope is that this flood will

make a really good job of restoring the Grand Canyon to its pre-dam condition. That's if short-sighted politics doesn't prevent the flood happening at all.

Dams have been on the Colorado River for nearly a century. However, there were no dams upstream of the Grand Canyon until 1963, when the Glen Canyon Dam was built. At the time, the idea of damming the river above the canyon didn't alarm anyone: it was a dam-building era and cheap, clean hydroelectricity seemed like a good thing.

But by the late 1970s, serious long-term effects were becoming apparent. Downstream of the dam, the Grand Canyon was suffering from a dearth of new sediment. People rafting down the river found the beaches they were used to camping on had eroded to half their original size. Meanwhile, the National Park Service noticed that the Colorado River's largest fish, the pikeminnow, had completely disappeared from the river system. And the humpback chub, a fish native to the Colorado, was endangered. Alien carnivorous fish such as brown trout and rainbow trout had spread through the entire river system, as had a non-native river tree called the tamarisk.

The reason for the geological changes was easy to identify. Before the dam, river water in the canyon carried so much sediment that the river actually had a cloudy rust colour, which earned it the name Colorado - Spanish for reddish - as it flowed through an area that used to be part of Mexico. But now the water is forced to come to a halt behind the dam before being released into the Grand Canyon. Over 90 per cent of the river's sediment is dropped here, leaving the river crystal clear. This is bad news for young humpback chub, which use cloudy water to hide from predators such as the carnivorous trout. Clear water makes the chub an easy target.

The lack of sediment also explains the beach erosion. Beaches depend on annual floods bringing a continuous supply of fresh fine sand, but the dam traps the majority of this. Small tributary rivers flowing unimpeded into the Colorado River below the dam bring some sediment, but not enough.

The dam has caused other geological problems. By limiting the river to a steady low flow all year round, the dam ended the canyon's seasonal floods. Rapids, which

depend upon flood waters to clear any debris from between the boulders, became choked.

These conditions probably set the chub back further, because they like constant, turbulent waters. Trout, on the other hand, thrive in the clear, placid pools between the rapids. "Chub's decline over the past 10 years correlates with increases in the trout population," says aquatic ecologist Michael Yard of the US Geological Survey's Grand Canyon Monitoring and Research Center in Flagstaff, Arizona. Yard says the chub suffered and the trout thrived most when demand for power was low and less water was run through the turbines into the canyon.

In 1989, under growing public concern and pressure from environmental groups, the US Bureau of Reclamation in Washington DC sent a task force into the canyon. The result was an environmental impact statement highlighting a multitude of problems in the Grand Canyon ecosystem. The team said that sediment, introduced by tributary rivers, was collecting along the Colorado's bottom due to the restricted water flow. This triggered a stunning and, at the time, revolutionary idea. Why not try to reverse some of the decline by opening the dam for a short time so that flood waters stir up the sediment on the river bed? This should, the team claimed, help rebuild the sandbars and clean the silt out of the backwater channels used by native fish. In July of 1995 the desperate bureau gave the go-ahead.

The punching of a few buttons on 26 March 1996 opened the floodgates, allowing 0.7 trillion litres of water to bypass the turbines and, at a cost of $2 million in lost electricity revenue, flow into the canyon at a rate of 1300 cubic metres per second for a week, just under half the rate at which the river used to flood. The National Park Service issued safety warnings and monitored the canyon.

Geologists hoped the flood would recirculate fine sediment at the bottom of the river. Ecologists hoped the flood might also sweep away young trout, which aren't adapted for flood conditions, and clear the way for a chub recovery by reconstructing their native habitat. "It was an enormous science experiment," says Robert Webb, a hydrologist at the USGS in Flagstaff.

In the weeks after the flood waters cleared, researchers were amazed by what they saw. The flood seemed to have restored the canyon to a near pristine state. Beaches that had been rocky and barren of sand turned into beautiful sandy hills. This presumably meant that flood waters also improved the shape of the river bed. Instead of sediment sitting at the base of the river, it had been piled high into beaches and sandbars.

The most impressive changes were seen in the rapids. The debris and sediment that had been choking them off was gone. "A lot of rapids were becoming quite dangerous to the [white-water] rafters," says Webb. "The flood cleared a lot up." In particular the canyon's largest rapid, Lava Falls, saw a startling transformation. Debris that had been constricting its white water for over a year was dislodged, increasing the width of the rapid by an average of 5 metres.

However, while many of the flood's positive effects have lasted with remarkable resilience since 1996, not everything went as planned. This summer, David Rubin and David Topping, also at the USGS in Flagstaff, published work showing that the badly eroded sandbars were not rebuilt using sediment from the whole river bed (*Eos*, vol 83, p 273). Instead, the sand came only from the edges of the river bed, at the base of the same sandbars, making them narrower but higher. The river may have looked better from the bank but this was at the expense of the views from the river bed. "It was a bit like using your credit card to bump up your bank account," says Topping. The same applied to beaches, which are just sandbars above the waterline.

The flood brought similarly mediocre results for the endangered species. While the chub were not harmed in any significant way by the flood, neither were the competing trout. And the raging waters actually exacerbated the tamarisk invasion, spreading its seeds all along the canyon's banks.

So although the idea behind the flood - that it could be used to recirculate sediment - was viable, this was tempered by the discovery that not nearly as much sediment was available at the base of the river as thought. To explain this, researchers looked back at the history of flows released during the 1980s and early 90s. Although the floodgates had remained closed, power companies had released extra water through the turbines whenever demand for power was high. It now seemed likely

these flows were fast enough to pick up sediment and carry it downstream, eventually out of the canyon. But their water level was too low for the sediment to reach the tops of sandbars and beaches, says Matt Kaplinski, research associate at Northern Arizona University in Flagstaff.

Other surprises were buried in the reams of data generated by the 1996 flood. Photographs and measurements of beach and sandbar size suggested that most of the rebuilding occurred at the start of the flood. After that, the flood waters actually eroded the beaches away. Flood waters are the only force that can get sediment from the river bottom and use it to build beaches, but like any fast-flowing water, they also move sediment downstream.

Experiments, however, are all about trial and error, and the canyon's geologists are using everything they learned from the 1996 flood to guide the 2002 one. They now know that the next flood only needs to last for two days. They also know that the dam should be kept at a very low output for a period of several months before the flood, so that sediment from downstream tributaries, such as the Paria River, can build up on the river bed.

This looks like the best way to fix the canyon's geological problems. What about the ecological issues? Running the flood in winter has a second advantage: the invading tamarisk river tree will not be in seed, so the torrent won't spread the tree, and could even damage smaller tamarisks. However, the discussion is complicated. The tamarisk may compete with native plants, but it also provides a superb nesting habitat for an endangered bird, the Southwestern willow flycatcher. In spite of the name, the birds seem to prefer tamarisk to willow. "Is a species automatically bad if it is non-native?" asks Webb.

As for the unwanted trout, Yard is keen to follow up the winter flood with a series of high flows in the spring, when the trout are reproducing. These would go through the turbines instead of the floodgates and flush through sediment that geologists would rather was on the river bed. These flows should hit the trout during their spawning period, hopefully enabling the chub to recover. But geologists are also worried that repeated spring floods could erode away fresh deposits. "It is possible that these spring flows could undo everything the winter flood builds," says Kaplinski.

Agreement between scientists is elusive enough but the team that makes the decision on whether the dam should be opened includes representatives from states along the Colorado River, Native American tribes, fishers, river rafters, environmentalists and, of course, power contractors. More formally known as the Glen Canyon Dam Adaptive Management Program, the team recommends to the federal government how the dam should operate.

Not everyone in the group is backing the flood. The State of Colorado, whose rivers feed the canyon, may try to fight any decision to flood in the courts, citing concerns over water conservation. No scientists support their concerns, but it's not clear how a US administration with the stated aim of making the country more energy independent will react to this pressure. But Randy Peterson, manager of the Adaptive Management Program, thinks Colorado will probably withdraw its opposition before the argument reaches that level. "We believe we can keep law suits out of the picture," he says.

The good news is that the power companies that share in the dam's operation have agreed to the flood, despite the fact that they'll incur greater losses than in 1996. According to Peterson, the companies won't just lose around $2 million during the controlled flood itself. The reduced flow throughout the preceding autumn will allow only a small portion of the hydroelectric generators to be used, at a cost of roughly $8 million. But if the series of high flows in the following spring get the go-ahead, the companies might be able to recoup some of their losses by running more water through than normal.

Just like the 1996 flood, the 2002 one is an experiment. But if a few months of restraint followed by a two-day torrent can undo years of ecological destruction, environmentalists may clamour for floods to become a regular fixture, and not just at the Grand Canyon. Jeff Mount, director of the Watershed Center at the University of California, Davis, is watching the events closely. "California has more than 1400 dams," he says, every one of which has sediment issues. "We need experiments like the ones they are doing at the Grand Canyon to give us the courage to try our own."

But what if it doesn't work? If the tributary rivers don't bring enough sediment into the system, researchers will have to get some elsewhere. One possibility is to

dredge the sediment caught behind the dam and dump it on the other side, before attempting another deliberate flood. But tests on this sediment have indicated it is high in naturally occurring heavy metals such as selenium. Geologists may have to consider the expensive option of bringing sediment in from elsewhere.

In the end, though, it's unlikely that huge power consumption comes without some environmental cost - whether that's to the geology or the wildlife of the canyon. Though researchers may be able to mitigate the damage by controlling the dam's operations, the only sure way to return the canyon to its natural state is to decommission the dam.

Matt Kaplan

Matt Kaplan is a science writer in Los Angeles

Appendix 2.C: TextC for the Summarization Tasks

Work Life Balance

Mary O'Hara

03 March 2002, The Observer, p.1

A special supplement produced by The Observer in association with the Department of Trade and Industry: The work-life balance campaign: Redressing an imbalance: Britain's workforce work the longest hours in Europe. In answer to this, the government has launched a campaign to promote a better balance between work and home life

That the UK's workforce spend an average four hours more each week doing their job than workers in the rest of Europe will come as no surprise to anyone. But the government's initiatives to help redress the imbalance could play a pivotal part in changing the culture of working life in Britain, with benefits for employers, workers, their families and the community as a whole.

Recognising the need for a greater balance between work and home life, the government launched its Work-Life Balance campaign in March 2000, making it a priority to promote the benefits of flexible working practices to business, workers and unions. It is concerned in particular with redrawing those work practices that make it harder for the parents of young children to juggle family with employment.

As part of the push, it launched a scheme called the Work-Life Balance Challenge Fund in March 2000 with over £10m government funding for the first three years. It was set up to advise and assist businesses wishing to introduce changes that will make working practices more flexible. Those employers who want to change can apply for funding and expert advice on how to formally introduce new practices.

The fund is open to all employers in England and Scotland with a separate fund for Wales aimed at small and medium-sized enterprises. Northern Ireland also has a scheme funded by a number of government departments.

The first wave saw 88 companies qualify for consultancy and funding; in the latest round another 89 have signed up. A total of 181 companies from sectors

including transport, manufacturing, retail, finance, telecommunications and the public sector have benefited; the average funding per project is £37,000.

Independent from, but working alongside government, is an alliance of employers called Employers for Work-Life Balance, made up of business leaders who have voluntarily taken on the mantle of promoting the benefits of changing work practices to others in the business community. Members include Nationwide Building Society and Lloyds TSB.

The alliance recognises key demographic triggers for change, pushing the need for greater involvement: 'The nature of the workforce has changed dramatically. In addition, dual-earner families are now the norm, we are experiencing an ageing population and a shift in the expectations of quality of work-life by younger generations.'

The Work-Life Balance initiative is, the government says, a chance for the UK to abandon a work culture based upon excessive hours, and an attempt to follow the lead of other countries that have successfully reduced the time the average citizen spends in the workplace.

The problem caused by excessive working hours pervades all types of businesses and all levels within them. A study by the Institute of Management and the University of Manchester Institute of Science and Technology in February last year found that 75% of managers felt they needed to 'burn the candle at both ends' to stay on top of their work, and 64% believed that long hours were 'a part of their organisation's culture'.

The overarching aim of the campaign says Alan Johnson, minister of state for employment relations and regions, is 'to help change the pervading working culture which makes people feel like they have to be seen to be working very long hours to demonstrate commitment.

'This is about quality of life, about getting the best out of people, about giving people some flexibility in their lives, and it's about proving to business that it can make a tangible difference to productivity. And it can.

'We all have a stake in this. The government has spent two years in exhaustive research with employers and employees as well as trade unions and other key groups.

There has been a very positive response. But we know it will take time. You can't change a culture overnight. It will take at least a generation.'

Britain's notorious long-hours culture does little for anyone, least of all employees who struggle to find time for their families, or much else. Neither does it achieve much for business output: UK companies are 25% less productive than their continental counterparts despite the longer hours.

A survey conducted by the Equal Opportunities Commission found that in more than half the cases where workers registered a complaint to the commission about lack of flexibility at work had their requests rejected by their employers. Findings also reported that one in three of those refused changes were either dismissed, made redundant, or forced to resign.

The work-life balance initiative, the government says, is aimed at cultivating conditions in which employers can see the benefits of adapting often stringent and traditional working practices, to enable workers to feel comfortable about asking for alternatives, without the worry of unfavourable treatment.

The issues of maternity and paternity leave, especially in small businesses, is a subject much debated, but, says Johnson: 'What is needed is to convince employers that finding time for commitments outside work does not have a negative effect on business. In fact, it can have an extraordinarily positive effect on staff morale and productivity.'

The government's proposals for working parents of young children was a key focus of ideas originating from a report by the Work and Parent Taskforce, a working group set up last year. A significant step along the path was achieved with the announcement last November that parents of young children will have the right to ask their employer for greater flexibility; and, significantly, for those asking for change, the employer is required to offer evidence of why a request has been refused.

The work-life balance targets are ambitious, but achievable, the government says. It estimates that of the 3.8m working parents with children under six, over half a million will ask for altered working hours, and it is forecasting a take-up of 82%.

Critics of the scheme say the targets are wildly overestimated, and that small businesses in particular will find it impossible to juggle the varying demands of their

workforce. Some supporters are even sceptical about the degree to which it is enforceable, pointing to the fact that under the new guidelines, employment tribunals will still not be able to force employers to grant more flexible hours, but only ensure their refusal follows a specified code and is based on commercial reasons.

Johnson says the criticisms are to be expected, bearing in mind the magnitude of the task, but argues that such has been the success of companies voluntarily buying into work-life balance schemes, that projections of the initiative falling flat on its face are exaggerated and unnecessarily negative.

'There are many companies already running programmes, which clearly demonstrates that this initiative can work,' he says. 'There is a fundamental business case and we have seen a number of companies trailblazing the way for others to follow. Companies, big and small, are on board, and as word gets around it will be clear that this is a win-win situation.

'Some of those organisations that have adapted to the idea of flexible working hours - especially for working parents - have produced dramatic results,' he adds.

Those concerned about the potential pitfalls of the government's aims say finding a work-life balance is easier said than done, and that some sectors may not be suited to the key methods of flexible working, for example, shorter working days, compressed working weeks, job share, or flexitime. Can social workers or other health professionals, for example, where there may be staff shortages or limited skilled staff, ever expect to see the benefits of such initiatives?

Johnson acknowledges that for some areas, for some time at least, there are practical issues concerning resources which need to be addressed. But he is confident that, by setting a 'realistic' timescale, many people in key national health service roles, for example, will be able to reap the rewards as other public sector workers already are.

Management at the housing benefit department of Merton council in south London, for example, had been concerned about staff retention and morale for some time when they decided to aim for more flexible work practices.

'We had recruitment problems, high levels of sickness and high staff turnover,' says Keith Davis, the council's assistant chief executive. 'There was a general feeling

that staff morale was low. We felt we needed to do something proactive.'

Working with the public services union Unison, the council applied for and won £50,000 from the government's Work-Life Balance Challenge Fund to pilot their own scheme to help staff reorganise their working hours to better suit their individual needs.

After liaison with a consultancy group, where staff made suggestions for how things might be improved, a number of the ideas were taken up, and the results, say the council, speak for themselves: sickness levels have fallen by half, productivity is up, a work backlog has finally been dealt with, and staff that the council was worried about losing have decided to stay on.

Another issue that exponents of work-life balance are having to address is whether non-professional workers, or those whose jobs cannot be made flexible, stand to gain. For example, how can a part-time worker stacking supermarket shelves benefit?

Supermarket chain ASDA has set up a work-life balance scheme which it says has dramatically changed the way the company operates, has included a wide variety of staff at different levels, and has boosted productivity. They cite benefits such as reduced levels of absenteeism, improved staff retention, motivation and customer service.

Other examples where strategies have worked include the AA, where productivity levels of teleworkers based at home was found to be 30% higher than office-based staff. And when times are busy, says the company, they can call on an extra pool of people who don't need to be in the office to take on work and thereby increase levels of customer service.

Nationwide, one of the founding members of Employers for Work-Life Balance, first began introducing changes to working arrangements in 1995. The company has seen an increase of 14% in employee satisfaction since, according to a staff survey. Employee turnover is just 9.8%, when the financial services sector average is 24%; its maternity return rate is 91.5% - up 30% in 10 years.

The company implemented a variety of practices to foster greater flexibility, including job share, term-time, homeworking, annualised hours and a compressed

working week. It estimates a saving of around £10m in 2000 in recruitment and training costs, thanks to higher levels of staff retention.

The cooperation and involvement of business, trade unions, employment specialists and employees has been, and will continue to be critical to the success of any work-life balance scheme.

The Work-Life Balance Forum campaigns for realistic, flexible change for British workers. Set up in 1998 by its now chair, Joanna Foster, it works, she says, in partnership with a cross-section of organisations to widen the net of work-life balance initiatives. But she believes much more can be achieved. 'Government at central, regional and local level has a major enabling role to play in helping to bring about change,' she says. ' Central to the government's values are the twin aims of encouraging an economically competitive society as well as a socially cohesive and caring one. The government now needs to develop the framework of policies across departments that benefit individuals and organisations. Its role includes creating the right environment and rewarding innovative and successful ways of doing things.'

The real cost to business of days lost because of sickness caused by the stress of too much work, low morale or because of limited flexibility affecting home life, is probably incalculable. But study after study from employment experts, trade unions, government and bodies such as the Industrial Society, has clearly demonstrated that increased flexibility equals happier, healthier, better motivated and more productive staff.

Add to this the cost of recruiting new staff if valuable talent leaves because of lack of flexibility (again particularly parents or people with care commitments) and the price paid by businesses that fail to adapt are there for all to see.

Ministers say that with a concerted effort, in time British workers will have more hours outside work, a better quality of life, and be better able to help increase UK productivity.

Appendix 3.A: Directions for Summarization Task One (English Version for the Students)

Directions for the students:

Time: 2 hours for the first task + 1 hour for the second task + time for reading the directions and the questionnaire

You are to do three tasks in this session (the first task, the second task and the questionnaire), with specific directions for each one. Each task is allocated different time (2 hours for the first task, 1 hour for the second task and no time limit for the third task to fill in the questionnaire). When you finish the first two tasks, you will be asked to fill in the questionnaire regarding the tasks.

Directions for TASK ONE (Time: 2 hours)

Please read the directions carefully before reading the text. You will be given 10 minutes to read the directions by yourselves, please make sure you understand the task. After 10 minutes, the researcher will briefly explain to you the directions in Chinese, which will probably take 5 minutes. If you've got any questions concerning the directions, please don't hesitate to ask me at the classroom.

1. Read the text carefully and quickly.
2. Write a summary of it in **300-350 words** in **English**[*].
3. Please suppose that you are writing the summary for your current classmates who have not read the text themselves and you are going to give them the written summary later. **They are NOT going to read the text themselves.**
4. Your summary will be judged on its overall quality. It should be coherent, concise, and self-contained. **A summary represents the condensation of the information accessible to you and reflects the macrostructure or the gist (central ideas or essence) of the text you summarize.**
5. The scoring template will also be based on your summary, please do your best in this summarization task. It is your summary protocol that will determine what the scoring template will be like.
6. You will be given **2 hours** for both reading the text and summarizing it in the **first task,** please make careful and full use of the time allowed. Time for reading the directions is extra.

---

[*] This is in an alternating order with the second task: first English then Chinese, or first Chinese then English. The students received different directions according to their summarization conditions (see Table 4.6).

7. You are strongly urged to write your draft summary on the scrap paper provided, and then copy it neatly and legibly onto the ANSWER SHEET within the time limit for this task.
8. You can also mark the text that will be available for you throughout the test session, in whatever methods you like, for example, underlining and highlighting, to help you summarize the text.
9. Not everyone is doing exactly the same task, so please don't panic, if you see others finish their task earlier than you. It could be simply because they are doing a different task.
10. Even if you finish the **first task** before time calls, please keep seated in the classroom, as you will be asked to do the second task and fill in a questionnaire regarding the two tasks later. Therefore, **it is suggested that you shouldn't hand in your summary before time calls.**
11. At the end of the **first task**, please hand in all the scrap paper, and of course your written summary, **with your names**. Please keep the text and the directions with you.
12. Please also read the following general rules of summarization, which may help you to do the summarization task.

General rules of summarization:

1. **Deletion -- delete the trivial and redundant information in the source text**
2. **Superordination – substitute a superordinate term for a list of items, and a superordinate action for a list of subcomponents of that action**
3. **Selection – select the topic sentence that already exists, select the important information**
4. **Invention – invent the topic sentence if it does not exist**
5. **(Re)construction – integrate the important information you've selected and invented into a coherent, concise and self-contained summary that represents and reflects the condensed central ideas or essence of the source text**
6. **Polishing your summary – finish your summary product with best care, make it readable and polished, and faithful to the source text**

Now begin your first task to read and summarize the text→

Appendix 3.B: Directions for Summarization Task Two (English Version for the Students)

Directions for TASK TWO (1 hour)

Time: 1 hour

All other directions for the second task are exactly the same as those in the first task you've just finished, except No.2, No.6, No. 10, and No.11. There is no extra time for reading the directions as in the first task.

2. Write a summary of it in **300-350 words** in **Chinese***.
6. You will be given **1 hour** for both reading the text and summarizing it in the **second task**, please make careful and full use of the time allowed. No extra time for reading the directions for this task.
10. Even if you finish the **second** task before time calls, please keep seated in the classroom, as you will be asked to fill in a questionnaire regarding the two tasks later. Therefore, **it is suggested that you shouldn't hand in your summary before time calls.**
11. At the end of the **second task**, please hand in all the scrap paper, the text, the directions and of course your written summary, **with your names**.

Thank you very much indeed for your participation in the research!

For those students who are doing the tasks at the computer rooms, please DO NOT SAVE OR TURN OFF Microsoft Word, leave it as it is. The researcher is going to save it as another file. Many thanks.

* This is in an alternating order with the first task: first English then Chinese, first Chinese then English. The students received different directions according to their summarization conditions (see Table 4.6).

Appendix 3.C: Directions for Summarization Task One (Chinese Version for the Students)

中文说明：

本次集合，大家共需完成三项任务，每项任务各有不同的指令。请大家务必仔细阅读各指令。第一项任务 2 个小时，第二项任务 1 个小时，第三项任务没有时间限制（估计会需要 20-30 分钟）。请大家依次完成这三项任务，每完成一项任务，按照指令规定上交部分或全部材料。

第一项任务指令 （2 小时）

- 仔细、快速地阅读原文；
- 请用英语*撰写一篇 300-350 字对原文的缩写；
- 假想您的同学自己没有阅读原文，但想了解文章的大意，您的缩写是为她/他撰写的；
- 我们将对您的缩写作整体的评价。缩写本身应该是连贯的、浓缩的、一篇完整的文章，它浓缩原文的重要信息、反映原文的宏观结构与主旨。
- 本次缩写的评分标准来源于您的缩写，每位同学的缩写共同决定评分标准；
- 第一项任务共 2 个小时（不包括阅读本指令的时间），请大家仔细、合理地利用好全部时间；
- 我们建议您先在草稿纸上缩写，然后在规定的时间内将缩写清楚地誊写到答题纸上；
- 您可以凭自己的喜好在原文上做任何记号，如下划线等，这样可能有助于您完成本次缩写；
- 每个同学的任务不尽相同，如看到其他同学比您早完成，请勿惊慌，很有可能这无非是因为他们做的是不同的任务，阅读的是不同的文章；
- 如果您提前完成此项任务，请您仍然坐在座位上，因为接下去还有两项任务需要完成。我们建议您最好不要提前交卷。
- 第一项任务 2 小时满后，请大家上交草稿纸和缩写，别忘了写上学号和姓名，不必上交原文和本指令。
- 下面请大家阅读缩写的一些基本原则，这或许有助于您完成缩写任务：

- **删除** --- 删除原文细节及冗余信息
- **升级分类** --- 把下一级的行为或项目上升一个或几个级别。如：牛、猪、羊等可以升级分类为家畜或动物；又如：他乘公共汽车到火车站，进入票房购买了火车票，排队等在月台，3 点钟的时候终于踏上了去伦敦的列车。这样一系列的行为可以升级分类为“他坐火车去伦敦了”。
- **选择** --- 选择原文已经存在的主题句，选择重要信息
- **创造** --- 若原文不存在主题句，创造主题句
- **构建（重构）** --- 把您已经选择或创造的重要信息融合成反映原文主旨的、连贯的、浓缩的、完整的一篇缩写
- **粉饰缩写** --- 仔细修改、装点缩写，使之不仅忠于原文，而且文体优美、可读性强。

---

* This is in an alternating order with the second task: first English then Chinese, or first Chinese then English. The students received different directions according to their summarization conditions (see Table 4.6).

Appendix 3.D: Directions for the Summarization Task (for the Experts)

Directions for experts:

Please read the directions carefully before reading the three texts, and make sure you understand the tasks. If you've got any questions concerning the directions, please don't hesitate to contact me at Guoxing.Yu@bristol.ac.uk.

- ✓ Read and summarize the three texts carefully, **one by one.** Only when you've finished summarizing the first text, then you go on with the next one, though you can choose whichever as the first text.
- ✓ Write a summary of each text in **300-350** words in **English**. Your summary should be coherent, concise and self-contained. **A summary represents the condensation of the information accessible to you and reflects the macrostructure or the gist (central ideas or essence) of the text you summarize.**
- ✓ There is **no time limit** for the tasks, please spend adequate time for each text, and write down the time you spend on each text, including reading and summarizing.
- ✓ Please suppose that you are writing the summaries for your current colleagues who have not read the texts themselves and you are going to give them the written summaries later. **They are not going to read the original texts themselves.**
- ✓ You are strongly urged to write your draft summaries on scrap papers, and then copy them neatly and legibly onto A4 paper, or you can Word® process them.
- ✓ You can also mark the texts that should be available for you throughout the summarization sessions, in whatever methods you like, for example, underlining and highlighting, to help you summarize the texts.
- ✓ Please return the directions, scrap papers with your draft summaries, the three texts, and of course the written summary for each text.
- ✓ Please also read the following general rules of summarization, which may help you to do the summarization tasks.

General rules of summarization:

1. **Deletion -- delete the trivial and redundant information in the source text**

2. **Superordination – substitute a superordinate term for a list of items, and a superordinate action for a list of subcomponents of that action**

3. **Selection – select the topic sentence that already exists, select the important information**

4. **Invention – invent the topic sentence if it does not exist**

5. **(Re)construction – integrate the important information you've selected and invented into a coherent, concise and self-contained summary that represents and reflects the condensed central ideas or essence of the source text**

6. **Polishing your summary – finish your summary product with best care, make it readable and polished, and faithful to the source text**

<u>Thank you very much indeed for your help!</u>

Appendix 4.A: Post-Summarization Questionnaire (TextA Group, English Version)

Your Name _____University _____

Please read each question carefully and **tick (√) one box only** which corresponds with what you think most reflects your situation, where appropriate. There are 3 questions where you are required to provide detailed answers in Chinese or/and English. There is no right or wrong answer, and no time limit. Please also note that it is **not a test of your language abilities.** This questionnaire has **4 pages**.

1. Was the text overall easy or difficult for you to

	very difficult	somewhat difficult	moderately easy/difficult	somewhat easy	easy
(**1.a**) read to understand?					
(**1.b**) read to summarize?					

2. Were you familiar with the topic of **educational history** before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all

[if you tick (not familiar at all), please go to Question **5**, others please carry on]

3. If you were familiar with the topic of **educational history** before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(**3.a**) read to understand the text?					
(**3.b**) read to summarize the text?					

4. To which activity do you think your familiarity with the topic of **educational history** before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

5. Were you familiar with the topic of **educational history of South East Asian** countries before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all

[if you tick (not familiar at all), please go to Question **8**, others please carry on]

6. If you were familiar with the topic of **educational history of South East Asian** countries before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(**6.a**) read to understand the text?					
(**6.b**) read to summarize the text?					

7. To which activity do you think your familiarity with the topic of **educational history of South East Asian** countries before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

8. Are you familiar with the following two tests you've just done?

	very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all
(**8.a**) English summarization					
(**8.b**) Chinese summarization					

9. Did you write summaries like these in your university courses?

Yes	No

10. How much do you think your ability to write the **summary in English** depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(**10.a**) English reading abilities?					
(**10.b**) English writing abilities?					

11. On which ability do you think your **English summary** depends most?

English reading abilities	English writing abilities	equally (in)dependent on the 2 types of abilities

12. How much do you think your ability to write the **summary in Chinese** depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(**12.a**) English reading abilities?					
(**12.b**) Chinese writing abilities?					
(**12.c**) English to Chinese translation abilities?					

13. On which ability do you think your **summary in Chinese** depends most?

English reading abilities	Chinese writing abilities	English to Chinese translation abilities

14. Which language do you prefer to use to summarize the text, English or Chinese?

English	Chinese	I don't mind which language

15. Can you explain the reasons for your answer to Question 14?

16. You were asked to summarize the same text in both Chinese and English. Which task do you think can better measure your English reading abilities?

English summarization	Chinese summarization	Equally well

17. In which order did you summarize the text in both languages?

English then Chinese	Chinese then English

18. In which order would you prefer to summarize the same text in both languages?

English then Chinese	Chinese then English	I don't mind the order

19. Can you explain the reasons for your answer to Question 18?

20. Are there any other comments you would like to make regarding the text and the summarization tasks?

The questions on this page are **only for students who read the text on computer**. Please don't answer if you read the text on paper.

21. How helpful was your level of computer familiarity for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(**21.a**) read to understand the text?					
(**21.b**) read to summarize the text?					

22. To which activity do you think your computer familiarity level was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

Thank you very much for your time again.

Appendix 4.B: Post-Summarization Questionnaire (TextA Group, Chinese Version)

姓名_____ 大学 _____

请您仔细阅读本问卷的每个问题，在最符合您的情况的方格上打√（注意： 每个问题**只选一个**答案，您选择的答案没有对与错之分）。本问卷有三个问题（即 No. 15, No. 19, No. 20）要求您详细回答。本问卷没有时间限制，也不测试您的语言能力，务请按照您的实际情况选择符合您的最佳答案。

本问卷共 4 页。

1. 您觉得要 1.a/1.b 您刚才阅读的那篇文章，对您来说

	非常困难	困难	中等难易度	还算容易	容易
(1.a) 理解?					
(1.b) 缩写?					

2. 阅读那篇文章之前，您是否熟悉**教育史**这个话题?

非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉

[如果您选择了“一点不熟悉”，下面请回答第 5 个问题，选择其他的同学，请继续]

3. 如果您对**教育史**这个话题有一定程度的熟悉，请问这对于您 3.a/3.b 有多大帮助?

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(3.a) 理解文章?					
(3.b) 缩写文章?					

4. 您对**教育史**这个话题一定程度的熟悉对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

5. 阅读那篇文章之前，您是否熟悉**东南亚国家的教育史**这个话题?

非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉

[如果您选择了“一点不熟悉”，下面请回答第 8 个问题，选择其他的同学，请继续]

6. 如果您对**东南亚国家的教育史**这个话题有一定程度的熟悉，请问这对于您 6.a/6.b 有多大帮助?

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(6.a) 理解文章?					
(6.b) 缩写文章?					

7. 您对**东南亚国家的教育史**这个话题一定程度的熟悉对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

8. 您对刚才的那两种缩写考试形式是否熟悉?

	非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉
(8.a) 用英文缩写					
(8.b) 用中文缩写					

9. 请问在大学的各类课程中您是否写过类似的缩写?

写过类似的缩写	没有写过类似的缩写

10. 您的**英文缩写**成绩在多大程度上取决于(或依赖于)您的 10.a/10.b?

	依赖性很强	依赖性相当强	略有依赖性(独立性)	独立性相当强	独力性很强
(10.a) 英文阅读能力?					
(10.b) 英文写作能力?					

11. 您的**英文缩写**成绩更取决于您的下述哪种能力?

英文阅读能力	英文写作能力	这两种能力的作用相同

12. 您的**中文缩写**成绩在多大程度上取决于(或依赖于)您的 12.a/12.b/12.c?

	依赖性很强	依赖性相当强	略有依赖性(独立性)	独立性相当强	独力性很强
(12.a) 英文阅读能力?					
(12.b) 中文写作能力?					
(12.c) 英文翻译成中文的能力?					

13. 您的**中文缩写**成绩更取决于您的下述哪种能力?

英文阅读能力	中文写作能力	英文翻译成中文的能力

14. 您更愿意用哪种语言缩写那篇文章?

英文	中文	本人无特别偏好

15. 请您详细说明第 14 题中您选择的理由

16. 在本次考试中，您分别用中文和英文缩写了那篇文章。请问哪种缩写形式更能反映您的英文阅读能力?

英文缩写	中文缩写	两种缩写形式能相等地反映我的英文阅读能力

17. 在本次考试中，您分别用中文和英文缩写了那篇文章。请问您缩写的语言顺序是?

先英文缩写再中文缩写	先中文缩写再英文缩写

18. 如果在本次考试中您可以选择的话，请问您更愿意用下列哪种语言顺序缩写?

先英文缩写再中文缩写	先中文缩写再英文缩写	本人无特别偏好

19. 请您详细说明第 18 题中您选择的理由

20. 关于本次考试的文章及缩写形式的其他任何评论：

本页问题只适用于在计算机房阅读的同学。如果您是在纸上阅读的，请不要回答下列问题。

21. 您的计算机熟练程度对 21.a/21.b 有多大帮助？

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(21.a) 理解文章?					
(21.b) 缩写文章?					

22. 您的计算机熟练程度对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

非常感谢您的再次合作！

Appendix 4.C: Post-Summarization Questionnaire (TextB Group, English Version)

Your Name _____University _____

Please read each question carefully and **tick (√) one box only** which corresponds with what you think most reflects your situation, where appropriate. There are 3 questions where you are required to provide detailed answers in Chinese or/and English. There is no right or wrong answer, and no time limit. Please also note that it is **not a test of your language abilities.** This questionnaire has **4 pages.**

1. Was the text overall easy or difficult for you to

	very difficult	somewhat difficult	moderately easy/difficult	somewhat easy	easy
(1.a) read to understand?					
(1.b) read to summarize?					

2. Were you familiar with the topic of **controlled flooding** before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all

[if you tick (not familiar at all), please go to Question **5**, others please carry on]

3. If you were familiar with the topic of **controlled flooding** before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(3.a) read to understand the text?					
(3.b) read to summarize the text?					

4. To which activity do you think your familiarity with the topic of **controlled flooding** before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

5. Were you familiar with the topic of **controlled flooding at Arizona's Grand Canyon** before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all

[if you tick (not familiar at all), please go to Question **8**, others please carry on]

6. If you were familiar with the topic of **controlled flooding at Arizona's Grand Canyon** before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(6.a) read to understand the text?					
(6.b) read to summarize the text?					

7. To which activity do you think your familiarity with the topic of **controlled flooding at Arizona's Grand Canyon** before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

8. Are you familiar with the following two tests you've just done?

	very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all
(8.a) English summarization					
(8.b) Chinese summarization					

9. Did you write summaries like these in your university courses?

Yes	No

10. How much do you think your ability to write the **summary in English** depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(10.a) English reading abilities?					
(10.b) English writing abilities?					

11. On which ability do you think your **English summary** depends most?

English reading abilities	English writing abilities	equally (in)dependent on the 2 types of abilities

12. How much do you think your ability to write the **summary in Chinese** depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(12.a) English reading abilities?					
(12.b) Chinese writing abilities?					
(12.c) English to Chinese translation abilities?					

13. On which ability do you think your **summary in Chinese** depends most?

English reading abilities	Chinese writing abilities	English to Chinese translation abilities

14. Which language do you prefer to use to summarize the text, English or Chinese?

English	Chinese	I don't mind which language

15. Can you explain the reasons for your answer to Question 14?

16. You were asked to summarize the same text in both Chinese and English. Which task do you think can better measure your English reading abilities?

English summarization	Chinese summarization	Equally well

17. In which order did you summarize the text in both languages?

English then Chinese	Chinese then English

18. In which order would you prefer to summarize the same text in both languages?

English then Chinese	Chinese then English	I don't mind the order

19. Can you explain the reasons for your answer to Question 18?

20. Are there any other comments you would like to make regarding the text and the summarization tasks?

The questions on this page are **only for students who read the text on computer.** Please don't answer if you read the text on paper.

21. How helpful was your level of computer familiarity for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(21.a) read to understand the text?					
(21.b) read to summarize the text?					

22. To which activity do you think your computer familiarity level was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

Thank you very much for your time again.

Appendix 4.D: Post-Summarization Questionnaire (TextB Group, Chinese Version)

姓名_____ 大学 _____

请您仔细阅读本问卷的每个问题，在最符合您的情况的方格上打√（注意： 每个问题**只选一个**答案，您选择的答案没有对与错之分）。本问卷有三个问题（即 No. 15, No. 19, No. 20）要求您详细回答。本问卷没有时间限制，也不测试您的语言能力，务请按照您的实际情况选择符合您的最佳答案。

本问卷共 4 页。

1. 您觉得要 1.a/1.b 您刚才阅读的那篇文章，对您来说

	非常困难	困难	中等难易度	还算容易	容易
(1.a) 理解?					
(1.b) 缩写?					

2. 阅读那篇文章之前，您是否熟悉**人造洪水(controlled flooding)**这个话题?

非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉

[如果您选择了“一点不熟悉”，下面请回答第 5 个问题，选择其他的同学，请继续]

3. 如果您对**人造洪水(controlled flooding)**这个话题有一定程度的熟悉，请问这对于您 3.a/3.b 有多大帮助?

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(3.a) 理解文章?					
(3.b) 缩写文章?					

4. 您对**人造洪水(controlled flooding)**这个话题一定程度的熟悉对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

5. 阅读那篇文章之前，您是否熟悉**美国亚利桑那州大峡谷人造洪水(controlled flooding at Arizona's Grand Canyon)**这个话题?

非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉

[如果您选择了“一点不熟悉”，下面请回答第 8 个问题，选择其他的同学，请继续]

6. 如果您对**美国亚利桑那州大峡谷人造洪水(controlled flooding at Arizona's Grand Canyon)**这个话题有一定程度的熟悉，请问这对于您 6.a/6.b 有多大帮助?

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(6.a) 理解文章?					
(6.b) 缩写文章?					

7. 您对**美国亚利桑那州大峡谷人造洪水(controlled flooding at Arizona's Grand Canyon)**这个话题一定程度的熟悉对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

8. 您对刚才的那两种缩写考试形式是否熟悉?

	非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉
(8.a) 用英文缩写					
(8.b) 用中文缩写					

9. 请问在大学的各类课程中您是否写过类似的缩写?

写过类似的缩写	没有写过类似的缩写

10. 您的**英文缩写**成绩在多大程度上取决于(或依赖于)您的 10.a/10.b?

	依赖性很强	依赖性相当强	略有依赖性(独立性)	独立性相当强	独力性很强
(10.a) 英文阅读能力?					
(10.b) 英文写作能力?					

11. 您的**英文缩写**成绩更取决于您的下述哪种能力?

英文阅读能力	英文写作能力	这两种能力的作用相同

12. 您的**中文缩写**成绩在多大程度上取决于(或依赖于)您的 12.a/12.b/12.c?

	依赖性很强	依赖性相当强	略有依赖性(独立性)	独立性相当强	独力性很强
(12.a) 英文阅读能力?					
(12.b) 中文写作能力?					
(12.c) 英文翻译成中文的能力?					

13. 您的**中文缩写**成绩更取决于您的下述哪种能力?

英文阅读能力	中文写作能力	英文翻译成中文的能力

14. 您更愿意用哪种语言缩写那篇文章?

英文	中文	本人无特别偏好

15. 请您详细说明第 14 题中您选择的理由

16. 在本次考试中，您分别用中文和英文缩写了那篇文章。请问哪种缩写形式更能反映您的英文阅读能力?

英文缩写	中文缩写	两种缩写形式能相等地反映我的英文阅读能力

17. 在本次考试中，您分别用中文和英文缩写了那篇文章。请问您缩写的语言顺序是?

先英文缩写再中文缩写	先中文缩写再英文缩写

18. 如果在本次考试中您可以选择的话，请问您更愿意用下列哪种语言顺序缩写?

先英文缩写再中文缩写	先中文缩写再英文缩写	本人无特别偏好

19. 请您详细说明第 18 题中您选择的理由

20. 关于本次考试的文章及缩写形式的其他任何评论：

本页问题只适用于在计算机房阅读的同学。如果您是在纸上阅读的，请不要回答下列问题。

21. 您的计算机熟练程度对 21.a/21.b 有多大帮助？

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(21.a) 理解文章?					
(21.b) 缩写文章?					

22. 您的计算机熟练程度对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

非常感谢您的再次合作！

Appendix 4.E: Post-Summarization Questionnaire (TextC Group, English Version)

Your Name _____University _____

Please read each question carefully and **tick (√) one box only** which corresponds with what you think most reflects your situation, where appropriate. There are 3 questions where you are required to provide detailed answers in Chinese or/and English. There is no right or wrong answer, and no time limit. Please also note that it is **not a test of your language abilities.** This questionnaire has **4 pages.**

1. Was the text overall easy or difficult for you to

	very difficult	somewhat difficult	moderately easy/difficult	somewhat easy	easy
(1.a) read to understand?					
(1.b) read to summarize?					

2. Were you familiar with the topic of **work-life balance campaign** before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all

[if you tick (not familiar at all), please go to Question 5, others please carry on]

3. If you were familiar with the topic of **work-life balance campaign** before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(3.a) read to understand the text?					
(3.b) read to summarize the text?					

4. To which activity do you think your familiarity with the topic of **work-life balance campaign** before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

5. Were you familiar with the topic of **work-life balance campaign in the UK** before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all

[if you tick (not familiar at all), please go to Question 8, others please carry on]

6. If you were familiar with the topic of **work-life balance campaign in the UK** before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(6.a) read to understand the text?					
(6.b) read to summarize the text?					

7. To which activity do you think your familiarity with the topic of **work-life balance campaign in the UK** before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

8. Are you familiar with the following two tests you've just done?

	very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all
(8.a) English summarization					
(8.b) Chinese summarization					

9. Did you write summaries like these in your university courses?

Yes	No

10. How much do you think your ability to write the **summary in English** depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(10.a) English reading abilities?					
(10.b) English writing abilities?					

11. On which ability do you think your **English summary** depends most?

English reading abilities	English writing abilities	equally (in)dependent on the 2 types of abilities

12. How much do you think your ability to write the **summary in Chinese** depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(12.a) English reading abilities?					
(12.b) Chinese writing abilities?					
(12.c) English to Chinese translation abilities?					

13. On which ability do you think your **summary in Chinese** depends most?

English reading abilities	Chinese writing abilities	English to Chinese translation abilities

14. Which language do you prefer to use to summarize the text, English or Chinese?

English	Chinese	I don't mind which language

15. Can you explain the reasons for your answer to Question 14?

16. You were asked to summarize the same text in both Chinese and English. Which task do you think can better measure your English reading abilities?

English summarization	Chinese summarization	Equally well

17. In which order did you summarize the text in both languages?

English then Chinese	Chinese then English

18. In which order would you prefer to summarize the same text in both languages?

English then Chinese	Chinese then English	I don't mind the order

19. Can you explain the reasons for your answer to Question 18?

20. Are there any other comments you would like to make regarding the text and the summarization tasks?

The questions on this page are **only for students who read the text on computer.** Please don't answer if you read the text on paper.

21. How helpful was your level of computer familiarity for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(21.a) read to understand the text?					
(21.b) read to summarize the text?					

22. To which activity do you think your computer familiarity level was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities

Thank you very much for your time again.

Appendix 4.F: Post-Summarization Questionnaire (TextC Group, Chinese Version)

姓名_____ 大学 _____

请您仔细阅读本问卷的每个问题，在最符合您的情况的方格上打√（注意：每个问题**只选一个**答案，您选择的答案没有对与错之分）。本问卷有三个问题（即 No. 15, No. 19, No. 20）要求您详细回答。本问卷没有时间限制，也不测试您的语言能力，务请按照您的实际情况选择符合您的最佳答案。

本问卷共 4 页。

1. 您觉得要 1.a/1.b 您刚才阅读的那篇文章，对您来说

	非常困难	困难	中等难易度	还算容易	容易
(1.a) 理解?					
(1.b) 缩写?					

2. 阅读那篇文章之前，您是否熟悉**谋求工作与生活的平衡的运动 (work-life balance campaign)**这个话题?

非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉

[如果您选择了“一点不熟悉”，下面请回答第 5 个问题，选择其他的同学，请继续]

3. 如果您对**谋求工作与生活的平衡的运动(work-life balance campaign)**这个话题有一定程度的熟悉，请问这对于您 3.a/3.b 有多大帮助?

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(3.a) 理解文章?					
(3.b) 缩写文章?					

4. 您对**谋求工作与生活的平衡的运动(work-life balance campaign)**这个话题一定程度的熟悉对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

5. 阅读那篇文章之前，您是否熟悉**英国谋求工作与生活的平衡的运动(work-life balance campaign in the UK)**这个话题?

非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉

[如果您选择了“一点不熟悉”，下面请回答第 8 个问题，选择其他的同学，请继续]

6. 如果您对**英国谋求工作与生活的平衡的运动(work-life balance campaign in the UK)**这个话题有一定程度的熟悉，请问这对于您 6.a/6.b 有多大帮助?

	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(6.a) 理解文章?					
(6.b) 缩写文章?					

7. 您对**英国谋求工作与生活的平衡的运动(work-life balance campaign in the UK)**这个话题一定程度的熟悉对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

8. 您对刚才的那两种缩写考试形式是否熟悉?

	非常熟悉	熟悉	中等熟悉程度	不太熟悉	一点不熟悉
(8.a) 用英文缩写					
(8.b) 用中文缩写					

9. 请问在大学的各类课程中您是否写过类似的缩写?

写过类似的缩写	没有写过类似的缩写

10. 您的**英文缩写**成绩在多大程度上取决于(或依赖于)您的 10.a/10.b?

	依赖性很强	依赖性相当强	略有依赖性(独立性)	独立性相当强	独力性很强
(10.a) 英文阅读能力?					
(10.b) 英文写作能力?					

11. 您的**英文缩写**成绩更取决于您的下述哪种能力?

英文阅读能力	英文写作能力	这两种能力的作用相同

12. 您的**中文缩写**成绩在多大程度上取决于(或依赖于)您的 12.a/12.b/12.c?

	依赖性很强	依赖性相当强	略有依赖性(独立性)	独立性相当强	独力性很强
(12.a) 英文阅读能力?					
(12.b) 中文写作能力?					
(12.c) 英文翻译成中文的能力?					

13. 您的**中文缩写**成绩更取决于您的下述哪种能力?

英文阅读能力	中文写作能力	英文翻译成中文的能力

14. 您更愿意用哪种语言缩写那篇文章?

英文	中文	本人无特别偏好

15. 请您详细说明第 14 题中您选择的理由

16. 在本次考试中，您分别用中文和英文缩写了那篇文章。请问哪种缩写形式更能反映您的英文阅读能力?

英文缩写	中文缩写	两种缩写形式能相等地反映我的英文阅读能力

17. 在本次考试中，您分别用中文和英文缩写了那篇文章。请问您缩写的语言顺序是?

先英文缩写再中文缩写	先中文缩写再英文缩写

18. 如果在本次考试中您可以选择的话，请问您更愿意用下列哪种语言顺序缩写?

先英文缩写再中文缩写	先中文缩写再英文缩写	本人无特别偏好

19. 请您详细说明第 18 题中您选择的理由

20. 关于本次考试的文章及缩写形式的其他任何评论：

本页问题只适用于在计算机房阅读的同学。如果您是在纸上阅读的，请不要回答下列问题。

21. 您的计算机熟练程度对 21.a/21.b 有多大帮助？

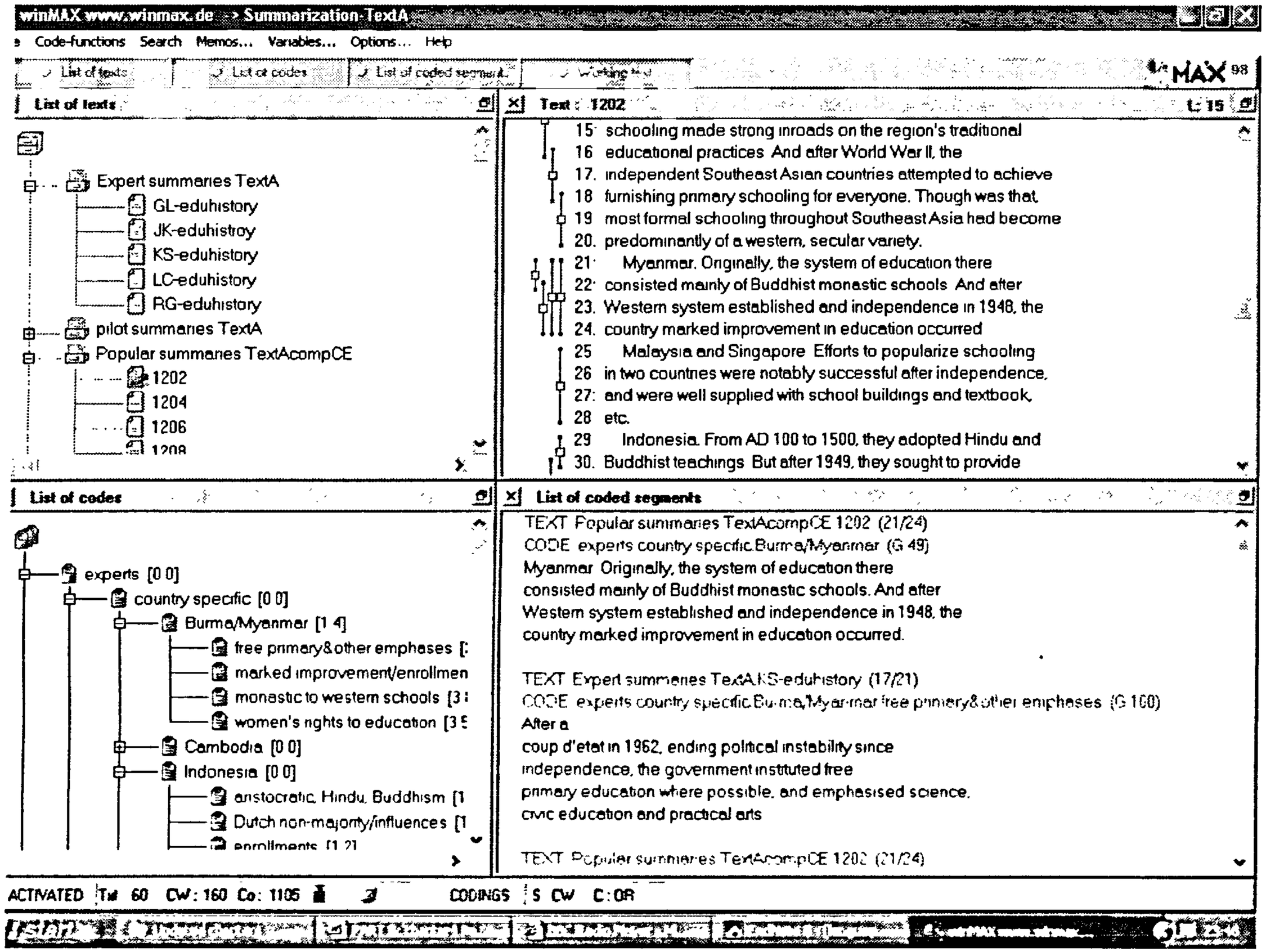
	非常有帮助	有帮助	中等帮助作用	不太有帮助	一点没帮助
(21.a) 理解文章?					
(21.b) 缩写文章?					

22. 您的计算机熟练程度对下面哪种行为更有帮助?

理解文章	缩写文章	对上述两种行为的帮助作用相同

非常感谢您的再次合作！

Appendix 5: A screenshot of winMAX programme



Appendix 6: A List of Questions for the Post-Summarization Interviews

The post-summarization interviews are based on and complementary to the post-summarization questionnaire (see Appendix 4) which will be available to the students at the interviews.

1. What are the most striking features do you think that made your text difficult or easy to understand and summarize?
2. Are you familiar with the topics of the text? To what extent do you think your (lack of) familiarity affected your understanding and summarization?
3. Did you find it difficulty to read the text on a computer? To what extent do you think your (lack of) familiarity in using computers affected your reading and summarization? (This question is only for those who read the text on computer)
4. What presentation mode would you like to choose (computer/print) if you can to read and summarize the text? What are the differences between summarizing the texts on computer and in print?
5. What were your reactions to summarize the text in Chinese/English without knowing that you were going to summarize the same text in another language (English/Chinese)? How did this affect your second summarization task in terms of your feelings and summarization process?
6. Could you please explain the reasons for your choice in the post-summarization questionnaire on "which summarization task, English or Chinese, required more of your reading abilities"? What are your comments on the English and Chinese summarization tasks?
7. Did you translate from your first summarization, from your memory, as your first summary protocol was not available for you?
8. If you translated your first summary (English/Chinese) to the other language (Chinese/English), did you find it difficult to keep all the ideas in the first summary, as we know, the two language systems are different, but you are asked to summarize the same text within the same word limit.
9. Did you summarize the same text in a similar way, regardless of the language you were using?
10. Could you please tell me in general how you summarized the text?
11. Which scoring guide (expert/popular) do you prefer to be judged against (Note: I need to explain to the interviewees how the expert and the popular scoring templates are defined and to be used in this project)?

Appendix 7.A: English Writing Task

English Writing Task

Directions:

1. You will have **60 minutes** to plan, write, and make any necessary changes to your essay. **Your essay will be graded on its overall quality.**
2. Read the topic carefully to make sure that you understand what you are asked to write about. If you do not write on the topic, your essay will not be scored.
3. Think before you write. We suggest you make notes/outlines to help you organize your essay, on scrap paper. We also suggest you write the draft(s) on the scrap paper, and then **copy your essay neatly and legibly on the answer sheet. Only essays on the answer sheet will be scored.**
4. Write clearly and precisely. How well you write is much more important than how much you write, but to cover the topic adequately, you are asked to write between **300 and 350 words in English**.
5. Plan carefully and make full use of the time allowed. You are strongly urged to allow a few minutes before time is called to read over your essay and make any changes.

Essay Topic

Some students like classes where teachers lecture (do all of the talking) in class. Other students prefer classes where the students do some of the talking. Which type of class do you prefer? Give specific reasons and details to support your choice.

Please write your final essay neatly and legibly in the ANSWER SHEET only

Appendix 7.B: Chinese Writing Task (Original Chinese Version)

中文写作题

考试说明:

1. 本题共 **45** 分钟。**45** 分钟后您必须停止答题。请合理安排时间。
2. 请大家仔细阅读考试说明和作文命题，切勿偏离主题。字数以 300-350 为宜。关键在于您写得多好，而不是您写了多少，您作文的整体质量才是评分的标准。
3. 您可以先把作文写在草稿纸上，但最后必须清楚整洁的誊写到规定的答题纸上。

作文命题

您是否同意下述观点？请据理力争您的看法。

教师的工资应该根据学生的学习掌握情况来评定。

草稿：（您必须把作文清楚整洁的誊写到答题纸上）

Appendix 7.C: Chinese Writing Task (Translated English Version)

Chinese Writing Task (English version)

Directions:

1. You will have **45 minutes** to plan, write, and make any necessary changes to your essay. **Your essay will be graded on its overall quality.** Plan carefully and make full use of the time allowed. You are strongly urged to allow a few minutes before time is called to read over your essay and make any changes.
2. Read the topic carefully to make sure that you understand what you are asked to write about. If you do not write on the topic, your essay will not be scored. Write clearly and precisely. How well you write is much more important than how much you write, but to cover the topic adequately, you are asked to write between **300 and 350 words in Chinese.**
3. Think before you write. We suggest you make notes/outlines to help you organize your essay, on scrap paper. We also suggest you write the draft(s) on the scrap paper, and then **copy your essay neatly and legibly on the answer sheet. Only essays on the answer sheet will be scored.**

Essay Topic

Do you agree or disagree with the following statement?
Teachers should be paid according to how much their students learn. Give specific reasons and examples to support your opinion.

Please write your final essay neatly and legibly in the ANSWER SHEET only

Appendix 8: Scoring Guide for the English and Chinese Writing Task

Scoring Guide for the English and Chinese Writing Tasks[1]

Two raters will use an augmentation method (e g. D^+, D, and D^-) to assign scores for each written protocol that is word-processed after the test, based on the following scoring guide In case of the score difference greater than 3[2] (e.g D and C^+, E^+ and F^-) assigned by the two raters, a third rater will use the same augmentation method and the scoring guide to judge the written protocol without knowing the previous scores by the first two raters. The average of the two most adjacent scores (difference less than 3) is reported. In case the third score is the average of the first two scores, the third score is then reported; in case the difference between any two of the three scores is still greater than 3, the three raters will negotiate face to face to assign a proper score for the questionable written protocol All raters will judge the written protocol, based on its overall quality.

Scores	Overall Comments	Task	Organization	Details/Examples	Facility in use of Language	Syntax and Vocabulary	Rater Difficulty
A	**A paper demonstrates effective communication with accuracy and clear competence in writing on both the rhetorical and syntactic levels, though it may have occasional minor errors**	effectively addresses the task	well organized and well developed logically	uses appropriate and sufficient details and examples to support effectively a clearly stated thesis or illustrate ideas	displays adequate and consistent facility in the use of the language	demonstrates syntactic variety and appropriate word choice	no difficulty in understanding is experienced by the rater
B	**A paper demonstrates good communication and competence in writing on both the rhetorical and syntactic levels, though it will probably have occasional errors**	adequately addresses almost all of the task, with some parts addressing the task more effectively than others	generally well organized and developed	uses specific examples and details to support a focused thesis or illustrate ideas	displays a general facility in the use of the language	demonstrates some syntactic variety and range of vocabulary	very little difficulty in understanding, although it may have minor flaws or occasional awkwardness that is largely free of serious errors in mechanics, grammar and usage
C	**A paper demonstrates adequate competence in writing on both the rhetorical and syntactic levels**	adequately addresses most of the task, but not always effectively	adequately organized and developed, though it may respond somewhat routinely or simplistically to the task	uses fewer specific details and examples to support a recognizable thesis or illustrate ideas	demonstrates adequate but possibly inconsistent facility in the use of the language	may contain some lexical and syntactic errors that occasionally obscure meaning convey	occasional difficulty in understanding is experienced by the rater
D	**A paper demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level or both**	inadequately addresses some of the task	inadequate organization or development	inappropriate or insufficient details and examples to support or illustrate ideas	demonstrates some but inadequate facility in the use of the language	a noticeably inappropriate choice of words/word forms, and an accumulation of errors in sentence structure and/or usage	some difficulty in understanding is experienced by the rater

[1] This scoring guide for the English Chinese writing tasks is developed after the TWE Scoring Guide of Educational Testing Service (revised 1990).

[2] The alphabetical scores (F^-, F, F^+, E^-, E, E^+, D^-, D, D^+...A^+) are transformed into numerical scores (1, 2, 3, 4, 5, 6, 7, 8, 9 ...18) correspondingly.

E	**A paper suggests incompetence in writing, and is seriously flawed by one or more of the following weaknesses**	inadequately addresses most of the task	serious or obvious disorganization or underdevelopment. Serious problems with focus, largely incoherent	little details, or irrelevant specifics	demonstrates almost no facility in the use of the language	serious and frequent errors in diction, phrasing, and sentence structure or usage	much difficulty in understanding is experienced by the rater.
F	**A paper demonstrates definitely incompetence in writing.**	fails to address the task properly	either incoherent or undeveloped, and produces only two or three incoherent sentences or a single paragraph largely paraphrasing the prompt of the writing task	no details or specifics	demonstrates no facility in the use of the language	severe and persistent writing errors in vocabulary and syntax	it is almost unintelligible.
0[3]	**A paper makes no attempts to write. It rejects the assignment, writes on another topic, or exhibits absolutely no response at all.**	---	---	---	---	---	---

[3] The score of zero is treated as system missing in subsequent data analysis of the research.

Appendix 9: The Translation Task

Translation from English to Chinese (70 minutes)

Directions:

Please read the following text (399 words) carefully and then translate it into Chinese. You shall first write your draft translation on the scrap paper provided and then copy your translation neatly and legibly onto the ANSWER SHEET.

Plan carefully and make full use of the 70 minutes. Please allow some minutes to read over your translation and make any changes before time is called.

Your translation will be marked against the overall quality on the following criteria:

- ✓ faithfulness to the original text
- ✓ correctness of your language
- ✓ elegance in your expression

英语翻译成中文 (70 分钟)

考试说明:

请仔细阅读下文(399 字)，然后把它翻译成中文，并将您的翻译从草稿纸上清楚地誊写到答题纸上。

我们将基于以下三个评分标准对您的翻译作整体的评价：

- ✓ 对原文的忠实程度
- ✓ 语言表达的准确度
- ✓ 语言表达的优美度

The rights and wrongs of treating anorexia (厌食症)

The case of Samantha Kendall, the anorexia nervosa (神经性厌食症) sufferer who discharged herself from hospital despite doctors' fears for her life, has highlighted the confusion in public thinking about this disturbing and perplexing disease. Ten years ago anorexia was still dismissed as nothing more than slimming gone too far. Today it is recognised as a treatable medical condition; but the degree to which treatment should be carried out without the patient's consent has become a topic of debate.

Researchers have suggested two psychiatric (精神病学的) explanations behind the onset of anorexia. One is that the patient, faced with an unacceptably stressful or difficult adult life, is trying to retreat into childhood or avoid leaving it. Another is that choosing what to eat — and specifically choosing not to eat — is often an attempt to exert control by people who feel that their lives are too constrained in other ways. But the truth is that for all the resources that have been devoted to its study, the syndrome (综合征) remains imperfectly understood.

It is beyond doubt, however, that anorexia is a severe psychiatric disorder. There is no other way to describe an illness that allows a patient to look in the mirror at her own emaciated (消瘦的, 憔悴的), starved body, and see someone obese staring back. Severe sufferers often deny that they are trying to kill themselves, but the diet they are pursuing is all too likely to make death inevitable.

The 1983 Mental Health Act provides for sufferers from severe psychiatric disorders to be held in hospital for treatment against their will if there is a danger that they will do harm to themselves or others. Yet even though one in 10 anorexia sufferers dies, doctors are sometimes reluctant to use their powers under the law. This is often because of a fear that treatment by compulsion is self-defeating, since force-fed victims of anorexia often return to starvation diets when they get home.

There is clearly work to be done in making the treatment of extreme anorexia -- which often involves leaving patients in isolation and without their clothes, and watching them as they eat and go to the lavatory -- more humane. But the shortcomings of the available treatments should not obscure the fact that the alternative to treatments can sometimes be death. If doctors made more use of the powers available to them, lives could be saved.

Appendix 10.A: Scoring Guide for the Translation (English to Chinese) Task

Scoring Guide for the Translation (English to Chinese) Task

Two raters will use an augmentation method (e.g. D^+, D, and D^-) to assign scores for each translation that is word-processed after the test, based on the following scoring guide. In case of the score differences greater than 3[1] (e.g. D and C^+, E^+ and F^-) assigned by the two raters, a third rater will use the same augmentation method and the scoring guide to judge the translation without knowing the previous scores by the first two raters. The average of the two most adjacent scores (difference less than 3) is reported. In case the third score is the average of the first two scores, the third score is then reported; in case the difference between any two of the three scores is still greater than 3, the three raters will negotiate face to face to assign a proper score for the questionable translation. All raters will judge the translation, based on its overall quality.

Scores	Overall Comments	Lexical Meaning	Style, Tone, and Nuances	Chinese Vocabulary and Syntax	Rater Understanding
A	**It demonstrates faithful and elegant translation of the original passage, with few minor mistranslations or omissions/additions**	faithfully reflects all the original passage with less than a couple of minor lexical errors or omissions/additions	adequately reflects the style, tone and nuances of the original passage	appropriately and elegantly uses Chinese vocabulary and syntax	No difficulty in understanding is experienced by the rater
B	**It demonstrates very good translation, although it will probably have occasional mistranslations or omissions/additions**	faithfully reflects almost all of the original passage, with some minor mistranslations or omissions/additions in comprehending individual words, phrases, sentences or ideas	generally reflects the style, tone and nuances of the original passage	demonstrates appropriate, though may not elegant, use of Chinese vocabulary and syntax	Very little difficulty in understanding is experienced by the rater
C	**It demonstrates good translation ability, with few significant mistranslations or omissions/additions**	generally reflects most of the original passage, with few significant mistranslations or omissions/additions in comprehending individual words, phrases, sentences or ideas, but not serious enough to alter the	keeps most of the style, tone and nuances of the original passage	demonstrates a few errors in choosing appropriate Chinese vocabulary and syntax, though the meaning is still conveyed	occasional difficulty in understanding is experienced by the rater

[1] The alphabetical scores (F^-, F, F^+, E^-, E, E^+, D^-, D, D^+…A^+) are transformed into numerical scores (1, 2, 3, 4, 5, 6, 7, 8, 9 …18) correspondingly.

		theme of the original passage			
D	**It demonstrates some developing competence in translating, but it remains flawed on either accuracy or expression or both**	basically reflects about half of the original passage, with some significant mistranslations, omissions or inappropriate additions that may change the theme of the original passage	The style, tone, and nuances are generally still reflected in the translation of the approximately half of the original passage	a noticeably inappropriate choice of Chinese vocabulary and syntax, though most of the meaning is still conveyed	The translation distorts about a half of the meaning of the original text, though the rater may not experience much difficulty in understanding it.
E	**It suggests incompetence in translating, and is seriously flawed by one or more of the following weaknesses**	basically reflects less than half of the original passage, with some serious mistranslations, omissions, or inappropriate additions that will change the theme of the original passage	The style, tone, and nuances are almost lost in the translation of the less than half of the original passage	contains severe and persistent errors in understanding the original passage, and in choosing appropriate Chinese vocabulary and syntax	The translation substantially distorts the meaning of the original text, though it may still be true that the rater doesn't experience much difficulty in understanding it.
F	**It demonstrates definitely incompetence in translating, and is seriously flawed by one or more of the following weaknesses**	produces only two or three Chinese sentences, no matter whether they are faithful reflection or mistranslation of the original passage, or inappropriate additions (but not absolutely irrelevant sentences). It substantively omits parts of the original passage	----	----	The translation is almost unintelligible. Or, the translation is not relevant to the original text.
0	**A paper in this category makes no attempts to translate. It rejects the assignment, exhibits absolutely no response at all, or write absolutely irrelevant sentences (English/Chinese)**	—	—	----	----

Appendix 10.B: A Chinese Translation of the Passage (Anorexia) by the Researcher

中文翻译

Samantha Kendall 是一名神经性厌食症患者，她不顾医生对她生命的担忧依然出院。这一事件昭示了民众对这既恼人又复杂的疾病理解上的混乱。十年前，人们认为厌食症无非是消瘦的过分而已。现在，厌食症被认为是一种可治疗的疾病；但是，在没有病人许可的情况下对病人进行治疗的程度却成了人们争议的焦点。

科研人员提出了造成厌食症的两种精神病学解释。第一种解释认为病人面对成人生活的压力或困难，感到难以接受，因而企图退避到童年生活，或企图避免跨出童年生活。另一种解释认为，选择吃什么---具体地说就是选择不吃东西--- 是病人实施自我控制的一种企图，这些病人通常认为自己的生活在其他许多方面受到太多的限制。但是，真正的事实是，尽管对这一疾病投入了大量的研究，我们对这病症却远未完全理解。

但是，毫无疑问，厌食症是一种严重的精神失调。我们可以这样形象地描述这一疾病：病人在镜中看到的不是自己日益消瘦的，而是肥胖的躯体。严重的患者通常否认自己是在残杀自己，但是他们的饮食却极有可能使死亡不可避免。

1983 心理健康法规定，严重的精神失常患者，如果有对本人或他人造成伤害的危险，即使本人反对，也必须接受医院治疗。然而，尽管厌食症患者有高达10%的死亡率，有时，医生还是不愿意依法实行强制治疗。这通常是由于担心强迫治疗也往往事与愿违，因为这些被强迫进食的病人（牺牲品）出院回到家后通常又恢复到先前的绝食性饮食。

目前对极度精神性厌食症患者的治疗通常是把他们隔离起来，脱光衣服，并对进餐和如厕进行监视，很显然，我们要采取措施，使治疗更人道些。但是，现有治疗方法的不足之处不应掩盖这样一个事实，即其他治疗方法有时意味着死亡。如果医生能够更多地使用自己的权力，生命应该能够拯救。

Appendix 11: Student Consent Form

Dear students,

I'm writing to explain briefly my PhD study and seek your consent to participate in the project on your summarization performances in different reading contexts. You will sit in several sessions, which are not necessarily tests (detailed directions for each session will be provided), for about 9 hours altogether within a month or so. Your participation is essential for the research project and is also an integral part of your courses/units as agreed with your class teachers in your university. You will be credited for the participation towards your grades in those courses/units.

- All of you will spend about 15 minutes filling in a 2-page questionnaire in Chinese on your computer familiarity, and signing the consent form at the end of this letter;
- All of you will be randomly assigned to a 3-hour session of the summarization task of one English text either on computer or in print (half of you will do the task at computer labs), followed by a post-summarization questionnaire in Chinese;
- All of you will sit in for a standardized reading test (55 minutes) and a task to write a short English essay on a given topic (60 minutes);
- All of you will then translate a short English article of 399 words into Chinese (70 minutes), and write a short Chinese essay of 300-350 words on a given topic (45 minutes);
- All of you will be given a practice paper to familiarize yourselves with the test formats in the subsequent session. You are asked to finish the practice paper on your own, at your spare time;
- In a week or so, all of you will have a second standardized reading test (75 minutes) that has the same test methods as in the practice paper you did a week ago;
- Some of you will be interviewed individually by the researcher (30 minutes or so);
- All of you will get feedback on your performances in the tests.

I hope you will enjoy and learn from the tests. If you have any questions, comments or complaints on the research project, please don't hesitate to contact me. Many thanks again for your help in advance. Please sign the consent form below and return it to me now.

Yu Guoxing (Email: Guoxing.Yu@bristol.ac.uk)

Graduate School of Education, University of Bristol

----------Please detach and return the form below and keep the letter for your reference----------

I have read the letter about the research project, and am willing to do my best in the tests and fill in the two Chinese questionnaires honestly.

University_____ Department _____ Student No._____

Name_____ Signature_____ Date_____

Appendix 12: Guidelines for evaluating the quality of students' summaries (Part one)

12.A: Right statements of textA (expert template, English)

	Right statements	Key words for quick reference
1	Indigenous culture, colonialism and the post - World War II era of political independence influenced the forms of education in the nations of Southeast Asia.	indigenous culture, colonialism, post-war independence
2	Before AD 1500, education throughout the region consisted chiefly of the transmission of cultural values through family and community living, supplemented by some formal teaching of each locality's dominant religion.	transmission, culture, family and community living, religion
3	With the advent of western colonization after 1500, western schooling began to make strong inroads/influences on the region's traditional educational practices.	western colonisation, inroads/influences
4	After World War II, each newly formed/independent nation in Southeast Asia attempted to achieve planned development, and schooling in all of these countries was organized in three main levels: primary, secondary, and higher education.	planned development, three levels (primary, secondary, higher)
5	But problems still existed such as school dropout, grade-repeater rates, providing sufficient school buildings and teachers to meet the expanding number of children/population.	problems, expanding children/population
6	The indigenous system of education in Burma/Myanmar consisted mainly of Buddhist monastic schools, and the western system after British occupation recognised women's right to education.	Burma/Myanmar, Buddhist/monastic schools, western/British, women's right
7	Under British rule, Malaysia and Singapore had four language streams; after independence, Malaysia government chose Malay as the teaching/educational medium, while Singapore promoted English as the main language	Malaysia, Singapore, four language streams, Malay, English, main language/medium
8	After years of civil wars, Cambodian reconstruction/rebuilding of education began in the mid 1980s.	Cambodia, civil wars, reconstruction/ Rebuilding
9	Vietnam education was strongly influenced by long Chinese domination – Confucian and Taoism, and then by French.	Vietnam, Chinese/Confucian Taoism, French
10	After independence, education system in Indonesia was predominantly western secular for a 90% Muslim population.	western secular, Muslim population

Note: Highlighted areas indicate the shared statements between the expert and the popular templates. When marking the summaries, the raters were not told which scoring criteria they were using, nor were they told which statements were the same between the two templates. They were given only a list of the 10 statements and the corresponding key words.

12.B: Right statements of textA (popular template, English)

	Right statements	Key words for quick reference
1	Indigenous culture, colonialism and the post - World War II era of political independence influenced the forms of education in the nations of Southeast Asia.	indigenous culture, colonialism, post-war independence
2	Before AD 1500, education throughout the region consisted chiefly of the transmission of cultural values through family and community living, supplemented by some formal teaching of each locality's dominant religion.	transmission, culture, family and community living, religion
3	With the advent of western colonization after 1500, western schooling began to make strong inroads/influences on the region's traditional educational practices.	western colonisation, inroads/influences
4	After World War II, each newly formed/independent nation in Southeast Asia attempted to achieve planned development, and schooling in all of these countries was organized in three main levels: primary, secondary, and higher education.	planned development, three levels (primary, secondary, higher)
5	In each nation a central ministry of education set schooling structures and curriculum requirements, to make the country more self-reliant in education.	central ministry of education, self-reliant
6	But problems still existed such as school dropout, grade-repeater rates, providing sufficient school buildings and teachers to meet the expanding number of children/population.	problems, expanding children/population
7	The indigenous system of education in Burma/Myanmar consisted mainly of Buddhist monastic schools, and the western system was established after British occupation. Following its independence, marked improvements in education occurred, and school enrolments increased dramatically.	Burma/Myanmar, Buddhist/monastic schools, western/British, marked improvement/ enrolment
8	Malaysia and Singapore had the common (British) educational roots (or diverged when Singapore left the coalition), and their efforts to popularise schooling were notably successful.	Malaysia, Singapore, common/British roots, notably successful
9	The Indonesia government sought to provide universal elementary schooling and a large measure of secondary and higher education for its socio-economic modernisation.	Indonesia, elementary, secondary, higher, modernisation
10	The new Republic of Philippines launched a series of national plans aimed at renovation and expansion of education to promote socio-economic modernisation.	Philippines, national plans, renovation and expansion, modernisation

12.C: Right statements of textC (expert template, English)

	Right statements	Key words for quick reference
1	British people work the longest hours each week, four hours longer than their European counterparts.	long/excessive hours
2	Despite the longer hours, UK companies are 25% less productive than their continental counterparts.	less productive
3	In response to this, the government launched the Work-life Balance Campaign in March 2000, to promote flexible working practices (or to change the excessive hours work culture).	launch, campaign
4	The Work-Life Balance Challenge Fund of £10m over three years has been set up to assist and advice employers to make working practices more flexible.	challenge fund, assist/advice, flexible practices
5	An alliance of business leaders (Employers for Work-Life Balance) has also been set up to promote flexible working practices.	alliance, business leaders
6	Many companies who are already running flexible working practices have produced dramatic results in improving staff morale and productivity.	dramatic results, improved staff morale and productivity
7	Many research studies also demonstrate that increased flexibility equals/leads to happier, healthier, better-motivated and more productive staff.	research studies, happier, healthier, better-motivated, more productive
8	However, critics claim that some sectors or small businesses may not be suitable for flexible working practices.	critics/criticism, some sectors, small business, not suitable
9	Working parents with young children will now have the right to ask their employer for greater flexibility.	young parents, rights
10	If such a request is rejected, the employer is required to offer evidence of why a request has been rejected.	reject, evidence/reasons

12.D Right statements of textC (popular template, English)

	Right statements	Key words for quick reference
1	British people work the longest hours each week (four hours longer than their European counterparts), and the problem of excessive hours pervades all types of business and all levels within them.	long/excessive hours, pervading problem
2	In response to this, the government launched the Work-life Balance Campaign in March 2000, to promote flexible working practices (or to change the excessive hours work culture).	launch, campaign
3	The government says this campaign (or work-life balance initiative) is a chance for the UK to abandon an excessive hours work culture, and an attempt to follow the lead of other countries that have successfully reduced the time of working.	chance to abandon, attempt to follow
4	The work-life balance initiative (or campaign) is aimed at cultivating conditions in which the employers can see the benefits of adapting stringent and often traditional working practices, and to enable employees to feel comfortable about asking for alternatives.	cultivate, conditions
5	The Work-Life Balance Challenge Fund of £10m over three years has also been set up to assist and advice employers to make working practices more flexible.	challenge fund, assist/advice, flexible practices
6	Many companies who are already running flexible working practices have produced dramatic results in improving staff morale and productivity.	already running, dramatic results, improved staff morale and productivity
7	Many research studies also demonstrate that increased flexibility equals/leads to happier, healthier, better-motivated and more productive staff.	research studies, happier, healthier, better-motivated, more productive
8	However, critics claim that some sectors (small businesses or non-professionals) may not be suitable for flexible working practices.	critics/criticism, some sectors, small business, not suitable
9	Culture cannot be changed overnight; it will take at least a generation.	culture change, overnight/generation
10	With a concerted effort in time British workers will have more time outside work, a better quality of life, and be better able to help increase UK productivity.	concerted effort, more time, better quality, better able, productivity

Appendix 12.E: Right statements of textA (expert template, Chinese)

1. 本土文化，殖民主义，二战后的政治独立都影响着东南亚国家的教育形式.
2. 公元1500年前，该地区的教育主要是通过家庭和社区生活的文化传递，并辅之以当地的主要宗教的教育.
3. 1500 年被西方国家殖民后，该地区的传统的教育方式受到西方国家教育模式的强烈冲击.
4. 第二次世界大战后，新兴独立的东南亚国家都通过规划有计划地发展教育，学校教育分为初等，中等，和高等教育三类.
5. 但是问题仍然存在，诸如学生撮学，留级生的比例，为日益增长的人口提供足够的学校设施和师资.
6. 缅甸国的本土教育主要有佛教寺院学校组成，其西方教学模式在被英国占领后建立起来，并认可妇女受教育的权利.
7. 在英国统治下，马来西亚和新加坡实行四种语言教学；独立后，马来西亚政府选定马来语作为教学语言，而新加坡则提倡英语作为主要语言.
8. 多年内战后，柬埔寨的教育重建始于八十年代中期.
9. 越南的教育模式首先是受到中国儒家和道教的多年影响，然后又受到法国的影响.
10. 独立后的印度尼西亚，该国占总人口 90%的穆斯林所受的是西方式的世俗教育.

Keywords for quick reference

1. 本土文化，殖民主义，战后政治独立
2. 家庭和社区生活，文化传递，宗教教育
3. 西方殖民，冲击/影响
4. 规划发展，初等，中等，和高等教育三类
5. 问题，人口增长
6. 缅甸，佛教寺院学校，西方/英国，妇女教育权利
7. 马来西亚，新加坡，四种语言，马来语，英语，教学/主要语言
8. 内战，柬埔寨，教育重建
9. 越南，中国/儒家/道教，法国
10. 西方世俗教育，穆斯林

Appendix 12.F: Right statements of textA (popular template, Chinese)

1. 本土文化，殖民主义，二战后的政治独立都影响着东南亚国家的教育形式.
2. 公元 1500 年前，该地区的教育主要是通过家庭和社区生活的文化传递，并辅之以当地的主要宗教的教育.
3. 1500 年被西方国家殖民后，该地区的传统的教育方式受到西方国家教育模式的强烈冲击.
4. 第二次世界大战后，新兴独立的东南亚国家都通过规划有计划地发展教育，学校教育分为初等，中等，和高等教育三类.
5. 为使教育上更加独立自主，每个东南亚国家都成立了中央教育委员会，负责学校的设置和教学大纲.
6. 但是问题仍然存在，诸如学生撤学，留级生的比例，为日益增长的人口提供足够的学校设施和师资.
7. 缅甸国的本土教育主要有佛教寺院学校组成，其西方教学模式在被英国占领后建立起来. 主权独立后，教育得到了长足的发展，学生入学率大幅度增长.
8. 马来西亚和新加坡同受英国殖民统治，在教育上同源同根(新加坡在脱离新马联盟后在教育上与马来西亚有所不同)，两国的教育发展都取得了令人瞩目的成功.
9. 印度尼西亚政府希求通过全民初等教育和大范围的中等，高等教育，以实现经济和社会现代化.
10. 新成立的菲律宾政府实施了一系列的国家教育创新和拓展计划，以促进经济和社会现代化.

Keywords for quick reference

1. 本土文化，殖民主义，战后政治独立
2. 家庭和社区生活，文化传递，宗教教育
3. 西方殖民，冲击/影响
4. 规划发展，初等，中等，和高等教育三类
5. 中央教育委员会，独立自主
6. 问题，人口增长
7. 缅甸，佛教寺院学校，西方/英国，长足的发展
8. 马来西亚，新加坡，同源同根，令人瞩目的成功
9. 印度尼西亚，初等教育，中等，高等教育，现代化
10. 菲律宾，计划，创新和拓展，现代化

Appendix 12.G: Right statements of textC (expert template, Chinese)

1. 英国工人工作时间最长，每周比其它欧洲国家的工人多四个小时。
2. 尽管（工人）工作时间长，英国公司却比欧洲大陆的公司生产效率低 25%
3. 为解决这一问题，在 2000 年三月，（英国）政府发起了工作与生活平衡的运动，旨在提倡灵活多样的工作方式。
4. (政府）设立的工作与生活平衡挑战基金高达 1000 万英镑，用于在未来三年内资助和建议雇方使工作方式更加灵活多样。
5. 大型单位联合创立的“工作与生活平衡雇方”亦旨在提倡和促进工作方式更加灵活多样。
6. 许多实行灵活工作方式的公司已经收到了很好的效果，员工的士气和生产效率得到了提高。
7. 许多研究也表明增加（工作方式的）灵活性会促进员工更加健康，幸福，提高他们的工作动力和生产效率。
8. 但是，有批评人士认为，灵活工作方式不适合某些工种和小型单位。
9. 现在有小孩的父母有权向他们的雇方要求更为灵活的工作方式。
10. 如果要拒绝这些父母的要求，雇方必须出示证据，说明为何要拒绝的理由。

Keywords for quick reference

1. 工作时间长
2. 生产效率低
3. 发起，运动
4. 挑战基金，帮助/建议，灵活多样的工作方式
5. 联盟，大型单位
6. 很好的效果，提高，员工士气和生产效率
7. 研究，更加幸福，健康，提高工作动力，生产效率
8. 批评，某些单位/工种，小单位，不适合
9. 年轻父母，权利
10. 拒绝，证据/理由

Appendix 12.II: Right statements of textC (popular template, Chinese)

1. 英国工人工作时间最长（每周比其它欧洲国家的工人多四个小时），工作时间过长的问题遍及各行各业的各个层次。
2. 为解决这一问题，在 2000 年三月，（英国）政府发起了工作与生活平衡的运动，旨在提倡灵活多样的工作方式（或改变过长工作时间的文化）。
3. （英国）政府指出，工作与生活平衡的运动是英国摈弃过长工作时间的文化的一次机会，也是效仿其它成功减少工作时间的国家的一次尝试。
4. 工作与生活平衡的运动旨在培育工作环境，使雇主看到调整传统的、刻板的工作方式的好处，同时使雇员在要求调整工作方式时不受拘束。
5. （政府）设立的工作与生活平衡挑战基金高达 1000 万英镑，用于在未来三年内资助和建议雇方使工作方式更加灵活多样。
6. 许多实行灵活工作方式的公司已经收到了很好的效果，员工的士气和生产效率得到了提高。
7. 许多研究也表明增加（工作方式的）灵活性会促进员工更加健康，幸福，提高他们的工作动力和生产效率。
8. 但是，有批评人士认为，灵活工作方式不适合某些工种和小型单位。
9. 文化的改变不可能一夜之间就能完成，它需要至少一代人（的努力）。
10. 大家齐心协力，英国工人最终一定能够有更多的工作之外的时间，更好的生活质量，并能更好的帮助英国提高生产效率。

Keywords for quick reference:

1. 工作时间长，问题（多，遍及）
2. 发起，运动
3. 机会，摈弃，尝试，学习（效仿）
4. 培育，条件/环境
5. 挑战基金，帮助/建议，灵活多样的工作方式
6. 很好的效果，提高，员工士气和生产效率
7. 研究，更加幸福，健康，提高工作动力，生产效率
8. 批评，某些单位/工种，小单位，不适合
9. 文化的改变，一夜/一代人
10. 齐心协力，更多的工作之外的时间，更好的生活质量，提高生产效率

Appendix 13.A: Guidelines for evaluating the quality of students' summaries (Part two, for textA and textC summaries)

Summarization Tasks Scoring Guide Part II

Two raters will use an augmentation method (e g. D^+, D, and D^-) to assign scores for each summary that is word-processed after the test, based on the following scoring guide. In case of the score differences greater than 3[1] (e.g. D and C^+, E^+ and F^-) assigned by the two raters, a third rater will use the same augmentation method and the scoring guide to judge the summary without knowing the previous scores by the first two raters. The average of the two most adjacent scores (difference less than 3) is reported. In case the third score is the average of the first two scores, the third score is then reported; in case the difference between any two of the three scores is still greater than 3, the three raters will negotiate face to face to assign a proper score for the questionable summary. All raters will judge the summary, based on its overall quality.

Scores	Overall Comments	Faithfulness to the Source Text (RSC + WSP)	SSS: summary and source text relationships scores[2]	Conciseness and Coherence (5% score) (+, 0, -)	Rater Understanding
A	**It demonstrates faithful and elegant summarization of the source text, with few minor misrepresentations or omissions/ additions of the source text**	• faithfully reflects all the important statements from the source text with very few minor lexical errors or omissions/ additions in the important statements[3]. • No significant wrong statements.	• predominantly in the summarizer's own words and sentence structures, in addition to the accurate use of the language from the source text. • also rearranges the order of the statements logically, have examples of integration and connectives, and have global interpretation of the source text.	• expresses the important statements concisely and in a coherent and logic manner; the summary *per se* is also self-contained, and logically organized to faithfully reflect the meaning of the source text. • The percentage of non-important statements is quite low.	• No difficulty in understanding is experienced by the rater
B	**It demonstrates very good summarization, although it will**	• faithfully reflects almost all the important statements from the source text, but with	• mostly in the the summarizer's own words and sentence structures, in	• almost all the important statements are delivered in an adequate organization –	• Very little difficulty in understanding is experienced by

[1] The alphabetical scores (F^-, F, F^+, E^-, E, E^+, D^-, D, D^+...A^+) are transformed into numerical scores (1, 2, 3, 4, 5, 6, 7, 8, 9 ...18) correspondingly.
[2] It is assumed that the summarizers use their own words to produce a Chinese summary of an English source text, therefore no SSS is assigned for Chinese summaries.
[3] See Scoring Guide Part I for the 10 most important right statements (either in expert or popular scoring templates for the 3 texts).

	probably have occasional misrepresentations or omissions/additions of the source text	some minor misrepresentations or omissions/additions in the important statements. • Few noticeable wrong statements, that may not be part of the important statements, could obscure occasionally the meaning of the source text.	addition to appropriate use of the language from the source text. • also re-orders (though still linear in their presentation) of the statements, attempts to integrate or use connectives.	coherent and logical, though may not be concisely presented. • There are a few instances of non-important statements of the source text.	the rater
C	**It demonstrates good summarization ability, with few significant misrepresentations or omissions/additions**	• generally reflects most of the important statements of the source text, • with some significant misrepresentations or omissions/additions in the important statements, but not serious enough to alter the general theme of the source text	• some use of the summarizer's own words and sentence structures, in addition to the adequate use of the language from the source text. • basically follows the order of source text with few cases of re-ordering and integration.	• the important statements are delivered in an acceptable manner – coherent and logical, though few important statements are mislocated or not concisely presented. • There are several instances of non-important statements of the source text.	• occasional difficulty in understanding is experienced by the rater
D	**It demonstrates some developing competence in summarization, but it remains flawed on either accuracy or expression or both**	• basically reflects about half of the important statements, with some significant misrepresentations, omissions or inappropriate additions that may change the theme of the source text	• predominantly verbatim copied, • follows the original order of the statements in the source text, shows rare instance of proper integration and connectives, and not global in their interpretation of the source text	• the summarizer shows rare concerns and attempts to logically present the important statements, and fails to address this properly because of either its verbatim copying or wrong links among the statements • There are almost equal portions of	• Some difficulty in understanding is experienced by the rater. • Or, the summary omits about a half of the meaning of the source text, though the rater may not experience much difficulty in understanding what's in the

				important and non-important statements.	summary.
E	**It suggests incompetence in summarization, and is seriously flawed by one or more of the following weaknesses**	• basically reflects less than half of the source text, with some serious misrepresentations, omissions, or inappropriate additions that will change the theme of the source text	• in addition to the weaknesses of predominantly verbatim copying the source text, • it is also flawed that the integrations or connectives are not logically presented.	• it demonstrates no concern of conciseness, coherence or logicality of the meaning conveyed, with reference to the conceptual map of the model summary. • There are higher portions of non-important statements than important ones.	• Much difficulty in understanding is experienced by the rater. • Or, the summary substantially distorts the meaning of the source text, though it may still be true that the rater doesn't experience much difficulty in understanding the summary.
F	**It demonstrates incompetence in summarization, and is seriously flawed by one or more of the following weaknesses**	• produces only two or three sentences, no matter whether they are faithful verbatim copying or misrepresentation of the source text, or inappropriate additions (but not absolutely irrelevant sentences) It substantively omits the important statements of the source text	• it lifts locally a couple of sentences from the source text, and fails to deliver the main important statements • No instances of integration or connectives at all.	• The few sentences are a random selection or verbatim copying from the source text, there is definitely no concern of conciseness, coherence or logicality of the meaning conveyed, with reference to the conceptual map of the model summary. • The few sentences produced are either important or non-important statements of the source text	• The summary is almost unintelligible • Or, the summary doesn't convey the main theme of the source text.

Appendix 13.B: Guidelines for evaluating the quality of students' summaries (Part two, for textB summaries)

Summarization Tasks Scoring Guide Part II for TextB (Let the River Run)

Two raters will use an augmentation method (e.g. D^+, D, and D^-) to assign a single quality score for each summary that is word-processed after the test, based on the following scoring guide. In case of the score differences greater than 3[1] (e.g. D and C^+, E^+ and F^-) assigned by the two raters, a third rater will use the same augmentation method and the scoring guide to judge the summary without knowing the previous scores by the first two raters. The average of the two most adjacent scores (difference less than 3) is reported. In case the third score is the average of the first two scores, the third score is then reported; in case the difference between any two of the three scores is still greater than 3, the three raters will negotiate face to face to assign a proper score for the questionable summary. All raters will judge the summary, based on its overall quality.

Scores	Overall Comments	Faithfulness to the Source Text[2]	summary and source text relationships scores[3]	Conciseness and Coherence	Rater Understanding
A	**It demonstrates faithful and elegant summarization of the source text, with few minor misrepresentations or omissions/ additions of the source text**	◆ faithfully reflects all the important statements from the source text with few minor lexical errors or omissions/ additions in the important statements. ◆ No significant wrong statements.	◆ predominantly in the summarizer's own words and sentence structures, in addition to the accurate use of the language from the source text. ◆ also rearranges the order of the statements logically, have examples of integration and connectives, and have global interpretation of the source text.	◆ expresses the important statements concisely and in a coherent and logic manner; the summary *per se* is also self-contained, and logically organized to faithfully reflect the meaning of the source text. ◆ The percentage of non-important statements is quite low.	◆ No difficulty in understanding is experienced by the rater
B	**It demonstrates very good**	◆ faithfully reflects almost all the	◆ mostly in the the summarizer's	◆ almost all the important statements	◆ Very little difficulty in understanding is

[1] The alphabetical scores (F^-, F, F^+, E^-, E, E^+, D^-, D, D^+…A^+) are transformed into numerical scores (1, 2, 3, 4, 5, 6, 7, 8, 9 …18) correspondingly.
[2] See Scoring Guide I of TextB for the sample summary.
[3] It is assumed that the summarizers use their own words to produce a Chinese summary of an English source text, therefore, in rating the Chinese summaries, this criteria is not applied.

	summarization, although it will probably have occasional misrepresentations or omissions/additions of the source text	important statements from the source text, with some minor misrepresentations or omissions/additions in the important statements. ◆ Few noticeable wrong statements, that may not be part of the important statements, could obscure the meaning of the source text.	own words and sentence structures, in addition to appropriate use of the language from the source text. ◆ also re-orders (though still linear in their presentation) of the statements, attempts to integrate or use connectives.	are delivered in an adequate organization – coherent and logical, though may not be concisely presented. ◆ There are a few instances of non-important statements of the source text.	experienced by the rater
C	**It demonstrates good summarization ability, with few significant misrepresentations or omissions/additions**	◆ generally reflects most of the important statements of the source text, ◆ with few significant misrepresentations or omissions/additions in the important statements, but not serious enough to alter the theme of the source text	◆ some use of the summarizer's own words and sentence structures, in addition to the adequate use of the language from the source text. ◆ basically follows the order of the source text with few cases of re-ordering and integration.	◆ the important statements are delivered in an acceptable manner – coherent and logical, though few important statements are mislocated or not concisely presented. ◆ There are several instances of non-important statements of the source text.	◆ occasional difficulty in understanding is experienced by the rater
D	**It demonstrates some developing competence in summarization, but it remains flawed on either accuracy or expression or both**	◆ basically reflects about half of the important statements, with some significant misrepresentations, omissions or inappropriate additions that may change the theme of the	◆ predominantly verbatim copied, ◆ follows the original order of the statements in the source text, shows rare instance of proper integration and connectives, and	◆ the summarizer shows rare concerns and attempts to logically present the important statements, and fails to address this properly because of either its verbatim copying or wrong links	◆ Some difficulty in understanding is experienced by the rater. ◆ Or, the summary omits about a half of the meaning of the source text, though the rater may not experience much

		source text	not global in their interpretation of the source text	among the statements • There are almost equal portions of important and non-important statements.	difficulty in understanding what's in the summary.
E	**It suggests incompetence in summarization, and is seriously flawed by one or more of the following weaknesses**	• basically reflects less than half of the source text, with some serious misrepresentations, omissions, or inappropriate additions that will change the theme of the source text	• in addition to the weaknesses of predominantly verbatim copying the source text, • it is also flawed that the integrations or connectives are not logically presented.	• it demonstrates no concern of conciseness, coherence or logicality of the meaning conveyed, with reference to the conceptual map of the model summary. • There are higher portions of non-important statements than important ones.	• Much difficulty in understanding is experienced by the rater. • Or, the summary substantially distorts the meaning of the source text, though it may still be true that the rater doesn't experience much difficulty in understanding the summary.
F	**It demonstrates definitely incompetence in summarization, and is seriously flawed by one or more of the following weaknesses**	• produces only two or three sentences, no matter whether they are faithful verbatim copying or misrepresentation of the source text, or inappropriate additions (but not absolutely irrelevant sentences). It substantively omits the important statements of the source text.	• it lifts locally a couple of sentences from the source text, and fails to deliver the main important statements. • No instances of integration or connectives at all.	• The few sentences are a random selection or verbatim copying from the source text, there is definitely no concern of conciseness, coherence or logicality of the meaning conveyed, with reference to the conceptual map of the model summary. • The few sentences produced are either important or non-important statements of the source text.	• The summary is almost unintelligible. • Or, the summary doesn't convey the main theme of the source text.

Appendix 13.C: A sample summary of *Let the River Run* (textB) for the raters (English version)

In1996, Arizona's Grand Canyon was deliberately flooded to undo years of environmental damages caused by Glen Canyon Dam built in 1963 to produce hydroelectricity. By the late 1970s, however, serious ecological and geological effects of the Dam were becoming apparent.

The Dam trapped over 90% of the sediment, and changed the river's environment. Its largest fish (pikeminnow) had disappeared; another native fish (humpback chub) was endangered. However, alien carnivorous trout had spread throughout the entire river, as had a non-native river tree (tamarisk). Because of the lack of the continuous supply of sand, the beaches downstream were eroded to half of their original size. The Dam also ended the river's natural floods to clear of debris that choked the rapids.

A task force recommended opening the floodgates for a short time to reverse the environmental damages. In1996, the floodgates were opened. In the weeks after the flood waters cleared, the canyon seemed pristinely restored, its beaches back and rapids clear of debris. However, while many of the flood's positive effects have lasted, not everything went as planned. The sandbars were not rebuilt from the whole riverbed. The flood had no significant effect on the fish population. It had also actually exacerbated the tamarisk invasion by spreading its seeds along the canyon,

Lessons from this are being included to guide 2002 flood. It would need last only two days in winter, with a very low output for several months beforehand to build riverbed sediments. Winter flooding has a second advantage: it won't spread tamarisks when they are not in seed. However, this ecological issue is complicated: the tamarisk may compete with native plants but is a superb nesting habitat for the endangered willow flycatcher. There are also fears of undoing everything the winter flood builds if there would be spring floods as suggested by some scientist to hit the trout during their spawning period.

Anyhow, the power companies have agreed to 2002 flood despite its causing them greater losses than in 1996. If successful, flooding may be called for elsewhere, but the only certain way to undo the damages is to decommission the dam.

349 words (excl. four words in brackets)

Appendix 13.D: A sample summary of *Let the River Run* (textB) for the raters (Chinese version)

为消除革兰大坝 1963 年建成后多年来对环境的破坏，1996 年，亚力桑那州大峡谷经历了一场人为的洪水。革兰大坝建成时旨在水力发电，但是到了七十年代末，大坝对生态和地质的影响也日益明显。

大坝拦截了超过 90%的泥沙流量，因而改变了河流环境。流域中最大的鱼种(pikeminnow)已经绝迹，另一种土生鱼种(humpback chub)亦受到威胁。但是，外来的食肉类鱼种（如 trout）和植物（如 tamarisk)却遍布整个流域。下游的河滩因为没有得到泥沙的经常性补给也萎缩到了只有原来的一半大小。大坝的建成也结束了河流原先的季节性洪水，导致湍滩碎石堆积 。

政府派出的工作组建议通过短期泻洪来扭转大坝对环境的破坏。1996 年闸门终于打开。洪水退后的几周，大峡谷似乎恢复到了原始状态，河滩回复，湍滩碎石不再堆积。尽管这次人为洪水的一些正面效应得以持续，但不是每件事都如计划的一样。抬高沙洲的泥沙并不是来自河床；洪水对鱼类也没有产生很大影响；洪水助播了 tamarisk 的种子，因此反而加剧了它的入侵。

1996 年的经验教训将用于指导 2002 年的泻洪。2002 年泻洪需安排在冬天且只需持续两天即可，在泻洪前数月大坝流量应控制在非常低的水平，以帮助河床泥沙的堆积。冬季泻洪的另一个好处在于，它可以防止助播 tamarisk 的种子，因为这时 tamarisk 没有结籽。但是这一生态问题却是复杂的：tamarisk 可能会影响土生植物的生长，但它却是频危的食虫鸟的极佳的栖息之地。有科学家建议在冬季泻洪之后，春季继续泻洪以打击此时处于产卵期的 trout，但是也有科学家担忧这样一来会使冬季的泻洪成果前功尽弃。

不管怎样，能源公司已经同意 2002 年的泻洪计划。如果革兰大坝的泻洪能够扭转环境破坏的话，类似的泻洪也可以在其它地区实施。但是，唯一能确保成功的方法只有一个，炸毁大坝。

Appendix 13.E: Scoring sheet

Rater: One/ Two/Three (please delete as appropriate)

SumNo	Statements Details													5%	SSS	HS
	Right statements and RSC score										RSC	WSP	Subtotal			
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						
	1	2	3	4	5	6	7	8	9	10						

Appendix 14: Descriptive statistics of the data from the computer familiarity questionnaire

Question ID	1 (less familiar)		→ 2 →		→ 3 →		→ 4 (more familiar)		Mean	Std. deviation
	Freq.	%	Freq.	%	Freq.	%	Freq.	%		
1	97	64.2	14	9.3	13	8.6	27	17.9	1.80	1.189
2	1	.6	9	5.8	43	27.6	103	66.0	3.59	.631
3	37	24.0	68	44.2	31	20.1	18	11.7	2.19	.936
4	112	71.8	9	5.8	24	15.4	11	7.1	1.58	.991
5	145	98.0	2	1.4	1	.7	-	-	1.03	.200
6	5	3.2	76	48.7	65	41.7	10	6.4	2.51	.667
7	1	.6	23	14.6	88	56.1	45	28.7	3.13	.667
8	13	8.3	83	52.9	54	34.4	7	4.5	2.35	.697
9	2	1.3	49	31.2	94	59.9	12	7.6	2.74	.611
10	3	1.9	39	24.8	79	50.3	36	22.9	2.94	.745
11	5	3.2	139	88.5	11	7.0	2	1.3	2.06	.387
12	18	11.5	120	76.4	19	12.1	-	-	2.01	.487
13	3	1.9	4	2.6	5	3.2	144	92.3	3.86	.538
14	14	8.9	26	16.6	69	43.9	48	30.6	2.96	.912
15	-	-	7	4.5	40	25.6	109	69.9	3.65	.564
16	-	-	8	5.1	45	28.8	103	66.0	3.61	.586
17	8	5.1	41	26.3	63	40.4	44	28.2	2.92	.865
18	9	5.8	32	20.5	60	38.5	55	35.3	3.03	.890
19	-	-	7	4.5	37	23.7	112	71.8	3.67	.558
20	88	56.4	34	21.8	18	11.5	16	10.3	1.76	1.018
21	5	3.2	44	28.2	62	39.7	45	28.8	2.94	.837
22	10	6.4	59	37.8	52	33.3	35	22.4	2.72	.886
23	78	50.3	72	46.5	4	2.6	1	.6	1.54	.584
24	78	50.3	55	35.5	15	9.7	7	4.5	1.68	.828
25	18	11.5	49	31.4	51	32.7	38	24.4	2.70	.967
26	-	-	24	15.4	116	74.4	16	10.3	2.95	.505
27	35	22.4	80	51.3	32	20.5	9	5.8	2.10	.809
28	60	38.5	79	50.6	16	10.3	1	.6	1.73	.666
29	36	23.1	65	41.7	44	28.2	11	7.1	2.19	.873
30	36	23.2	93	60.0	24	15.5	2	1.3	1.95	.662
31	1	.6	23	14.7	103	66.0	29	18.6	3.03	.601
32	No training (39), 25%				Training (117), 75%					

Appendix 15: Factor analysing the data from the computer familiarity questionnaire: some statistics

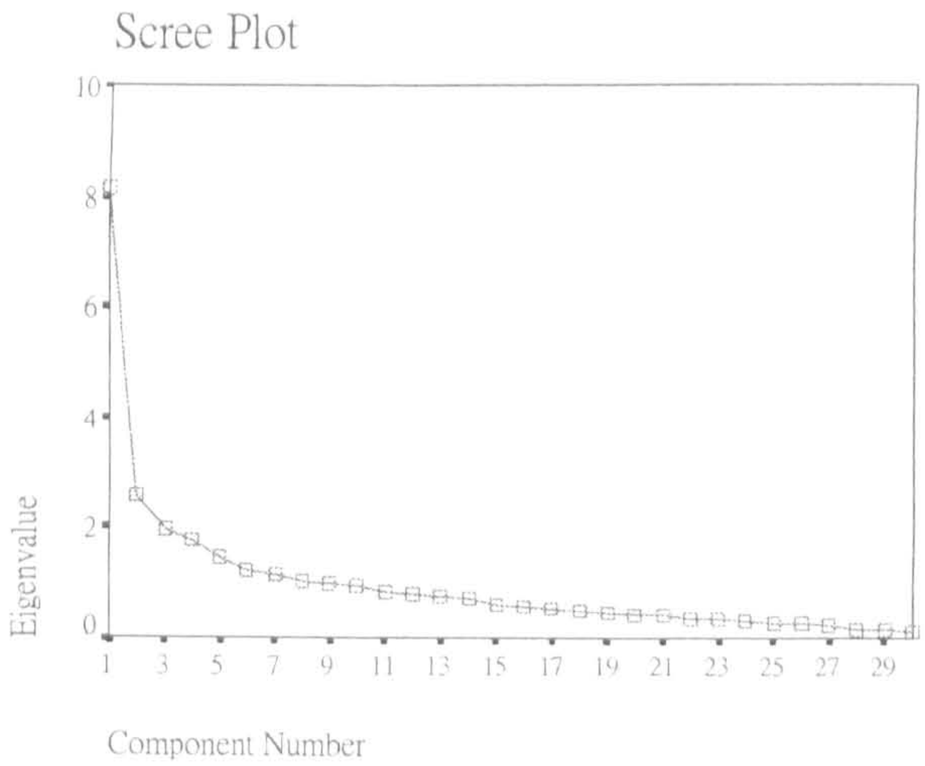


Figure: Scree plot

Component	Initial Eigenvalues		
	Total	% of variance	Cumulative %
1	8.126	27.087	27.087
2	2.542	8.473	35.560
3	1.929	6.430	41.990
4	1.769	5.895	47.885
5	1.460	4.866	52.751
6	1.240	4.133	56.884
7	1.160	3.866	60.750
8	1.031	3.436	64.185
9	1.009	3.362	67.547

Table: Initial eigenvalues greater than one

Factor	Verbal Descriptor
1	Uses of public-accessed computers other than home: frequency of using computer equipments and software for various purposes
2	Uses of self-possessed computers at home: familiarity with computer equipments and software for word-processing and computer games, and the self-initiated problem-solving activities

Table: Verbal descriptors of the two factors of promax oblique rotations

Rotated Component Matrix[a]			Pattern Matrix[a]		
	Component			Component	
	1	2		1	2
CFQ1 Availability home	-.201	.654	Availability home	-.368	.757
CFQ2 Availability uni. comp. labs	.731	-8.019E-03	Availability uni. comp. labs	.795	-.200
CFQ3 Availability others	.383	.316	Availability others	.343	.241
CFQ4 ownership home	-.298	.722	ownership home	-.489	.856
CFQ6 familiarity computer	.292	.540	familiarity computer	.193	.506
CFQ7 familiarity mouse	.146	.495	familiarity mouse	4.494E-02	.495
CFQ8 familiarity Word English	.443	.463	familiarity Word English	.374	.384
CFQ9 familiarity Word Chinese	.333	.558	familiarity Word Chinese	.233	.515
CFQ10 familiarity screen reading	.230	.451	familiarity screen reading	.146	.426
CFQ11 no. CBT tests	3.708E-02	.132	no. CBT tests	9.917E-03	.133
CFQ12 general self-evaluation	.305	.501	general self-evaluation	.216	.460
CFQ13 frequency mobile	5.516E-02	8.794E-02	frequency mobile	3.965E-02	8.034E-02
CFQ14 frequency ATM	.204	.265	frequency ATM	.161	.232
CFQ15 frequency computer	.704	.316	frequency computer	.691	.156
CFQ16 frequency Internet	.672	.392	frequency Internet	.639	.247
CFQ17 frequency VCD/DVD	.461	.352	frequency VCD/DVD	.419	.259
CFQ18 frequency Emails	.658	.399	frequency Emails	.623	.258
CFQ19 frequency mouse	.642	.351	frequency mouse	.616	.210
CFQ20 frequency computer games	.264	.522	frequency computer games	.167	.494
CFQ21 frequency Word Chinese	.734	.108	frequency Word Chinese	.771	-7.511E-02
CFQ22 frequency Word English	.744	7.428E-02	frequency Word English	.790	-.114
CFQ23 frequency spreadsheets	.549	-5.634E-02	frequency spreadsheets	.608	-.204
CFQ24 frequency graphics	.582	.188	frequency graphics	.588	5.038E-02
CFQ25 frequency chat room	.411	.283	frequency chat room	.381	.198
CFQ26 possibility self sort out	7.174E-02	.401	possibility self sort out	-1.419E-02	.413
CFQ27 possibility help button	.201	.630	possibility help button	7.328E-02	.626
CFQ28 possibility manual/magazine	.117	.625	possibility manual/magazine	-1.679E-02	.642
CFQ29 possibility Internet search help	.187	.536	possibility Internet search help	8.012E-02	.529
CFQ31 possibility give up	.167	.304	possibility give up	.111	.283
CFQ32 training last two years	-.206	-.259	training last two years	-.165	-.225

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
[a] Rotation converged in 3 iterations.

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.
[a] Rotation converged in 3 iterations.

Table: Comparisons of the loadings in varimax and promax rotations.

Reliability analyses of computer familiarity scale

Item-total Statistics

Item	scale mean if item deleted	scale variance if item deleted	corrected item-total correlations	squared multiple correlations	Alpha if item deleted
CFQ1	37.9730	42.1761	.2515	.6304	.8539
CFQ2	36.1959	43.1518	.4618	.4904	.8342
CFQ3	37.5811	41.3063	.4286	.2324	.8362
CFQ4	38.1824	43.8236	.2017	.6718	.8519
CFQ8	37.4392	42.3704	.5083	.3883	.8315
CFQ15	36.1284	41.8950	.7058	.7779	.8248
CFQ16	36.1757	41.5880	.7197	.7570	.8236
CFQ17	36.8716	41.1059	.4996	.3150	.8312
CFQ18	36.7703	39.2258	.6607	.5403	.8207
CFQ19	36.1081	42.3556	.6470	.6366	.8274
CFQ21	36.8514	40.4267	.5933	.5601	.8256
CFQ22	37.0676	40.1995	.5764	.6288	.8263
CFQ23	38.2365	44.4131	.3425	.3642	.8395
CFQ24	38.1149	41.8711	.4785	.4125	.8326
CFQ25	37.0878	40.9922	.4441	.3033	.8353

Alpha = .8425 Standardized item alpha = .8658

Appendix 16: Raters performances in marking the summaries

A: Rating performances of three groups of raters for all summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
mean	9 1724	8 8621	11 3548	11 4839	8 8788	8 0909	9 1875	9 1875	10 54	10.26	11 8333	11 433	9 9091	8 6364	10 6154	9 7308
s.d	2 5783	2 3562	2.8466	2 8269	3 5157	3 6603	1 7859	1 7678	2 0821	2 2389	2 0356	3 0021	2 8433	2 7241	1 6752	1 9091
Corr.	0 8330		0 8934		0 9285		0 7241		0 7004		0 8812		0 8588		0.7668	
Alpha	0.9069		0 9437		0 9625		0 84		0 8225		0 9003		0 9236		0 8638	
Std. alpha	0 9089		0 9437		0 9630		0 84		0 8238		0 9368		0 9240		0 8680	
No.	29		31		33		32		50		30		55		26	

Table: Rating performance of Group1 raters (i.e. Raters 1&2) for all texts summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
mean	8 7442	9 000	11	10 2571	8 7037	8 4444	9 3235	9 3235	9 8958	10 3542	11 5429	10 9714	9 10	9 825	9 5	10
s.d	2 2466	2 2509	2 1004	1 7547	3 1722	2 4390	2 8361	2 7603	2 1326	2 41	2 3432	1 6357	2 2395	2 4061	2 6585	2 4495
Corr.	0 7778		0 7741		0 8379		0 7952		0 7967		0 7869		0 7742		0 7886	
Alpha	0 8750		0 8647		0 8949		0 8857		0 8833		0 8497		0 8715		0 8801	
Std. alpha	0 8750		0 8727		0 9118		0 8859		0 8868		0 8807		0 8727		0 8818	
No.	31		35		27		34		48		35		40		38	

Table: Rating performance of Group2 raters (i.e. Raters 2&3) for all texts' summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1
mean	10 175	10 2	10 0294	10 5	8 625	8.9250	9 3824	10 1471	10 8136	10 9153	10 6	11 6	9 9839	9 6129	10 4722	10 9167
s.d	2 1109	2 6814	2.2894	2 1213	2 0342	2 3685	2 2964	3 1635	2 0127	2 2689	1 769	1 7534	2 4926	2 2350	2.1178	2 5565
Corr.	0 8770		0 8891		0 7923		0 8930		0 7554		0 7798		0 7257		0 8413	
Alpha	0 9204		0 9399		0 8784		0 9184		0 8571		0 8762		0 8382		0 9051	
Std. alpha	0 9345		0 9413		0 8841		0 9435		0 8606		0 8763		0 8411		0 9138	
No.	40		34		40		34		59		35		62		36	

Table: Rating performance of Group3 raters (i.e. Raters3&1) for all texts' summaries

B: Rating performances of three groups of raters for textA summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
mean	9 7647	9 4118	11 875	12 125	10 1111	9 8889	9 9444	9 6667	11 20	9 733	12 25	12 125	10 8421	9 3684	11 1538	10 0
s.d	2 6582	2.2655	3 1385	2 8018	2 8674	2 6983	1 6968	1 9097	1 8974	2 4631	1 8439	2 7049	2 3866	2 6291	1 9513	2 3805
Corr.	0 8681		0 9268		0 9140		0 7564		0 8529		0 8621		0 8686		0 8611	
Alpha	0 9231		0 9588		0 9541		0 8579		0 9039		0 8904		0 9274		0 9156	
Std. alpha	0 9294		0 9620		0 9551		0 8613		0 9206		0 9260		0 9297		0 9254	
No.	17		16		18		18		15		16		19		13	

Table: Rating performance of Group 1 raters (i.e. Raters1&2) for text A summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
Mean	9 7895	9 1579	11 1111	10 2778	9 6471	8 4706	10 3889	9 7778	9 5789	9 4737	11 3889	10 6667	9 3750	9 3750	9 9524	10 3333
s.d	1 9883	2 4555	2 0548	1 9646	2 2897	2 2945	2 0903	2 2375	2 0633	2 4351	1 8830	1 3284	2 3910	2 6552	1 8296	1 6533
Corr.	0 9062		0 8225		0 8663		0 7490		0 8601		0 6898		0 8585		0 7163	
Alpha	0 9398		0 9021		0 9284		0 8554		0 9180		0 7878		0 9211		0 8322	
Std. alpha	0 9508		0 9026		0 9284		0 8565		0 9248		0 8164		0 9239		0 8347	
No.	19		18		17		18		19		18		16		21	

Table: Rating performance of Group 2 raters (i.e. Raters2&3) for text A summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1
Mean	11 0588	11 7059	10 6316	10 8947	9 2222	9 7778	10 1765	11 5294	11 2105	12 2105	11 0526	12 2632	10 0556	10 7222	11	12 0526
s.d	2 3041	2 5682	2 4315	2 3545	1 8960	1 9268	2 4808	2 6953	1 8732	1 9316	1 8401	1 7902	1 9545	1 5265	2	1 9571
Corr.	0 8903		0 9244		0 8033		0 9479		0 8316		0 8220		0 6758		0 7948	
Alpha	0 9391		0 9605		0 8909		0 9715		0 9078		0 9021		0 7920		0 8856	
Std. alpha	0 9420		0 9607		0 8909		0 9733		0 9080		0 9023		0 8006		0 8857	
No.	17		19		18		17		19		19		18		19	

Table: Rating performance of Group 3 raters (Raters3&1) for textA summaries

C: Rating performance of three groups of raters for textB summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
mean	No RSC for textB (see also research design)								10 5909	10 8182	No ephs		10 2273	9 0909	No cphs	
s.d									1 9678	2 0386			2 4089	1 9739		
Corr.									0 6928				0 7966			
Alpha									0 8182				0 8771			
Std. alpha									0.8185				0 8868			
No.									22				22			

Table: Rating performance of Group1 raters (Raters1&2) for textB summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
mean	No RSC for textB (see also research design)								10 9375	11 8750	No ephs		8 8824	10 1176	No cphs	
s.d									1 8062	2 2174			1 9648	2 1179		
Corr.									0 7303				0 6945			
Alpha									0 8340				0 8183			
Std. alpha									0 8441				0 8197			
No.									16				17			

Table: Rating performance of Group2 raters (Raters2&3) for textB summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1
mean	No RSC for textB (see also research design)								11 2105	10 6316	No ephs		11 9444	10 6111	No cphs	
s.d									2.3233	2 1912			1 9844	1 3779		
Corr.									0 7364				0 7446			
Alpha									0 8473				0 8219			
Std. alpha									0 8482				0 8563			
No.									19				18			

Table: Rating performance of Group3 raters (Raters3&1) for textB summaries

D: Rating performances of three groups of raters for textC summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
mean	8 3333	8 0833	10 8	10 8	7 4000	5 9333	8 2143	8 5714	9 6923	9 9231	11 3571	10 6429	8 1429	6 9286	10 0769	9 4615
s.d	2 3094	2 3533	2 4842	2 7826	3 7378	3 5550	1 4239	1 3986	2 3232	2 2532	2 2051	3 2251	3 3936	3 2691	1 1875	1 3301
Corr.	0 7472		0 8514		0 9375		0 5518		0 7433		0 8846		0 8539		0 5032	
Alpha	0 8552		0 9166		0 9671		0 7111		0 8525		0 9037		0 9208		0 6667	
Std. alpha	0 8553		0 9198		0 9677		0 7112		0 8528		0 9388		0 9212		0 6695	
No.	12		15		15		14		13		14		14		13	

Table: Rating performance of Group 1 raters (Raters1&2) for textC summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3
mean	7 1667	8 75	10 8824	10 2353	7 1000	8 4000	8 1250	8 8125	9 0769	9 7692	11 7059	11 2941	9 0000	10 1429	8 9412	9 5882
s.d	1 6422	1 9598	2 2046	1 5624	3 9001	2 7968	3 1385	3 2500	2 2532	1 7394	2 8010	1 8962	2 7689	2 6726	3 3998	3 1831
Corr.	0 8615		0 7343		0 9636		0 8325		0 7278		0 8410		0 8558		0 8062	
Alpha	0 9179		0 8186		0 9544		0 9083		0 8264		0 8769		0 9220		0 8916	
Std. alpha	0 9256		0 8468		0 9815		0 9086		0 8425		0 9136		0 9223		0 8927	
No.	12		17		10		16		13		17		7		17	

Table: Rating performance of Group 2 raters (Raters2&3) for textC summaries

rating criteria	RSC								HS							
	ee		ep		ce		cp		ee		ep		ce		cp	
rater	3	1	3	1	3	1	3	1	3	1	3	1	3	1	3	1
mean	9 5217	9 0870	9 2667	10 0000	8 1364	8 2273	8 5882	8 7647	10 0952	10 0000	10 0625	10 8125	8 5769	8 1538	9 8824	9 6471
s.d	1 7286	2 2139	1 9074	1 7321	2 0539	2 5058	1 8391	3 0522	1 7001	2 1679	1 5692	1 3276	2 2481	2 3442	2 1472	2 5967
Corr.	0 8309		0 8000		0 7616		0 8390		0 7732		0 6460		0 6656		0 8889	
Alpha	0 8926		0 8866		0 8551		0 8517		0 8578		0 7783		0 7988		0 9322	
Std. alpha	0 9076		0 8889		0 8647		0 9124		0 8721		0.7849		0 7992		0 9412	
No.	23		15		22		17		21		16		26		17	

Table: Rating performance of Group 3 raters (Raters3&1) for textC summaries

Appendix 17: Frequency of 5% scores

ESumExp5%	TextA		TextC		Total (textsA&C)	
	Frequency	Percent (%)	Frequency	Percent (%)	Frequency	Percent (%)
-1.0	13	24.5	2	4.3	15	15
-.5	17	32.1	6	12.8	23	23
0	6	11.3	14	29.8	20	20
.5	10	18.9	9	19.1	19	19
1.0	7	13.2	16	34.0	23	23
Total	53	100	47	100	100	100

Table: Frequency of ESumExp5% score

ESumPop5%	TextA		TextC		Total (textsA&C)	
	Frequency	Percent (%)	Frequency	Percent (%)	Frequency	Percent (%)
-1.0	12	22.6	2	4.3	14	14
-.5	7	13.2	6	12.8	13	13
0	12	22.6	14	29.8	26	26
.5	14	26.4	9	19.1	23	23
1.0	8	15.1	16	34.0	24	24
Total	53	100	47	100	100	100

Table: Frequency of ESumPop5% score

CSumExp5%	TextA		TextC		Total (textsA&C)	
	Frequency	Percent (%)	Frequency	Percent (%)	Frequency	Percent (%)
-1.0	4	7.5	2	4.3	6	6
-.5	10	18.9	11	23.4	21	21
0	14	26.4	11	23.4	25	25
.5	19	35.8	14	29.8	33	33
1.0	6	11.3	9	19.1	15	15
Total	53	100	47	100	100	100

Table: Frequency of CSumExp5% score

CSumPop5%	TextA		TextC		Total (textsA&C)	
	Frequency	Percent (%)	Frequency	Percent (%)	Frequency	Percent (%)
-1.0	3	5.7	2	4.3	5	5
-.5	7	13.2	10	21.3	17	17
0	13	24.5	10	21.3	23	23
.5	18	34.0	14	29.8	32	32
1.0	12	22.6	11	23.4	23	23
Total	53	100	47	100	100	100

Table: Frequency of CSumPop5% score

Appendix 18: RSC scores before and after adjustments

Rating criteria	RSC			
	EE	EP	CE	CP
Mean	9.445	10.740	8.625	9.43
Standard deviation	2.3321	2.2769	2.7683	2.3688
Min. – Max.	5-16.5	6-17	2.5-15.5	2-15.5
Kolmogorov-Smirnov Z^*	1.006	0.720	0.96	1.057
Sig.	0.2645	0.6785	0.3165	0.2135
No. summaries	100			

Note: * Kolmogorov-Smirnov test (2-tailed) of normal distribution.

Table: RSC before adjustments of WSP

Rating criteria	RSC			
	EE	EP	CE	CP
Mean	9.235	10.53	8.435	9.240
Standard deviation	2.4021	2.3114	2.8592	2.4541
Min. – Max.	4.0-16.5	4.5-17.0	1.5-15.5	2.0-15.5
Kolmogorov-Smirnov Z^*	1.090	0.648	0.809	1.138
Sig.<	0.1865	0.7955	0.5295	0.1505
No. summaries	100			

Note: * Kolmogorov-Smirnov test (2-tailed) of normal distribution.

Table: RSC after adjustments of WSP

Rating criteria	RSC			
	EE	EP	CE	CP
Mean	46.27	52.94	42.605	46.85
Standard deviation	11.9611	11.2925	14.82	12.5621
Min. – Max.	21-80	23.5-81.5	7.5-79.5	9.0-79.5
Kolmogorov-Smirnov Z^*	1.055	0.404	0.654	0.755
Sig.<	0.2165	0.9975	0.7865	0.6185
No. cases	100			

Note: * Kolmogorov-Smirnov test (2-tailed) of normal distribution.

Table: RSC after adjustments of both WSP and 5% (in percentage)

Appendix 19: Correlations between the key variables of this research

Correlations

		TOEFL reading	FCE reading all parts	Chinese writing	English writing	translation	ESumExp statement score final PC after 5% and WSP	ESumExp HS/OQS	ESumPop statement score final PC after 5% and WSP	ESumPop HS/OQS	CSumExp statement final PC after 5% and WSP	CSumExp HS/OQS	CSumPop statement score final PC after 5% and WSP	CSumPop HS/OQS
TOEFL reading	Pearson Correlation	1	.400**	.110	.130	.471**	.253*	.130	.294**	.164	.257**	.237**	.233*	.216*
	Sig (2-tailed)		.000	.177	.107	.000	.011	.106	.003	.103	.010	.003	.020	.031
	N	156	153	152	156	153	100	156	100	100	100	156	100	100
FCE reading all parts	Pearson Correlation	.400**	1	.128	.164*	.309**	.022	.091	.048	-.009	.221*	.203*	.181	.113
	Sig (2-tailed)	.000		.116	.043	.000	.827	.260	.639	.932	.029	.012	.074	.269
	N	153	154	151	153	152	98	154	98	98	98	154	98	98
Chinese writing	Pearson Correlation	.110	.128	1	.227**	.209**	.180	.156	.153	.188	.140	.152	.067	-.024
	Sig (2-tailed)	.177	.116		.005	.009	.076	.054	.131	.064	.170	.060	.510	.812
	N	152	151	153	152	153	98	153	98	98	98	153	98	98
English writing	Pearson Correlation	.130	.164*	.227**	1	.073	.108	-.056	.140	.123	.107	.061	.106	.043
	Sig (2-tailed)	.107	.043	.005		.367	.283	.488	.164	.224	.290	.448	.294	.673
	N	156	153	152	156	153	100	156	100	100	100	156	100	100
translation	Pearson Correlation	.471**	.309**	.209**	.073	1	.179	.092	.111	.094	.102	.120	.106	.087
	Sig (2-tailed)	.000	.000	.009	.367		.076	.259	.272	.355	.317	.138	.296	.390
	N	153	152	153	153	154	99	154	99	99	99	154	99	99
ESumExp statement score final PC after 5% and WSP	Pearson Correlation	.253*	.022	.180	.108	.179	1	.880**	.465**	.345**	.483**	.458**	.384**	.285**
	Sig (2-tailed)	.011	.827	.076	.283	.076		.000	.000	.000	.000	.000	.000	.004
	N	100	98	98	100	99	100	100	100	100	100	100	100	100
ESumExp HS/OQS	Pearson Correlation	.130	.091	.156	-.056	.092	.880**	1	.400**	.369**	.446**	.434**	.316**	.243*
	Sig (2-tailed)	.106	.260	.054	.488	.259	.000		.000	.000	.000	.000	.001	.015
	N	156	154	153	156	154	100	157	100	100	100	157	100	100
ESumPop statement score final PC after 5% and WSP	Pearson Correlation	.294**	.048	.153	.140	.111	.465**	.400**	1	.892**	.215*	.197*	.408**	.299**
	Sig (2-tailed)	.003	.639	.131	.164	.272	.000	.000		.000	.032	.050	.000	.003
	N	100	98	98	100	99	100	100	100	100	100	100	100	100
ESumPop HS/OQS	Pearson Correlation	.164	-.009	.188	.123	.094	.345**	.369**	.892**	1	.144	.156	.337**	.282**
	Sig (2-tailed)	.103	.932	.064	.224	.355	.000	.000	.000		.153	.120	.001	.004
	N	100	98	98	100	99	100	100	100	100	100	100	100	100
CSumExp statement final PC after 5% and WSP	Pearson Correlation	.257**	.221*	.140	.107	.102	.483**	.446**	.215*	.144	1	.904**	.670**	.544**
	Sig (2-tailed)	.010	.029	.170	.290	.317	.000	.000	.032	.153		.000	.000	.000
	N	100	98	98	100	99	100	100	100	100	100	100	100	100
CSumExp HS/OQS	Pearson Correlation	.237**	.203*	.152	.061	.120	.458**	.434**	.197*	.156	.904**	1	.626**	.575**
	Sig (2 tailed)	.003	.012	.060	.448	.138	.000	.000	.050	.120	.000		.000	.000
	N	156	154	153	156	154	100	157	100	100	100	157	100	100
CSumPop statement score final PC after 5% and WSP	Pearson Correlation	.233*	.181	.067	.106	.106	.384**	.316**	.408**	.337**	.670**	.626**	1	.888**
	Sig (2-tailed)	.020	.074	.510	.294	.296	.000	.001	.000	.001	.000	.000		.000
	N	100	98	98	100	99	100	100	100	100	100	100	100	100
CSumPop HS/OQS	Pearson Correlation	.216*	.113	-.024	.043	.087	.285**	.243*	.299**	.282**	.544**	.575**	.888**	1
	Sig (2-tailed)	.031	.269	.812	.673	.390	.004	.015	.003	.004	.000	.000	.000	
	N	100	98	98	100	99	100	100	100	100	100	100	100	100

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Appendix 20: Effects of language and language order on RSC

20.A: RSC of expert templates (EERSC/CERSC)

1

Box's M=9.574, F=1.023, sig.<0.4185

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	7.999	0.0065	0.077	0 800
LANG*LANGORD	3.313	0.0725, n.s.	0.033	0.437
LANG*TXT	1.161	0.2845, n.s.	0.012	0.187
LANG*LANGORD*TXT	1.316	0.2545	0.014	0.206

Mean difference between languages (3.861, std. error=1.365; English=46.079, Chinese=42.218, sig.<0.0065)

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	0.504	0.4805	0.005	0.108
TXT	15.784	0.0005	0.141	0.976
LANGORD*TXT	0.082	0.7755	0.001	0.059

Mean difference between texts (8.628, std. error=2.172; textA=48.462, textC=39.834, sig.<0.0005)

Mean difference between language orders (-1.542, std. error=2.172; English then Chinese=43.378, Chinese then English=44.919, n.s.)

2

Box's M=4.275, F=0.457, sig.<0.9045

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	8 862	0.0045	0.085	0.838
LANG*LANGORD	4.730	0.0325	0.047	0.577
LANG*PRESMODE	4.376	0.0395	0 044	0.544
LANG*LANGORD*PRESMODE	3.079	0.0835, n.s.	0.031	0.412

Mean difference between languages (3.97, std. error=1.334; English=46.346, Chinese=42.376, sig.<0.0045)

Significant interactive effects

LANG*LANGORD

LANGORD	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
English/ Chinese	English	46.785	1.737	43.337	50.232
	Chinese	39.914	2.103	35.740	44.088
Chinese/ English	English	45.907	1.694	42.545	49.269
	Chinese	44.837	2.051	40.766	48.908

LANG*PRESMODE

PRESMODE	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
Computer	English	45.468	1.678	42.137	48.800
	Chinese	44.288	2.032	40 254	48.321
Paper	English	47.223	1.752	43.746	50.700
	Chinese	40.463	2.121	36.254	44 673

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	0.747	0.3905	0 008	0.137
PRESMODE	0.196	0.6595	0.002	0.072
LANGORD*PRESMODE	0.228	0.6345	0.002	0 076

Mean difference between language orders (-2.022, std. error=2.34; English

then Chinese=43.349, Chinese then English=45.372, n.s.)

3

Box's M=6.620, F=0.707, sig.<0.7035

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	9.239	0.0035	0.088	0.853
LANG*TXT	2.029	0.1585	0.021	0.292
LANG*PRESMODE	4.590	0.0355	0.046	0.564
LANG*TXT*PRESMODE	5.192	0.0255	0.051	0.616

Mean difference between the two languages (4.073, std. error=1.340; English=46.073, Chinese=42, sig.<0.0035)

Significant interactive effects

LANG*PRESMODE

PRESMODE	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
Computer	English	45.462	1.606	42.275	48.648
	Chinese	44.260	1.883	40.522	47.998
Paper	English	46.684	1.684	43.341	50.027
	Chinese	39.741	1.976	35.819	43.662

LANG*TXT*PRESMODE

TXT	PRESMODE	LANG	Mean	Std. error	95% confidence interval	
					Lower	Upper
Edu. history	Computer	English	48.885	2.271	44.378	53.392
		Chinese	46.538	2.663	41.252	51.825
	Paper	English	50.296	2.228	45.874	54.719
		Chinese	48.315	2.613	43.127	53.502
Work life balance	Computer	English	42.038	2.271	37.531	46.545
		Chinese	41.981	2.663	36.694	47.267
	Paper	English	43.071	2.526	38.056	48.086
		Chinese	31.167	2.963	25.284	37.049

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	17.254	0.0005	0.152	0.984
PRESMODE	0.586	0.4465	0.006	0.118
TXT*PRESMODE	2.267	0.1355	0.023	0.320

Mean difference between texts (8.944, std. error=2.153)

4

Note: One cell has only 10 participants (see also Research Design), therefore the results should be interpreted with caution.

Box's M=11.783, F=0.520, sig.<0.9645

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	10.838	0.0015	0.105	0.903
LANG*TXT	1.698	0.1965	0.018	0.252
LANG*PRESMODE	5.743	0.0195	0.059	0.660
LANG*LANGORD	4.196	0.0435	0.044	0.527
LANG*TXT*PRESMODE	4.875	0.0305	0.050	0.589
LANG*TXT*LANGORD	1.052	0.3085	0.011	0.174
LANG*PRESMODE*LANGORD	2.507	0.1175	0.027	0.347
LANG*TXT*PRESMODE*LANGORD	1.574	0.2135	0.017	0.237

Mean difference between languages (4.290, std. error=1.303; English=46.131, Chinese=41.814)

Significant interactive effects

LANG*PRESMODE

Presentation mode	Language	mean	Std. error	95% confidence interval	
				Lower	Upper
Computer	English	45.468	1 636	42.218	48.718
	Chinese	44.3	1.872	40.582	48 019
Paper	English	46.794	1.720	43.378	50.210
	Chinese	39.381	1.968	35.473	43.289

LANG*LANGORD

Language order	Language	Mean	Std. error	95% confidence interval	
				Lower	Upper
English then Chinese	English	46.645	1.696	43.276	50 014
	Chinese	39.685	1.941	35.830	43.540
Chinese then English	English	45 617	1.661	42.319	48.915
	Chinese	43.996	1.900	40.223	47.770

LANG*TXT*PRESMODE

Text	Presentation mode	Language	Mean	Std. error	95% confidence interval	
					Lower	Upper
Educ. History	Computer	English	48.885	2.311	44.295	53 474
		Chinese	46.538	2.644	41.288	51.789
	Paper	English	50.504	2 282	45 973	55 036
		Chinese	47.667	2.610	42.482	52.851
Work life balance	Computer	English	42.051	2.318	37.448	46.654
		Chinese	42.063	2.652	36.796	47.329
	Paper	English	43.084	2.574	37.972	48.196
		Chinese	31.095	2.945	25.246	36.945

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	16.199	0.0005	0.150	0.978
PRESMODE	0.671	0.4155	0 007	0.128
LANGORD	0.560	0.4565	0.006	0.115
TXT*PRESMODE	2.090	0.1525	0.022	0.299
TXT*LANGORD	0.036	0 8505	0.000	0.054
PRESMODE*LANGORD	0.189	0.6645	0.002	0 072
TXT*PRESMODE*LANGORD	0.188	0 6665	0.002	0 071

Mean difference between texts (8.825, std. error=2.193; textA=48.398, textC=39.573, sig.<0.0005)

Appendix 20.B: RSC of popular templates (EPRSC/CPRSC)

1

Box's M=10.377, F=1.109, sig.<0.3525.

Tests of within-subjects effects:

Source	F	Sig. <	Partial η^2	Observed power
LANG	25.645	0.0005	0.211	0.999
LANG*LANGORD	3.725	0.0575	0.037	0.480
LANG*TXT	4.308	0.0415	0.043	0.538
LANG*LANGORD*TXT	1.998	0.1615	0.020	0.288

Mean difference between languages (6.366, std. error=1.257; English=52.833, Chinese=46.466, sig.<0.0005)

Interactive effects:
LANG*TXT

LANG	TXT	Mean	Std. error	95% confidence interval	
				Lower	Upper
English	TextA	54.902	1.525	51.876	57.929
	TextC	51.145	1.589	47.992	54.299
Chinese	TextA	50.763	1.617	47.554	53.973
	TextC	41.788	1.685	38.444	45.132

LANG*LANGORD (approaching significance level)

LANGORD	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
English/ Chinese	English	54.868	1.584	51.725	58.011
	Chinese	46.075	1.650	42.800	49.350
Chinese/ English	English	50.797	1.559	47.702	53.893
	Chinese	46.858	1.625	43.633	50.082

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	0.757	0.3865	0.008	0.138
TXT	12.756	0.0015	0.117	0.942
LANGORD*TXT	2.474	0.1195	0.025	0.344

Mean difference between texts (6.748, std. error=1.889)
Mean difference between language orders (1.644, std. error=1.889; English then Chinese=50.472, Chinese then English=48.827, n.s.)

2

Box's M=16.751, F=1.791, sig.<0.0655

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	22.719	0.0005	0.191	0.997
LANG*LANGORD	4.306	0.0415	0.043	0.538
LANG*PRESMODE	0.723	0.3975	0.007	0.134
LANG*LANGORD*PRESMODE	0.096	0.7575	0.001	0.061

Mean difference between languages (6.164, std. error=1.293; English=52.956, Chinese=46.792, sig.<0.0005)

Significant interactive effects
LANG*LANGORD

LANGORD	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
English/ Chinese	English	54.846	1.621	51.628	58.065
	Chinese	46.000	1.814	42.398	49.601
Chinese/ English	English	51.066	1.581	47.927	54.205
	Chinese	47.585	1.770	44.073	51.098

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	0.293	0.5895	0.003	0.084
PRESMODE	0.935	0.3365	0.010	0.160
LANGORD*PRESMODE	0.015	0.9025	0.000	0.052

Mean difference between language orders (1.097, std. error=2.026; English then Chinese=50.423, Chinese then English=49.326, n.s.)

3 Box's M=5.963, F=0.637, sig.<0.7665

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	24.8	0.0005	0.205	0.999
LANG*TXT	5.009	0.0285	0.050	0.601
LANG*PRESMODE	0.897	0.3465	0.009	0.155
LANG*TXT*PRESMODE	1.330	0.2525	0.014	0.208

Mean difference between the two languages (6.4, std. error=1.285; English=52.801, Chinese=46.401, sig.<0.0005)

Significant interactive effects

LANG*TXT

TXT	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
Edu. history	English	54.836	1.547	51.765	57.908
	Chinese	51.312	1.592	48.152	54.473
Work life balance	English	50.766	1.652	47.486	54.045
	Chinese	41.489	1.700	38.114	44.864

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	13.319	0.0005	0.122	0.951
PRESMODE	1.779	0.1855	0.018	0.262
TXT*PRESMODE	0.377	0.5415	0.004	0.093

Mean difference between texts (6.947, std. error=1.904; textA=53.074, textC=46.127)

4 Note: One cell has only 10 participants (see also Research Design), therefore the results should be interpreted with caution.

Box's M=26.123, F=1.153, sig.<0.2835

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	26.208	0.0005	0.222	0.999
LANG*TXT	4.439	0.0385	0.046	0.550
LANG*PRESMODE	1.216	0.2735	0.013	0.194
LANG*LANGORD	3.577	0.0625	0.037	0.465
LANG*TXT*PRESMODE	1.184	0.2795	0.013	0.190
LANG*TXT*LANGORD	2.029	0.1585	0.022	0.291
LANG*PRESMODE*LANGORD	0.260	0.6115	0.003	0.080
LANG*TXT*PRESMODE*LANGORD	1.496	0.2245	0.016	0.228

Mean difference between languages (6.469, std. error=1.264; English=52.820, Chinese=46.351)

Significant interactive effects
LANG*TXT

Presentation mode	Language	mean	Std. error	95% confidence interval	
				Lower	Upper
Edu. History	English	54.887	1.556	51.796	57.979
	Chinese	51.081	1.568	47.966	54.195
Work life balance	English	50.754	1.660	47.457	54.051
	Chinese	41.622	1.672	38.301	44.944

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	12.758	0.0015	0.122	0.942
PRESMODE	1.733	0.1915	0.018	0.256
LANGORD	0.770	0.3825	0.008	0.140
TXT*PRESMODE	0.251	0.6185	0.003	0.079
TXT*LANGORD	2.577	0.1125	0.027	0.355
PRESMODE*LANGORD	0.098	0.7555	0.001	0.061
TXT*PRESMODE*LANGORD	1.674	0.1995	0.018	0.249

Mean difference between texts (6.796, std. error=1.903; textA=52.984, textC=46.188, sig.<0.0015)

Appendix 21: Effects of language and language order on HS

21.A: HS of expert templates (EEHS/CEHS)

1

Box's M=18.683, F=1.203, sig.<0.2605

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	26.438	0.0005	0.149	0.999
LANG*LANGORD	1.689	0.1965	0.011	0.252
LANG*TXT	1.840	0.1625	0.024	0.379
LANG*LANGORD*TXT	1.040	0.3565	0.014	0.229

Mean difference between languages (0.975, std. error=0.190; English=10.443, Chinese=9.467, sig.<0.0005)

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	0.456	0.5015	0.003	0.103
TXT	9.388	0.0005	0.111	0.977
LANGORD*TXT	0.137	0.8725	0.002	0.071

Mean difference between language orders (-0.196, std. error=0.29; English then Chinese=9.857, Chinese then English=10.053, n.s.);
Mean difference between texts (pairwise comparisons: textA/textB=-0.285, n.s.; textA/textC=1.192, sig.<0.0045; textB/textC=1.477, sig.<0.005).

2

Box's M=3.867, F=0.419, sig.<0.9265

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	24.535	0.0005	0.138	0.998
LANG*LANGORD	1.950	0.1655	0.013	0.284
LANG*PRESMODE	1.890	0.1715	0.012	0.277
LANG*LANGORD*PRESMODE	0.003	0.9605	0.000	0.050

Mean difference between languages (0.942, std. error=0.19; English=10.480, Chinese=9.538, sig.<0.0005)

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	0.258	0.6125	0.002	0.080
PRESMODE	0.004	0.9495	0.000	0.050
LANGORD*PRESMODE	0.022	0.8835	0.000	0.052

Mean difference between language orders (-0.155, std. error=0.304; English then Chinese=9.932, Chinese then English=10.086, n.s.)

3

Box's M=12.781, F=0.823, sig.<0.6535

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	29.216	0.0005	0.162	1.000
LANG*TXT	2.552	0.0815	0.033	0.504
LANG*PRESMODE	2.603	0.1095	0.017	0.361
LANG*TXT*PRESMODE	3.750	0.0265	0.047	0.678

Mean difference between the two languages (1.005, std. error=0.186; English=10.448, Chinese=9.443, sig.<0.0005)

Significant interactive effects
LANG*TXT*PRESMODE

TXT	PRESMODE	LANG	Mean	Std. error	95% confidence interval	
					Lower	Upper
Edu. history	Computer	English	10.329	0.399	9.539	11.115
		Chinese	9.538	0.436	8.677	10.4
	Paper	English	10.778	0.392	10.004	11.551
		Chinese	10.407	0.428	9.562	11.253
Let river run	Computer	English	11.607	0.371	10.333	11.801
		Chinese	10.317	0.406	9.515	11.118
	Paper	English	10.833	0.392	10.060	11.607
		Chinese	9.907	0.428	9.062	10.753
Work life balance	Computer	English	9.538	0.399	8.750	10.327
		Chinese	8.962	0.436	8.100	9.823
	Paper	English	10.143	0.444	9.266	11.020
		Chinese	7.524	0.485	6.566	8.482

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	9.750	0.0005	0.114	0.981
PRESMODE	0.008	0.9285	0.000	0.051
TXT*PRESMODE	1.445	0.2395	0.019	0.305

Mean difference between texts (textA/textB=-0.271, n.s.; textA/textC=1.205, sig.<0.0045; textB/textC=1.476, sig.<0.0005)

4 Note: One cell has only 10 participants (see also Research Design), therefore the results should be interpreted with caution.

Box's M=34.249, F=0.968, sig.<0.5205
Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	30.157	0.0005	0.172	1.000
LANG*TXT	2.290	0.1055	0.031	0.459
LANG*PRESMODE	3.522	0.0635	0.024	0.462
LANG*LANGORD	1.815	0.1805	0.012	0.268
LANG*TXT*PRESMODE	3.468	0.0345	0.046	0.641
LANG*TXT*LANGORD	1.013	0.3665	0.014	0.224
LANG*PRESMODE*LANGORD	0.016	0.8995	0.000	0.052
LANG*TXT*PRESMODE*LANGORD	2.897	0.0585	0.038	0.559

Mean difference between languages (1.014, std. error=0.185; English=10.449, Chinese=9.435)

Significant interactive effects
LANG*TXT*PRESMODE

Text	Presentation mode	Language	Mean	Std. error	95% confidence interval	
					Lower	Upper
Educ. History	Computer	English	10.327	0.406	9.524	11.130
		Chinese	9.538	0.433	8.682	10.395
	Paper	English	10.800	0.401	10.007	11.593
		Chinese	10.296	0.428	9.450	11.142
Work life balance	Computer	English	11.069	0.386	10.306	11.833
		Chinese	10.403	0.412	9.589	11.217
	Paper	English	10.833	0.401	10.040	11.627
		Chinese	9.867	0.428	9.021	10.713

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	9.420	0.0005	0.115	0.977
PRESMODE	0.051	0.8215	0.000	0.056
LANGORD	0.428	0.5145	0.003	0.100
TXT*PRESMODE	1.377	0.2565	0.019	0.293

	TXT*LANGORD	0.102	0.9035	0.001	0.065
	PRESMODE*LANGORD	0.006	0.9365	0.000	0.051
	TXT*PRESMODE*LANGORD	0.753	0.4735	0.010	0.176
	Mean difference between texts (textA/textB=-0.303, n.s.; textA/textC=1.198, sig.<0.0045; textB/textC=1.501, sig.<0.0005)				

21.B: HS of popular templates (EPHS/CPHS)

1

Box's M=10.936, F=1.169, sig.<0.3105.

Tests of within-subjects effects:

Source	F	Sig. <	Partial η^2	Observed power
LANG	20.988	0.0005	0.179	0.995
LANG*LANGORD	1.388	0.2425	0.014	0.215
LANG*TXT	0.955	0.3315	0.010	0.162
LANG*LANGORD*TXT	2.682	0.1055	0.027	0.368

Mean difference between languages (1.142, std. error=0.249; English=11.298, Chinese=10.156, sig.<0.0005)

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	2.173	0.1445	0.022	0.309
TXT	7.648	0.0075	0.074	0.782
LANGORD*TXT	4.637	0.0345	0.046	0.568

Mean difference between texts (0.883, textA=11.168, textC=10.286)
Mean difference between language orders (0.471, std. error=0.319; English then Chinese=10.962, Chinese then English=10.492, n.s.)

Significant interactive effects
LANGORD*TXT

TXT	LANGORD	Mean	Std. error	95% confidence interval	
				Lower	Upper
Edu. History	English/CHN	11.060	0.318	10.428	11.692
	Chinese/ENG	11.277	0.301	10.680	11.874
Work life balance	English/CHN	10.865	0.325	10.220	11.510
	Chinese/ENG	9.707	0.332	9.048	10.365

2

Box's M=12.207, F=1.305, sig.<0.2285
Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	19.644	0.0005	0.170	0.992
LANG*LANGORD	1.802	0.1835	0.018	0.264
LANG*PRESMODE	0.360	0.5505	0.004	0.091
LANG*LANGORD*PRESMODE	0.123	0.7265	0.001	0.064

Mean difference between languages (1.123, std. error=0.253; English=11.330, Chinese=10.206, sig.<0.0005)

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	1.370	0.2455	0.014	0.212
PRESMODE	0.001	0.9805	0.000	0.050
LANGORD*PRESMODE	0.024	0.8775	0.000	0.053

Mean difference between language orders (-0.397, std. error=0.339; English then Chinese=10.967, Chinese then English=10.569, n.s.)

3

Box's M=10.657, F=1.139, sig.<0.3315
Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	20.154	0.0005	0.174	0.994
LANG*TXT	1.225	0.2715	0.013	0.195
LANG*PRESMODE	0.395	0.5315	0.004	0.095
LANG*TXT*PRESMODE	0.646	0.4235	0.007	0.125

Mean difference between the two languages (1.141, std. error=0.254;

English=11.297, Chinese=10.156, sig.<0.0005)

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	7.357	0.0085	0.071	0.766
PRESMODE	0.101	0.7515	0.001	0.061
TXT*PRESMODE	0.530	0.4685	0.005	0.111

Mean difference between texts (0.894, std. error=0.330; textA=11.173, textC=10.280)

4 Note: One cell has only 10 participants (see also Research Design), therefore the results should be interpreted with caution.

Box's M=23.575, F=1.041, sig.<0.4085

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	21.039	0.0005	0.186	0.995
LANG*TXT	0.976	0.3265	0.010	0.165
LANG*PRESMODE	0.527	0.4705	0.006	0.111
LANG*LANGORD	1.348	0.2495	0.014	0.210
LANG*TXT*PRESMODE	0.480	0.4905	0.005	0.105
LANG*TXT*LANGORD	2.539	0.1155	0.027	0.351
LANG*PRESMODE*LANGORD	0.033	0.8555	0.000	0.054
LANG*TXT*PRESMODE*LANGORD	0.683	0.4115	0.007	0.129

Mean difference between languages (1.163, std. error=0.253; English=11.305, Chinese=10.142)

No significant interactive effects of LANG with other associated factors

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	7.257	0.0085	0.073	0.760
PRESMODE	0.068	0.7955	0.001	0.058
LANGORD	2.321	0.1315	0.025	0.326
TXT*PRESMODE	0.356	0.5525	0.004	0.091
TXT*LANGORD	4.668	0.0335	0.048	0.571
PRESMODE*LANGORD	0.144	0.7065	0.002	0.066
TXT*PRESMODE*LANGORD	1.425	0.2365	0.015	0.219

Mean difference between texts (0.874, std. error=0.324; textA=11.16, textC=10.287, sig.<0.0085)

Significant interactive effects

TXT*LANGORD

Text	Language order	Mean	Std. error	95% confidence interval	
				Lower	Upper
Edu. History	English/Chinese	11.057	0.322	10.417	11.697
	Chinese/English	11.263	0.305	10.658	11.869
Work life balance	English/Chinese	10.884	0.333	10.222	11.546
	Chinese/English	9.689	0.336	9.022	10.357

Appendix 22: Effects of language and language order on the lengths of summaries

1

Box's M=14.107, F=0.097, sig.<0.5565.

Tests of within-subjects effects:

Source	F	Sig. <	Partial η^2	Observed power
LANG	454.204	0.0005	0.758	1.000
LANG*LANGORD	4.276	0.0405	0.029	0.538
LANG*TXT	6.291	0.0025	0.080	0.892
LANG*LANGORD*TXT	3.830	0.0245	0.050	0.688

Mean difference between languages (195.293, std. error=9.164; English=300.937, Chinese=496.231, sig.<0.0005)

Significant interactive effects

LANG*LANGORD

LANGORD	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
English then Chinese	English	317.414	7.098	303.384	331.443
	Chinese	493.758	14.282	465.530	521.986
Chinese then English	English	284.461	7.368	269.898	299.024
	Chinese	498.703	14.825	469.402	528.005

LANG*TXT

TXT	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
Edu. history	English	305.018	8.934	287.361	322.674
	Chinese	541.620	17.975	506.094	577.147
Let the river run	English	307.556	8.531	290.694	324.418
	Chinese	465.841	17.166	431.914	499.768
Work life balance	English	290.239	9.107	272.240	308.238
	Chinese	481.231	18.323	445.017	517.445

LANG*LANGORD*TXT

Text	Language order	Language	Mean	Std. error	95% confidence interval	
					Lower	Upper
Educ. History	English then Chinese	English	319.304	13.015	293.581	345.028
		Chinese	501.087	26.187	449.330	552.844
	Chinese then English	English	290.731	12.241	266.537	314.925
		Chinese	582.154	24.630	533.474	630.833
Let the river run	English then Chinese	English	329.938	11.034	308.129	351.746
		Chinese	488.812	22.201	444.933	532.692
	Chinese then English	English	285.174	13.015	259.450	310.898
		Chinese	442.870	26.187	391.113	494.626
Work life balance	English then Chinese	English	303.000	12.741	277.818	328.182
		Chinese	491.375	25.635	440.708	542.042
	Chinese then English	English	277.478	13.015	251.755	303.202
		Chinese	471.087	26.187	419.330	522.844

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	1.088	0.2995	0.007	0.179
TXT	3.367	0.0375	0.044	0.628
LANGORD*TXT	2.521	0.0845	0.034	0.498

Summarization language order (E/C or C/E) did not have significant effects on the summary length (mean difference=14.004, std. error=13.426; E/C=405.586, C/E=391.582, n.s.). Neither was there significant interactive effect of LANGORD with TXT. There was significant main effect of TXT on the averaged summary length (F=3.367, sig.<0.0375), however, the pairwise comparisons indicated that there was on significant difference in the averaged

summary length between the three texts. In other words, it seemed that the three texts (mean difference: textA/textB=36.621, sig.<0.0765; textA/textC=37.548, sig.<0.0795; textB/textC=0.936, n.s.) were in a homogeneous group (c.f. analyses of TXT effects on summary length in Chapter 10).

2 Box's M=13.192, F=1.429, sig.<0.1695

Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	425.743	0.0005	0.743	1.000
LANG*LANGORD	5.711	0.0185	0.037	0.611
LANG*PRESMODE	7.755	0.0065	0.050	0.790
LANG*LANGORD*PRESMODE	0.182	0.6705	0.001	0.071

Mean difference between languages (194.334, std. error=9.418; English=301.293, Chinese=495.627).

Significant interactive effects

LANG*LANGORD

LANGORD	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
English then Chinese	English	318.015	7.038	304.105	331.924
	Chinese	489.841	14.412	461.359	518.323
Chinese then English	English	284.571	7.346	270.053	299.090
	Chinese	501.413	15.043	471.684	531.142

LANG*PRESMODE

Presentation mode	LANG	Mean	Std. error	95% confidence interval	
				Lower	Upper
Computer	English	307.640	6.986	293.833	321.446
	Chinese	528.201	14.306	499.929	556.473
Paper	English	294.946	7.396	280.331	309.562
	Chinese	463.052	15.144	433.124	492.981

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANGORD	0.664	0.4165	0.004	0.128
PRESMODE	8.413	0.0045	0.054	0.822
LANGORD*PRESMODE	0.190	0.6635	0.001	0.072

Mean difference between language orders (10.936, std. error=13.418; English then Chinese=403.928, Chinese then English=392.992, n.s.)
However, there was significant main effect of PRESMODE on summary length (see Chapter 9 *Text presentation mode and computer familiarity*)

3 Box's M=23.06, F=1.483, sig.<0.1025

Source	F	Sig. <	Partial η^2	Observed power
LANG	450.918	0.0005	0.757	1.000
LANG*TXT	7.281	0.0015	0.091	0.933
LANG*PRESMODE	8.295	0.0055	0.054	0.816
LANG*TXT*PRESMODE	1.599	0.2065	0.022	0.334

Mean difference between the two languages (-194.289, std. error=9.15; English=301.25, Chinese=495.539)

Significant interactive effects:

LANG*TXT

Text	Language	Mean	Std. error	95% confidence interval	
				Lower	Upper
Edu. History	English	304.158	9.144	286.085	322.230
	Chinese	544.043	17.409	509.635	578.450

	Chinese	467.522	16.453	435.004	500.040
Work life balance	English	288.676	9.388	270.121	307.231
	Chinese	475.053	17.873	439.728	510.378

LANG*PRESMODE

Presentation mode	Language	Mean	Std. error	95% confidence interval	
				Lower	Upper
Computer	English	308.604	7.169	294.435	322.773
	Chinese	529.245	13.649	502.269	556.221
Paper	English	293.896	7.624	278.827	308.965
	Chinese	461.833	14.515	433.144	490.523

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	3.918	0.0225	0.051	0.698
PRESMODE	9.944	0.0025	0.064	0.880
TXT*PRESMODE	2.671	0.0735	0.036	0.523

Mean differences between texts using pairwise comparisons (textA/textB=34.881, n.s.; textA/textC=42.236, sig.<0.0325; textB/textC=7.355, n.s.)
Mean differences between presentation modes (41.06, std. error=13.021; computer=418.924, paper=377.865)

4 Note: One cell has only 10 cases.
Box's M=31.159, F=0.878, sig.<0.6685
Tests of within-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
LANG	468.694	0.0005	0.771	1.000
LANG*TXT	6.901	0.0015	0.090	0.919
LANG*PRESMODE	9.174	0.0035	0.062	0.853
LANG*LANGORD	4.837	0.0305	0.034	0.589
LANG*TXT*PRESMODE	1.494	0.2285	0.021	0.314
LANG*TXT*LANGORD	4.012	0.0205	0.055	0.709
LANG*PRESMODE*LANGORD	0.056	0.8145	0.000	0.056
LANG*TXT*PRESMODE*LANGORD	1.112	0.3325	0.016	0.242

Mean difference between languages (193.345, std. error=8.931; English=300.466, Chinese=493.811)

4 Significant interactive effects
LANG*TXT

Text	Language	Mean	Std. error	95% confidence interval	
				Lower	Upper
Edu. History	English	304.992	8.924	287.347	322.637
	Chinese	541.081	17.098	507.274	574.887
Let river run	English	307.265	8.530	290.399	324.130
	Chinese	463.821	16.343	431.508	496.133
Work life balance	English	289.142	9.160	271.031	307.254
	Chinese	476.532	17.551	441.831	511.232

LANG*PRESMODE

Presentation mode	Language	Mean	Std. error	95% confidence interval	
				Lower	Upper
Computer	English	306.926	7.023	293.040	320.812
	Chinese	527.321	13.456	500.717	553.926
Paper	English	294.007	7.464	279.250	308.764
	Chinese	460.301	14.300	432.028	488.574

LANG*LANGORD

Language order	Language	Mean	Std. error	95% confidence interval	
				Lower	Upper
English then Chinese	English	316.771	7.130	302.675	330.868
	Chinese	490.474	13.660	463.465	517.482
Chinese then English	English	284.162	7.362	269.606	298.717
	Chinese	497.149	14.105	469.261	525.036

LANG*TXT*LANGORD

Text	Language order	Language	Mean	Std. error	95% confidence interval	
					Lower	Upper
Edu. History	English then Chinese	English	319.254	13.007	293.536	344.972
		Chinese	500.008	24.921	450.734	549.281
	Chinese then English	English	290.731	12.222	266.565	314.896
		Chinese	582.154	23.417	535.854	628.453
Let river run	English then Chinese	English	329.075	11.039	307.249	350.900
		Chinese	485.706	21.149	443.890	527.522
	Chinese then English	English	285.455	13.007	259.737	311.172
		Chinese	441.936	24.921	392.662	491.209
Work life balance	English then Chinese	English	301.986	12.902	276.476	327.495
		Chinese	485.707	24.719	436.833	534.581
	Chinese then English	English	276.299	13.007	250.581	302.017
		Chinese	467.356	24.921	418.083	516.629

Tests of between-subjects effects

Source	F	Sig. <	Partial η^2	Observed power
TXT	4.031	0.0205	0.055	0.711
PRESMODE	9.652	0.0025	0.065	0.870
LANGORD	1.016	0.3155	0.007	0.170
TXT*PRESMODE	2.541	0.0825	0.035	0.501
TXT*LANGORD	2.681	0.0725	0.037	0.524
PRESMODE*LANGORD	0.063	0.8025	0.000	0.057
TXT*PRESMODE*LANGORD	2.308	0.1035	0.032	0.462

Mean difference between texts using pairwise comparisons (textA/textB=37.494, sig.<0.0515 approaching significance level; textA/textC=40.199, sig.<0.0405; textB/textC=2.706, n.s.)

Mean difference between presentation modes (39.97, std. error=12.866; computer=417.124, paper=377.154)

Appendix 23: Descriptive statistics of PSQ data

1. Was the text overall easy or difficult for you to

	very difficult	somewhat difficult	moderately easy/difficult	somewhat easy	easy
(1.a) read to understand?	1	35	85	32	3
(1.b) read to summarize?	9	72	59	13	0

2. Were you familiar with the topic of **A*** before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all
0	11	33	81	31

3. If you were familiar with the topic of **A** before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(3.a) read to understand the text?	11	55	34	24	1
(3.b) read to summarize the text?	7	34	45	34	3

4. To which activity do you think your familiarity with the topic of **A** before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities
78	14	32

5. Were you familiar with the topic of **B*** before reading the text?

very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all
1	3	13	57	78

6. If you were familiar with the topic of **B** before reading the text, how helpful was this for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(6.a) read to understand the text?	15	26	27	12	1
(6.b) read to summarize the text?	6	28	26	19	1

7. To which activity do you think your familiarity with the topic of **B** before reading the text was more helpful?

read to understand the text	read to summarize the text	equally helpful (or equally not helpful) to the 2 activities
43	15	22

8. Are you familiar with the following two tests you've just done?

	very familiar	somewhat familiar	of average familiarity	not too familiar	not familiar at all
(8.a) English summarization	5	38	55	53	3
(8.b) Chinese summarization	9	9	49	80	10

Note: A refers to the general topic of the source text, B the specific topic.

9. Did you write summaries like these in your university courses?

Yes	No
115	41

10. How much do you think your ability to write the summary in English depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(10.a) English reading abilities?	44	81	27	4	0
(10.b) English writing abilities?	22	57	65	11	1

11. On which ability do you think your English summary depends most?

English reading abilities	English writing abilities	equally (in)dependent on the 2 types of abilities
71	23	59

12. How much do you think your ability to write the summary in Chinese depends on your

	highly dependent	fairly dependent	moderately (in)dependent	fairly independent	highly independent
(12 a) English reading abilities?	49	75	26	3	2
(12 b) Chinese writing abilities?	24	63	60	7	2
(12 c) English to Chinese translation abilities?	31	68	44	6	5

13. On which ability do you think your summary in Chinese depends most?

English reading abilities	Chinese writing abilities	English to Chinese translation abilities
61	33	60

14. Which language do you prefer to use to summarize the text, English or Chinese?

English	Chinese	I don't mind which language
84	42	29

15. Can you explain the reasons for your answer to Question 14?

16. You were asked to summarize the same text in both Chinese and English. Which task do you think can better measure your English reading abilities?

English summarization	Chinese summarization	Equally well
41	71	41

17. In which order did you summarize the text in both languages?

English then Chinese	Chinese then English
82	75

18. In which order would you prefer to summarize the same text in both languages?

English then Chinese	Chinese then English	I don't mind the order
83	36	34

19. Can you explain the reasons for your answer to Question 18?

20. Are there any other comments you would like to make regarding the text and the summarization tasks?

21. How helpful was your level of computer familiarity for you to

	very helpful	somewhat helpful	of average help	not too helpful	not helpful at all
(21.a) read to understand the text?	6	34	7	27	4
(21.b) read to summarize the text?	4	25	15	25	7

22. To which activity do you think your computer familiarity level was more helpful?

read to understand the text	Read to summarize the text	equally helpful (or equally not helpful) to the 2 activities
29	20	28

Appendix 24: Reasons for preferring to English summarization tasks: breakdown of responses

No. of participants	**Major theme (a) : Direct/straight copying from or referring to the source texts without full understanding**
4	Could copy some English sentences straight from the source text.
1	English summarization only requires a general understanding of the text; you don't have to understand completely every bit of the text.
1	When you come across new words which would affect your understanding of the text, you could copy directly from the source text into your English summaries.
1	I do more English than Chinese writing, and read more English than Chinese texts. And you can also copy some information directly from source text to your English summaries.
1	When writing an English summary, you can have reference to the source and even copy sentences exactly the same as in the source even if you do not really understand.
1	You can copy directly from the source texts. I am not satisfied with my not-so-good Chinese writing abilities; I can not express myself clearly in Chinese.
1	In English summarization, no translation is required, so you can copy directly from the source text when you do not understand.
1	This (English summarization) is just like squeezed juice, you can see it in its original form. It is not easy to go wrong, go in a wrong direction.
1	It is not necessary to have a deep understanding of the source text, when you summarize it in English. You can find and edit the topic sentences. Use it straight from the source. In Chinese summarization, you need to have a full understanding of the source text, and therefore it is more challenging.
1	Being able to copy from the source text can compensate my poor abilities in translation.
1	You can copy directly from the source, no translation is required. It does not really matter even if you do not understand.
1	It is easier to summarize in English, because you are less likely to make ambiguous statements/writings.
1	It is more convenient in English summarization to use proper words and sentence structures, and therefore can convey the original meaning correctly. English summarization can also guarantee correct use of grammar.
1	From English to English, it is easy.
1	When you read an English text, you are thinking in an "English" way.
1	I think I can write better in English than in Chinese. In English summarization, you can also imitate the syntax of the original text.
1	I think it easier to summarize an English text in English.
1	Some words are ready made.
1	I used more reasoning in English than Chinese to finish the two summarization tasks. What's more, there are things already there in the source text that we only need to modify them and then use them in the English summary.
	Major themes : (b) additional processing such as translation and (c) the issue of translatability between the two languages
2	When you summarize it in English, you can have direct reference to the source text. You only need to make some slight changes to the original sentences. However, the Chinese summary would have to be conveyed in a "polished" way, which requires more serious planning and thinking, as a result it is more challenging.
2	The source text is in English, so you do not have to translate in English summarization tasks.
2	You only need to understand the source text and then re-organize it for an English summary; however, Chinese summarization requires additional translation abilities.

2	Easier. You can find what you need in the source text, and do not need to use your own words. However, in Chinese summarization tasks, you need to learn how to translate them into proper Chinese.
2	English and Chinese are two different languages, and use different systems to convey the same meaning.
1	It is more natural to summarize an English text in English because you do not have to translate it into Chinese.
1	It is easier to write an English summary than writing a Chinese summary from an English source text. More efforts needed in writing the Chinese summary. In English summarization, you only need to have a general understanding of the source.
1	In English summarization, there are resources you can copy directly from the original text; however, Chinese summarization requires to some extent not only your translation but also Chinese writing abilities.
1	When writing an English summary, you can copy the important sentences from the source text. Even if you have difficulty in understanding some parts of the text, you can still finish the English summarization task, but in Chinese summarization, it is a different situation. It is more difficult to finish; it requires translation and a good command of how to organize your language.
1	It is more convenient to summarize it in English, because you can copy and edit the key or topic sentences. However, in Chinese summarization, additional abilities such as translation and Chinese writing are essential. My Chinese is poor.
1	In Chinese summarization, there is an additional processing step, that is, to translate from English to Chinese; and in English summarization, you could have direct reference to the source text.
1	In Chinese summarization task, there is a complex process of translation from English to Chinese; furthermore, it actually has a high demand of your translation abilities.
1	In Chinese summarization, you need to think very carefully to choose the most appropriate Chinese words to express your meaning you got from an English text. It is just the same as you drive around a corner; it really takes too much time to do that.
1	Summarizing in English can improve my English writing abilities, and summarizing in Chinese requires a high command of translation and Chinese writing abilities
1	When you are translating, you may not understand the new words; but when these new words are in the source text, within a certain context, you can probably guess what they are, though not their exact meaning.
1	It is easier and more faithful (to the source text) to summarize it in a language of its original form. If language other than its original form is required, translation abilities are required.
1	Chinese summarization is more difficult, because it requires re-arranging your ideas to make it coherent, in addition to the understanding of the source text first of all.
1	In addition to understanding and generalizing information from source text, which is essential to successful English summarization, Chinese summarization also requires your translation abilities to a certain extent.
1	Because I do not have background information about this (work-life balance) campaign, and I do not know how to translate some of the technical terms into proper Chinese. What's more, I am not good at Chinese writing, can't express myself in a proper way.
1	You can think in English and save a step of translation into Chinese. Sometimes, it is only in English that the original meaning can be conveyed faithfully, exactly the same as the source.
1	Because you do not have to switch your language during summarization, it can save time. Translation means distortion. It is inevitable that your translation can be slightly different from the source.
1	It is easier to summarize an English text in English, you only need to find the topic sentences and generalize them. However, in Chinese summarization, you also need to translate into Chinese.

1	We have written more English than Chinese summaries, and my abilities in translation and writing are not satisfactory.
1	As an English major in the university, we should write English summaries. It is also possible to use the original words. Sometimes, writing a Chinese summary can be laborious; you have to translate.
1	No need for switching your language or the structure.
1	You can use original sentences from the source text, no need for translation.
1	After these years of learning English, it seems that our reasoning abilities in Chinese are infringed and getting worse. What's more, we are more used to writing summaries in English, it is quite hard to accept Chinese summarization tasks.
1	Translation is required when you produce a Chinese summary of an English text. If you do not know some words, you can not express them in Chinese at all.
1	The process of English summarization is less complex; you can find the major theme of the source text directly from the text, and do not have to translate them into Chinese. A lot of troubles therefore are avoided.
1	Chinese summarization has a higher demand of translation and reading abilities, and it also requires good organization skills in Chinese.
1	I prefer English summarization because it is not possible to have proper Chinese translations for some English words and phrases. English summary can better reflect the original meaning of the source text.
1	Some key English words/phrases are already there in the source text, when they are being translated into Chinese, it is so difficult. Even if they are translated, the Chinese summary does not look like Chinese.
1	When you summarize it in Chinese, you have to re-organize the ideas substantially; but when you summarize it in English, you have far more chances to delete or even add something which are already in the source text, it is much easier.
1	I do not have good translation ability, and I am more familiar with English summarization tasks. English summarization tasks can also save your time and energy.
1	Chinese summarization entails translation. It would be more challenging to summarize (in Chinese) a challenging source text.
1	After years of learning English, I have got a strange "disease" – I can not have an appropriate Chinese equivalent for the English words instantly. I find it is easier to summarize in English, it is more natural.
1	Chinese summarization has an additional processing step – translation. In this process, if you make some mistakes, the end product (the Chinese summary) could deviate from the original meaning. English summarization is easier because you can edit the original sentences already in the source text.
1	It is not necessary to switch your languages.
1	No need to translate it into Chinese.
1	The source text is in English. It is not easy, if not impossible, to replace some English words with proper Chinese equivalents.
1	It is more convenient to summarize it in English than in Chinese. In particular, it is so difficult to translate it and have a proper Chinese sentence structure. Some proper names are simply not translatable.
1	Some sentences are not translatable.
1	Translation from English to Chinese is difficult. However, in English summarization, you can select what you need from the source.
1	Although some words were difficult to translate, they did not affect the overall understanding the source text. It is easier.
1	No need to translate.
1	Without additional process of translation, it would be more convenient and time-saving.
	Miscellaneous
2	More familiar with writing a summary in English.

1	I am English major in the university, and have almost forgotten how to write in Chinese.
1	I am majored in English language, it is relatively easier to understand the source text and to write an English summary.
1	Not good at Chinese writing
1	Easier to organize your language.
1	I can express myself better in English.
1	I can express myself better in English. My Chinese language abilities are just so-so.
1	Firstly, I felt it easier to summarize in English after I had already done the Chinese summarization task and understood the source text and had a clear plan what to write next. Secondly, some original sentences can be lifted from the source text.

Appendix 25: Reasons for preferring to Chinese summarization tasks: breakdown of responses

Number of participants	**Major Themes A&B** **Familiarity with the Chinese language and Chinese summarization tasks;** **Facility of the mother tongue to write concise Chinese summaries**
5	More familiar with Chinese
3	Easier to organize in the mother tongue.
2	More familiar with Chinese summarization, and easier to produce a Chinese summary.
2	Can express ourselves better in Chinese
1	Easier, more direct and familiar.
1	Comparatively speaking, I like Chinese better than English.
1	Understanding the source text first of all, and then it would be easier to write the summary in Chinese, our mother tongue.
1	Easier to summarize in Chinese. I know more Chinese than English words.
1	After all, our Chinese language abilities are much more advanced than English abilities. It would be easier to express in Chinese and the Chinese summaries would also be more concise.
1	We can use Chinese more freely and easily, and what's more, we are more familiar with summarizing in Chinese.
1	You feel more at home writing in Chinese
1	Chinese is our mother tongue, you feel easy and confident in using the language.
1	Chinese is the mother tongue.
1	Chinese is the mother tongue; I can use it at ease. What's more, I like Chinese better than English.
1	Chinese is the mother tongue, I am more familiar with Chinese; and I am good at Chinese. Another far more important reason is that I think the Chinese language is more beautiful and can convey more information with the same amount of words.
1	Your can better express yourself - express what you want to express in Chinese. However, you can not express freely in English because of some difficulty in vocabulary, syntax and grammar, etc.
1	Compared to our mother tongue, reading and summarizing in English was more difficult. When you are writing a summary in English, you are not competent enough to use one English word that can replace faithfully a whole sentence or even a paragraph. However, it is possible in Chinese.
1	Chinese summarization is not exactly the same as translation; it needs some generalisation of the ideas from the source text. Since Chinese is the mother tongue, it is easier to use; and you also feel free and more confident to use your mother tongue. Using English is another case.
1	Personally, I find Chinese summarization easier.
1	Personally, I think my Chinese writing abilities are much much better than my English reading abilities, and I am poor in writing in English.
1	Easier, do not have to worry about grammatical mistakes.
1	I found it very difficult to use short English sentence to produce a gist.
1	Chinese summaries can better reflect our degree of understanding of the source text. A summary needs to be concise and dense. Chinese summary can better meet this requirement of a summary. On the other hand, my preference of Chinese summarization may also indicate that I am not good at expressing myself in English freely.
1	You can use your own words to write a gist of the source text, without having to worry about English grammatical mistakes resulting from your lack or unfamiliarity with the English language.

1	When you have understood the source text, it is in Chinese that you can produce its gist. This is mainly because I haven't fully master the English language.
1	It is easier to express your ideas in Chinese, while it is far more difficult to organize your ideas in English. I am not good at expressing myself in English.
1	My English writing is poor.
	Major Theme C: **Natural ingredient of Chinese summarization in the process of comprehension**
1	When I summarized it in English, I felt there were a lot that I considered important and that should be in the English summary. However, when summarizing it in Chinese, I only needed to write down the summary that was **already in my mind**.
1	Summarization is a process of understanding and transferring from written form, and then uptaking and finally producing in another written form. It consists of three phases which all require the switching of languages. The Chinese summarization tasks apparently can skip some of the steps, and therefore it is easier to produce a Chinese summary.
1	As long as you have some understanding of the source text, it is easier to summarize in Chinese. In English summarization, you have to think in an "English" way, in terms of vocabulary, grammar and sentence structure.
1	I started to write the English summary while I was still reading the text, and therefore, I did not fully grasped the overall structure of the source text. When I summarized it in Chinese, I knew well its overall structure, and therefore found it easier to summarize in Chinese.
1	I am used to writing in Chinese; even the thinking process is also in Chinese.
	Miscellaneous
1	It is fun to translate.
1	Both English and Chinese summarization tasks test our English language abilities, although Chinese summarization task also requires translation.

Appendix 26: Reasons for "don't mind" which language to use for the summarization tasks: breakdown of responses

Number of participants	**Major Theme A:** **Understanding is the primary prerequisite for successful summarization, be it in English or in Chinese**
4	As long as you understand the source text, it does not matter which language to use.
1	Successful summarization relies on understanding of the source text. It is based on the understanding that we decide which part to keep in a summary. Language used, English or Chinese, is only the means to express your understanding of the text.
1	If you can have a general understanding of the text, it does not really matter.
1	No matter which language to use, you need to understand the text first. It makes no difference which language you use.
1	Generally speaking, which language to summarize the source text does not make too much difference if you can understand it. Neither my Chinese nor English writing is good. These two writing skills are related to a great extent.
1	It is fundamentally the same, English and Chinese summarization both require your understanding of the text first of all.
1	Language is to communicate. Both languages can communicate. Both are OK.
1	To write a good summary, whether in English or Chinese, it is absolutely important that we first of all need to understand the source text. I think both are fine.
	Major Theme B: **Both English and Chinese summarization tasks are already challenging enough; Establishing a balanced view of the advantages and disadvantages of either language**
1	Both English and Chinese summarization is challenging.
1	Neither am I good at. Writing summary is laborious.
1	Chinese summarization may be easier to organize your ideas. After all, our Chinese language abilities are more advanced than English. However, we are more familiar with English summarization tasks which we also do in our university study.
1	Both have got their merits. Chinese is the mother tongue and we would find it easier to organize the language. English summarization can avoid translating those words that can not be easily translated, and you can also modify and use directly the original sentences from the source text.
1	In English summarization, you do not have to translate those technical and proper terms. It could be very difficult if you have to express the English technical and proper terms in Chinese. What's more, generally speaking, it is easy to write in Chinese because English is a foreign language after all.
1	In English summarization, you can copy directly from source text; and in Chinese summarization, you feel more at home in writing, therefore there is no big difference between them. I don't really mind which language to use.
1	Easier to write Chinese summary because Chinese is our mother tongue, but it is helpful for improving our English if we write a summary in English.
	Miscellaneous
1	Although I believe the summarization tasks can test my language abilities, I can not finish the tasks because of my low language abilities.
1	It is very rare that we did either of the summarization tasks, so I am not used to such tasks and I have not developed any preference yet.
1	Seldom did I write summaries

Appendix 27: Reasons for preferences to language order: breakdown of responses

(a) Reasons for the preference to *English then Chinese*

Number of participants	**Major Theme A:** **Natural direction of reading comprehension and summarization processes – from English to Chinese;**
6	During English summarization process, I had a general understanding of the text, and this general understanding is helpful for Chinese summarization, making it to the point. It would be more difficult if you started with Chinese summarization.
3	This is right the thinking process.
3	This is the order of my thinking process.
1	It is right the order of my comprehension process (from English to Chinese), and therefore facilitates to *highly* condense the source text in Chinese.
1	It would sound too abrupt to start with Chinese summarization.
1	I think English summarization is easier and more convenient than Chinese summarization. English summarization actually serves as a good premise for the Chinese summarization task later.
1	When you write the English summary, you have already read carefully the source text several times. If you can write a good English summary, it won't be difficult to write a Chinese summary.
1	If I had summarized it first in Chinese, the English summary would be confined by the Chinese summary and would not look like English.
1	Direct, and natural order
1	A text of 6 pages long is difficult to understand, we need to first of all list the key or important English sentences from the source text and then summarize them in English and then in Chinese.
	Major Theme B: **Easier direction of translation – from English to Chinese**
6	Translate the English summary into Chinese, that's it.
5	English summarization helps you to have a clear thread of the development of the source text. Chinese summary can be the translation of the English one.
3	When you write the Chinese summary later, it is more or less the same as translating the English summary into Chinese.
1	Summarization in English (first) strengthens your understanding of the source text, and when you write the Chinese summary later, it is more or less the same as translating the English summary into Chinese.
1	You can translate the English summary into Chinese easier than you translate a Chinese summary into English.
1	You can translate the English summary into Chinese, and that becomes a Chinese summary then.
1	It would be a Chinese summary after some slight changes to an English one.
1	It is easier to translate English into Chinese than to translate Chinese into English.
	Miscellaneous
1	Because the source text is in English.
1	Easier
1	"Easy" to "difficult"

(b) Reasons for preference to Chinese then English

Number of participants	**Major Theme A:** **Facility of Chinese summarization to the subsequent English summarization task**

4	This is easier.
3	It would be from "difficult" to "easy", and also helps with understanding the source text. I like get difficult things done first.
2	Chinese summarization can better help me understand the source text than English summarization.
2	Chinese summarization helps to understand the text, deeper understanding of it.
1	Chinese summary provides a helpful structure for English summarization later.
1	During Chinese summarization process, I had a general understanding of the source text. And when I write the summary in English, I could find the sentences needed for the English summary in the source text, quickly and easily.
1	During Chinese summarization, you have not only understood the text but also organized your summary, therefore, when you write an English summary later, it would be much easier. Otherwise, you would still have to re-organize your sentences and ideas.
1	With the Chinese summary in hand, you feel confident and easier to write an English summary.
1	The Chinese summary can convey the gist of the source text, and then you just need to translate it into English; or you can follow the structure/organization of the Chinese summary when you are writing the English summary. Otherwise, it would seem aimless, lifting one English sentence here and another one there from the source text.
	Major Theme B: Thinking in a Chinese way
1	Very often, unconsciously, I translate an English text into Chinese when I am reading.
1	After all, we are Chinese, and think in a Chinese way.
	Miscellaneous
1	Easier to control the number of words, after all, Chinese is our mother tongue.

(c) Reasons for "do not mind" which language order to use

4	Once you have understood the text, it does not matter which language or which language order you choose to summarize it.
2	The first summarization must be helpful for the second summarization task!
1	Both orders have advantages and disadvantages, therefore, no difference.
1	I do not have the experience (summarization in English then Chinese, and in Chinese then English), so I don't know how to tell the difference.

Appendix 28: Independent samples *t*-tests on the effects of text presentation mode on summarization performances

Scores	Mean (n)		Levene's Test	*t*-test for equality of means					
	Computer	Paper		t	df	sig.(2-tailed)	Mean difference	Std. error difference	95% confidence interval of the difference
EERSC	45.462 (52)	47.135 (48)	F=0.121, n.s.	-0 697	98	0.487	-1.674	2.4004	-6.4373~3.0896
EPRSC	53.462 (52)	52.375 (48)	F=1.038, n s.	0.479		0.633	1.087	2.2692	-3.4166~5.5896
CERSC	44.260 (52)	40.813 (48)	F=1.983, n s.	1.164		0.247	3.447	2.9611	-2.4290~9.3233
CPRSC	48.279 (52)	45.302 (48)	F=0.093, n.s.	1.186		0.238	2.977	2.5093	-2.0028~7.9563
EEHS/OQS	10.348 (82)	10.620 (75)	F=0.843, n.s.	-0.824	155	0.411	-0.272	0.3308	-0.9258~0.3809
EPHS	11.269 (52)	11.365 (48)	F=1.169, n.s.	-0.238	98	0.812	-0.095	0.4001	-0.8894~0.6987
CEHS/OQS	9.640 (82)	9 420 (75)	F=1.179, n s.	0.578	155	0 564	0.220	0 3807	-0.5319~0 9273
CPHS	10.288 (52)	10.125 (48)	F=0.180, n.s.	0.370	98	0.712	0.163	0.4423	-0.7142~1.0411
E. Length*	313.70 (82)	307.23 (75)	F=0.390, n.s.	0.532	155	0.595	6.47	12.155	-17.542~30.479
C. Length*	530.28 (82)	482.52 (75)	F=0.003, n.s.	2.014	155	0.046	47.76	23.719	0.907~94.614

* If one outlier (ID: 4102) for both English and Chinese summary lengths was excluded in the independent *t*-tests, the significance values for the means differences between computer and paper presentation mode would be 0.352 and 0.0085 for English and Chinese summary lengths respectively.

Appendix 29 Independent samples *t*-tests of the effects of computer familiarity on summarization performances

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std Error Difference	Lower	Upper
EERSC	.374	.544	.716	50	.477	2.412	3.3673	-4.3518	9.1751
EPRSC	.000	.985	.593	50	.556	2 052	3.4583	-4.8939	8 9983
CERSC	2.099	.154	2.587	50	.013	9.302	3 5959	2.0793	16.5244
CPRSC	1.526	.222	2.180	50	.034	7.617	3 4940	.5988	14 6347
EEHS	.732	.395	-1.049	80	.297	-.492	4687	-1 4244	.4411
EPHS	.644	.426	.765	50	.448	.452	.5904	-.7342	1.6374
CEHS	.844	.361	.961	80	.339	.466	.4849	-.4990	1.4310
CPHS	.013	.911	2.293	50	.026	1.363	.5941	.1692	2.5558
E S L	1.519	.221	.342	80	733	5.37	15.708	-25.887	36 633
C S L	2.410	.125	1 209	80	.230	36 63	30 285	-23 642	96 895

Appendix 30: Multivariate statistics of the effects of text presentation mode on summarization performances

(1)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
PRESMODE	0.941	1.464	0.219	0.059	0.439
TXT	0.810	5.457	0.001	0.190	0.970
PRESMODE*TXT	0.936	1.577	0.187	0.064	0.470
↑Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT Box's M=26.717, F=0.822, sig.<0.741					
↓Design Two: Intercept+PRESMODE+LANGORD+PRESMODE*LANGORD Box's M=28.121, F=0.866, sig.<0.676					
PRESMODE	0.949	1.243	0.298	0.051	0.376
LANGORD	0.925	1.886	0.119	0.075	0.551
PRESMODE*LANGORD	0.945	1.359	0.254	0.055	0.409
Dependent variables: EERSC, EPRSC, CERSC and CPRSC					

(2)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
PRESMODE	0.989	0.548	0.580	0.011	0.138
TXT	0.911	4.652	0.012	0.089	0.771
PRESMODE*TXT	1.000	0.022	0.978	0.000	0.053
↑Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT Box's M=6.026, F=0.644, sig.<0.7605					
↓Design Two: Intercept+PRESMODE+LANGORD+PRESMODE*LANGORD Box's M=5.875, F=0.628, sig.<0.7745					
PRESMODE	0.988	0.590	0.585	0.012	0.146
LANGORD	0.969	1.495	0.216	0.031	0.312
PRESMODE*LANGORD	0.996	0.185	0.902	0.004	0.078
Dependent variables: EERSC and EPRSC					

(3)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
PRESMODE	0.966	1.648	0.198	0.034	0.340
TXT	0.821	10.384	0.000	0.179	0.986
PRESMODE*TXT	0.947	2.684	0.073	0.053	0.521
↑Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT Box's M=9.73, F=1.04, sig.<0.405					
Dependent variables: CERSC and CPRSC					

Note: For **Design Two** (*presentation mode* and summarization *language order* as between-subjects factors), some serious violations of the multivariate test assumptions were noticed, as shown in the Box's M test of equality of covariance matrices (Box's M=16.559, F=1.77, sig.<0.0685) and Levene's test of equality of error variances ($F_{3,191}$=3.644, sig.<0.0165) even after excluding the outliers. Therefore, the results of these multivariate tests are not reported.

(4)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
PRESMODE	0.964	0.875	0.482	0.036	0.269
TXT	0.862	3.725	0.007	0.138	0.871
PRESMODE*TXT	0.918	2.072	0.091	0.082	0.597
↑Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT, Box's Test of equality of covariance matrices: Box's M=27.381, F=0.843, sig.<0.711					
↓Design Two: Intercept+PRESMODE+LANGORD+PRESMODE*LANGORD,					

Box's Test of equality of covariance matrices: Box's M=31.874, F=0.981, sig.<0.495					
PRESMODE	0.969	0.747	0.563	0.031	0.232
LANGORD	0.926	1.864	0.123	0.074	0.546
PRESMODE*LANGORD	0.984	0.385	0.819	0.016	0.135
Dependent variables: EEHS, EPHS, CEHS and CPHS					

(5)

Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η^2	Observed power
TXT	31.823	2	15.912	3.844	0.024	0.048	0.690
PRESMODE	2.912	1	2.912	0.704	0.403	0.005	0.133
TXT*PRESMODE	5.313	2	2.656	0.642	0.528	0.008	0.156
↑Design One: Intercept+TXT+PRESMODE+TXT*PRESMODE							
↓Design Two: Intercept+LANGORD+PRESMODE+LANGORD*PRESMODE							
LANGORD	0.480	1	0.480	0.111	0.740	0.001	0.063
PRESMODE	3.076	1	3.076	0.709	0.401	0.005	0.133
LANGORD*PRESMODE	0.116	1	0.116	0.027	0.871	0.000	0.053
Dependent variable: EEHS (of textA, textB and textC summaries)							

(6)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
PRESMODE	0.982	0.851	0.430	0.018	0.192
TXT	0.961	1.933	0.150	0.039	0.392
PRESMODE*TXT	0.999	0.030	0.970	0.001	0.054
↑Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT Box's M=9.850, F=1.053, sig.<0.395					
↓Design Two: Intercept+PRESMODE+LANGORD+PRESMODE*LANGORD Box's M=8.123, F=0.868, sig.<0.553					
PRESMODE	0.981	0.924	0.401	0.019	0.206
LANGORD	0.963	1.828	0.166	0.037	0.373
PRESMODE*LANGORD	0.998	0.099	0.906	0.002	0.065
Dependent variables: EEHS and EPHS					

(7)

Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η^2	Observed power
TXT	106.590	2	53.295	10.788	0.000	0.125	0.989
PRESMODE	4.124	1	4.124	0.835	0.362	0.005	0.148
TXT*PRESMODE	33.293	2	16.646	3.370	0.037	0.043	0.628
↑Design One: Intercept+TXT+PRESMODE+TXT*PRESMODE							
↓Design Two: Intercept+LANGORD+PRESMODE+LANGORD*PRESMODE							
LANGORD	6.882	1	6.882	1.206	0.274	0.008	0.194
PRESMODE	2.284	1	2.284	0.400	0.528	0.003	0.096
LANGORD*PRESMODE	4.895E-02	1	4.895E-02	0.009	0.926	0.000	0.051
Dependent variable: CEHS							

(8)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power[#]
PRESMODE	0.995	0.241	0.786	0.005	0.087
TXT	0.869	7.142	0.001	0.131	0.925
PRESMODE*TXT	0.940	3.036	0.053	0.060	0.575

↑Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT, Box's Test of equality of covariance matrices: Box's M=7.616, F=0.814, sig.<0.603.					
↓Design Two: Intercept+PRESMODE+LANGORD+PRESMODE*LANGORD, Box's Test of equality of covariance matrices: Box's M= 7.106, F=0.758, sig.<0.655. Hypothesis df=2, error df=91 (4 univariate outliers: ID3310, 3312, 4114, 3406 were excluded)					
PRESMODE	0.998	0.102	0.903	0.002	0.065
LANGORD	0.952	2.307	0.105	0.048	0.457
PRESMODE*LANGORD	0.980	0.928	0.399	0.020	0.206
Dependent variables: CEHS and CPHS					

(9)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
PRESMODE	0.927	5.684	0.004	0.073	0.857
TXT	0.890	4.332	0.002	0.057	0.930
PRESMODE*TXT	0.962	1.421	0.227	0.019	0.440
↑Design One: Intercept+PRESMODE+TXT+PRESMODE*TXT, Box's Test of equality of covariance matrices: Box's M=23.06, F=1.483, sig.<0.102					
↓Design Two: Intercept+PRESMODE+LANGORD+PRESMODE*LANGORD, Box's Test of equality of covariance matrices: Box's M=13.192, F=1.429, sig.<0.169					
PRESMODE	0.938	4.863	0.009	0.062	0.795
LANGORD	0.904	7.738	0.001	0.096	0.946
PRESMODE*LANGORD	0.998	0.112	0.894	0.002	0.067
Dependent variables: E.S.L and C.S.L					

Appendix 31: Multivariate statistics of the effects of computer familiarity on summarization performances

(1)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM LEVEL	0.883	1.493	0.220	0.117	0.425
TXT	0.908	1.143	0.348	0.092	0.330
COMPFAM*TXT	0.975	0.290	0.883	0.025	0.108
↑Design A: Intercept+COMPFAMLEVEL+TXT+COMPFAMLEVEL*TXT Box's M=42.993, F=1.162, sig.<0.249					
↓Design B: Intercept+COMPFAMLEVEL+LANGORD+COMPFAMLEVEL*LANGORD Box's M=25.16, F=0.701, sig.<0.886					
COMPFAM LEVEL	0.852	1.956	0.118	0.148	0.543
LANGORD	0.945	0.653	0.628	0.055	0.197
COMPFAM*LANGORD	0.839	2.166	0.088	0.161	0.593
Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM score	0.928	0.895	0.475	0.072	0.262
TXT	0.912	1.113	0.362	0.088	0.322
↑Design C: Intercept+COMPFAM score+TXT Box's Test of Equality of covariance matrices: M=11.945, F=1.091, sig.<0.365					
↓Design D: Intercept+COMPFAM score+LANGORD Box's Test of Equality of covariance matrices: M=6.876, F=0.628, sig.<0.791					
COMPFAM score	0.887	1.466	0.228	0.113	0.419
LANGORD	0.930	0.860	0.495	0.070	0.253
Dependent variables: EERSC, EPRSC, CERSC and CPRSC					

(2)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM LEVEL	0.998	0.044	0.957	0.002	0.056
TXT	?	?	?	?	?
COMPFAM*TXT	0.996	0.101	0.904	0.004	0.064
↑Design A: Intercept+COMPFAMLEVEL+TXT+COMPFAMLEVEL*TXT Box's M=8.471, F=0.856, sig.<0.5645					
↓Design B: Intercept+COMPFAMLEVEL+LANGORD+COMPFAMLEVEL*LANGORD Box's M=5.730, F=0.586, sig.<0.810					
COMPFAM LEVEL	0.987	0.301	0.742	0.013	0.095
LANGORD	0.967	0.791	0.459	0.033	0.177
COMPFAM*LANGORD	0.859	3.853	0.028	0.141	0.670
Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM score	0.978	0.546	0.583	0.022	0.135
TXT	0.942	1.485	0.237	0.058	0.301
↑Design C: Intercept+COMPFAM score+TXT, Box's M=2.362, F=0.753, sig.<0.52					
↓Design D: Intercept+COMPFAM score+LANGORD, Box's M=0.257, F=0.082, sig.<0.97					
COMPFAM score	0.953	1.175	0.318	0.047	0.245
LANGORD	0.959	1.018	0.369	0.041	0.217
Dependent variables: EERSC and EPRSC (RSC of English summaries)					

(3)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM LEVEL	0.901	2.752	0.087	0.099	0.489

TXT	0.960	0.972	0.386	0.040	0.209
COMPFAM*TXT	0.979	0.508	0.605	0.021	0.129
↑Design A: Intercept+COMPFAMLEVEL+TXT+COMPFAMLEVEL*TXT Box's M=16.954, F=1.713, sig.<0.085					
↓Design B: Intercept+COMPFAMLEVEL+LANGORD+COMPFAMLEVEL*LANGORD Box's M=12.23, F=1.25, sig.<0.2595					
COMPFAM LEVEL	0.855	3.971	0.025	0.145	0.684
LANGORD	0.994	0.142	0.868	0.006	0.071
COMPFAM*LANGORD	0.916	2.161	0.127	0.084	0.420
Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM score	0.942	1.466	0.241	0.058	0.298
TXT	0.961	0.966	0.388	0.039	0.208
↑Design C: Intercept+COMPFAM score+TXT, Box's M=4.232, F=1.35, sig.<0.256					
↓Design D: Intercept+COMPFAM score+LANGORD, Box's M=5.187, F=1.654, sig.<0.175					
COMPFAM score	0.893	2.867	0.067	0.107	0.536
LANGORD	0.983	0.420	0.659	0.017	0.114
Dependent variables: CERSC and CPRSC					

(4)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power#
COMPFAM LEVEL	0.798	2.842	0.035	0.202	0.728
TXT	0.968	0.367	0.831	0.032	0.126
COMPFAM LEVEL*TXT	0.965	0.405	0.804	0.035	0.135
↑Design A: Intercept+COMPFAM LEVEL+TXT+COMPFAM LEVEL*TXT, Box's Test of equality of covariance matrices: Box's M=40.76, F=1.102, sig.<0.322					
↓Design B: Intercept+COMPFAM LEVEL+LANGORD+COMPFAM LEVEL*LANGORD, Box's Test of equality of covariance matrices: Box's M=26.573, F=0.740, sig.<0.846					
COMPFAM LEVEL	0.780	3.174	0.022	0.220	0.780
LANGORD	0.970	0.347	0.844	0.030	0.121
COMPFAM LEVEL*LANGORD	0.863	1.782	0.149	0.137	0.500
↓Design C: Intercept+COMPFAM score+TXT Box's Test of equality of covariance matrices: Box's M=10.582, F=0.966, sig.<0.471					
COMPFAM score	0.881	1.561	0.201	0.119	0.444
TXT	0.978	0.260	0.902	0.022	0.102
↓Design D: Intercept+COMPFAM score+LANGORD, Box's Test of equality of covariance matrices: Box's M=6.472, F=0.591, sig.<0.823					
COMPFAM score	0.851	2.014	0.108	0.149	0.559
LANGORD	0.971	0.343	0.848	0.029	0.121
Dependent variables: EEHS, EPHS, CEHS and CPHS					

(5)

Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η^2	Observed power
TXT	21.813	2	10.907	2.579	0.082	0.064	0.500
COMPFAM LEVEL	1.278	1	1.278	0.302	0.584	0.004	0.084
TXT*COMPFAM LEVEL	8.578	2	4.293	1.015	0.367	0.026	0.221
↑Design A: Intercept+TXT+COMPFAM LEVEL+TXT*COMPFAM LEVEL							
↓Design B: Intercept+LANGORD+COMPFAM LEVEL+LANGORD*COMPFAM LEVEL							
Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η^2	Observed power
LANGORD	0.436	1	0.436	0.099	0.753	0.001	0.061
COMPFAM LEVEL	3.417	1	3.417	0.778	0.380	0.010	0.140

LANGORD*COMPFAM LEVEL	16.132	1	16.132	3.657	0.059	0.045	0.473
Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η^2	Observed power
TXT	32.345	2	16.172	3.803	0.027	0.089	0.676
COMPFAM score	0.112	1	0.112	0.026	0.872	0.000	0.053
↑Design C: Intercept+TXT+COMPFAM score							
↓Design D: Intercept+LANGORD+COMPFAM score							
LANGORD	0.618	1	0.618	0.134	0.715	0.002	0.065
COMPFAM score	0.375	1	0.375	0.081	0.776	0.001	0.059
Dependent variable: EEHS							

(6)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM LEVEL	0.992	0.195	0.824	0.008	0.079
TXT	0.977	0.564	0.572	0.023	0.138
COMPFAM LEVEL*TXT	0.973	0.660	0.522	0.027	0.154
↑Design A: Intercept+COMPFAM LEVEL+TXT+COMPFAM LEVEL*TXT Box's Test of equality of covariance matrices: M=8.180, F=0.827, sig.<0.592.					
↓Design B: Intercept+COMPFAM LEVEL+LANGORD+COMPFAM LEVEL*LANGORD Box's Test of equality of covariance matrices: M=9.359, F=0.957, sig.<0.474					
COMPFAM LEVEL	0.990	0.228	0.797	0.010	0.084
LANGORD	0.984	0.387	0.681	0.016	0.109
COMPFAM LEVEL*LANGORD	0.867	3.603	0.035	0.133	0.639
Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM score	0.968	0.793	0.458	0.032	0.178
TXT	0.981	0.457	0.636	0.019	0.120
↑Design C: Intercept+COMPFAM score+TXT Box's Test of Equality of covariance matrices: M=0.918, F=0.293, sig.<0.831					
↓Design D: Intercept+COMPFAM score+LANGORD Box's Test of Equality of covariance matrices: M=0.808, F=0.258, sig.<0.856					
COMPFAM score	0.951	1.232	0.301	0.049	0.256
LANGORD	0.982	0.435	0.650	0.018	0.117
Dependent variables: EEHS and EPHS					

(7)

Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η^2	Observed power
TXT	14.959	2	7.478	1.783	0.175	0.045	0.362
COMPFAM LEVEL	11.545	1	11.545	2.753	0.101	0.035	0.374
TXT*COMPFAM LEVEL	32.896	2	16.448	3.922	0.024	0.094	0.690
↑Design A: Intercept+TXT+COMPFAM LEVEL+TXT*COMPFAM LEVEL							
↓Design B: Intercept+LANGORD+COMPFAM LEVEL+LANGORD*COMPFAM LEVEL							
LANGORD	4.105	1	4.105	0.849	0.360	0.011	0.149
COMPFAM LEVEL	5.972	1	5.972	1.235	0.270	0.016	0.195
LANGORD*COMPFAM LEVEL	3.759	1	3.759	0.777	0.381	0.010	0.140
Source	Type III sum of squares	df	Mean square	F	Sig.	Partial η^2	Observed power
TXT	34.016	2	17.008	3.784	0.027	0.088	0.674

COMPFAM score	12.582	1	12.582	2.799	0.098	0.035	0.379
↑Design C: Intercept+TXT+COMPFAM score							
↓Design D: Intercept+LANGORD+COMPFAM score							
LANGORD	3.557	1	3.557	0.737	0.393	0.009	0.136
COMPFAM score	5.091	1	5.091	1.056	0.307	0.013	0.174
Dependent variable: CEHS							

(8)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM LEVEL	0.831	4.774	0.013	0.169	0.768
TXT	0.991	0.212	0.810	0.009	0.081
COMPFAM LEVEL*TXT	1.000	0.008	0.992	0.000	0.051
↑Design A: Intercept+COMPFAM LEVEL+TXT+COMPFAM LEVEL*TXT Box's Test of equality of covariance matrices: M=15.231, F=1.539, sig.<0.128.					
↓Design B: Intercept+COMPFAM LEVEL+LANGORD+COMPFAM LEVEL*LANGORD Box's Test of equality of covariance matrices: M=6.824, F=0.698, sig.<0.712					
COMPFAM LEVEL	0.802	5.811	0.006	0.198	0.849
LANGORD	0.990	0.227	0.798	0.010	0.083
PRESMODE*LANGORD	0.985	0.357	0.701	0.015	0.104
COMPFAM score	0.896	2.784	0.072	0.104	0.523
TXT	0.997	0.068	0.934	0.003	0.060
↑Design C: Intercept+COMPFAM score+TXT Box's Test of Equality of covariance matrices: M=4.454, F=1.420, sig.<0.235					
↓Design D: Intercept+COMPFAM score+LANGORD Box's Test of Equality of covariance matrices: M=3.907, F=1.246, sig.<0.291					
COMPFAM score	0.866	3.702	0.032	0.134	0.652
LANGORD	0.988	0.303	0.740	0.012	0.095
Dependent variables: CEHS and CPHS					

(9)

Effect	Wilks' Λ value	F	Sig.	Partial η^2	Observed power
COMPFAM LEVEL	0.992	0.303	0.739	0.008	0.096
TXT	0.945	1.043	0.387	0.028	0.323
COMPFAM LEVEL*TXT	0.972	0.515	0.725	0.014	0.171
↑Design A: Intercept+COMPFAM LEVEL+TXT+COMPFAM LEVEL*TXT, Box's Test of equality of covariance matrices: Box's M=23.203, F=1.403, sig.<0.136					
↓Design B: Intercept+COMPFAM LEVEL+LANGORD+COMPFAM LEVEL*LANGORD, Box's Test of equality of covariance matrices: Box's M=5.185, F=0.548, sig.<0.840					
COMPFAM LEVEL	0.986	0.528	0.592	0.014	0.134
LANGORD	0.894	4.452	0.015	0.106	0.748
COMPFAM LEVEL*LANGORD	0.975	0.961	0.387	0.025	0.211
COMPFAM score	0.999	0.030	0.970	0.001	0.054
TXT	0.958	0.818	0.516	0.021	0.257
↑Design C: Intercept+COMPFAM score+TXT Box's Test of equality of covariance matrices: Box's M=9.68, F=1.552 sig.<0.157					
↓Design D: Intercept+COMPFAM score+LANGORD, Box's Test of equality of covariance matrices: Box's M=0.179, F=0.058, sig.<0.982					
COMPFAM score	0.985	0.596	0.554	0.015	0.146
LANGORD	0.891	4.631	0.013	0.109	0.766

Appendix 32: Statistics of the effects of text type on summarization performances (RSC of textA and textC summaries)

<table>
<tr><td>Model*</td><td></td><td colspan="3">Dependent variables</td></tr>
<tr><td></td><td></td><td>(1) EERSC, EPRSC, CERSC & CPRSC</td><td>(2) EERSC & EPRSC</td><td>(3) CERSC & CPRSC</td></tr>
<tr><td>I</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.823</td><td>5.102</td><td>0.001</td><td>0.177</td><td>0.959</td></tr>
</table>
Box's M=11.826, F=1.130, sig.<0.3355</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.909</td><td>4.838</td><td>0.010</td><td>0.091</td><td>0.789</td></tr>
</table>
Box's M=1.944, F=0.633, sig.<0.5935</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.841</td><td>9.199</td><td>0.000</td><td>0.159</td><td>0.973</td></tr>
</table>
Box's M=5.278, F=1.720, sig.<0.160</td></tr>
<tr><td></td><td>U</td><td colspan="3">EERSC (mean difference=7.104, F=9.545, sig.<0.0035, partial η²=0.089, observed power=0.864)
EPRSC (mean difference=4.002, F=3.197, sig.<0.075 n.s., partial η²=0.032, observed power=0.425)
CERSC (mean difference=10.294, F=13.542, sig.<0.0005, partial η²=0.121, observed power=0.954)
CPRSC (mean difference=9.472, F=16 360, sig.<0.0005, partial η²=0.143, observed power=0.980)</td></tr>
<tr><td>II</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.810</td><td>5.457</td><td>0.001</td><td>0.190</td><td>0.970</td></tr>
</table>
Box's M=26.717, F=0.822, sig.<0.7415</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.911</td><td>4.652</td><td>0.012</td><td>0.089</td><td>0.771</td></tr>
</table>
Box's M=6.026, F=0.644, sig.<0.765</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0 821</td><td>10.384</td><td>0.000</td><td>0.179</td><td>0.986</td></tr>
</table>
Box's M=9.730, F=1.04, sig.<0.405</td></tr>
<tr><td></td><td>U</td><td colspan="3">EERSC (mean difference=7.036, F=9.142, sig.<0.0035, partial η²=0.087, observed power=0.849)
EPRSC (mean difference=4.071, F=3.233, sig.<0.075 n s., partial η²=0.033, observed power=0.429)
CERSC (mean difference=10.853, F=15.812, sig.<0 0005, partial η²=0.141, observed power=0.976)
CPRSC (mean difference=9.823, F=17.785, sig.<0.0005, partial η²=0.156, observed power=0.987)
TXT*PRESMODE interactive effect on CERSC in (1), (3) [F=5 320, sig.<0.0235, partial η²=0 053, observed power=0 627, textA (computer=46.538, paper=48.315, textC (computer=41.981, paper=31.167)]</td></tr>
<tr><td>III</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ</th><th>F</th><th>Sig.</th><th>Partial</th><th>Observed</th></tr>
</table></td><td>
<table>
<tr><th>Wilks'Λ</th><th>F</th><th>Sig.</th><th>Partial</th><th>Observed</th></tr>
</table></td><td>
<table>
<tr><th>Wilks'Λ</th><th>F</th><th>Sig.</th><th>Partial</th><th>Observed</th></tr>
</table></td></tr>
</table>

		0.823	4.995	0.001	0.177	0.955	0.907	4.859	0.010	0.093	0.790	0.839	9.084	0.000	0.161	0.972
		Box's M=30.323, F=0.934, sig.<0.5705					Box's M=2.765, F=0.296, sig.<0.976					Box's M=15.994, F=1.71, sig.<0.081				
	U	EERSC (mean difference=7.157, F=9.509, sig.<0.0035, partial η^2=0.09, observed power=0.863) EPRSC (mean difference=4.139, F=3.468, sig.<0.065 n.s, partial η^2=0.035/0.034, observed power=0.454/0.442) CERSC (mean difference=10.099, F=13.122, sig.<0.0005, partial η^2=0.120, observed power=0.948) CPRSC (mean difference=9.358, F=16.333, sig.<0.0005, partial η^2=0.145, observed power=0.979) **TXT*LANGORD interactive effect on CPRSC in (1), (3)** [F=4.206, sig.<0.0435, partial η^2=0.042, observed power=0.528]														
IV	M	Wilks'Λ value	F	Sig.	Partial η^2	Observed power	Wilks'Λ value	F	Sig.	Partial η^2	Observed power	Wilks'Λ value	F	Sig.	Partial η^2	Observed power
		0.865	3.668	0.008	0.135	0.865	0.943	2.898	0.060	0.057	0.554	0.880	6.537	0.002	0.120	0.900
		Box's M=11.826, F=1.13, sig.<0.3355					0.930	3.624	0.030	0.070	0.657	Box's M=5.278, F=1.72, sig.<0.160				
							Box's M=1.944, F=0.633, sig.<0.593 Note: effects of text type was at a borderline of significance, p<0.05991758.									
	U	EERSC (mean difference=5.790, F=5.851, sig.<0.0175, partial η^2=0.057, observed power=0.668) EPRSC (mean difference=2.154, F=0.886, sig.<0.349 n.s., partial η^2=0 009, observed power=0.154) CERSC (mean difference=8.783, F=9.067, sig.<0.0035, partial η^2=0.085, observed power=0 846) CPRSC (mean difference=8.478, F=11.924, sig.<0.0015, partial η^2=0.109, observed power=0.928) **TOEFL effect on EPRSC in (1), (2)** [F=6.746, sig.<0.0115, partial η^2=0.065, observed power=0.730]														
V	M	Wilks'Λ value	F	Sig.	Partial η^2	Observed power	Wilks'Λ value	F	Sig.	Partial η^2	Observed power	Wilks'Λ value	F	Sig.	Partial η^2	Observed power
		0.854	3.929	0.005	0 146	0.890	0.946	2.659	0.075	0.054	0.516	0.864	7.414	0.001	0.136	0.934
		Box's M=26.717, F=0 822, sig.<0.741					0.917	4.265	0.017	0.083	0.732	Box's M=9.73, F=1.04, sig <0.405				
		The main effect of TOEFL Reading was approaching significance					Box's M=6 026, F=0 644, sig.<0.760									

<table>
<tr><td></td><td></td><td colspan="3">level (F=2.264, sig.<0.0685, partial η²=0.09).</td><td colspan="3">Note: effect of text type was not significant.</td><td></td></tr>
<tr><td></td><td>U</td><td colspan="3"></td><td colspan="3"></td><td></td></tr>
<tr><td></td><td></td><td colspan="3">EERSC (mean difference=5.639, F=5.348, sig.<0.0235,</td><td colspan="3">partial η²=0.053, observed power=0.629)</td><td></td></tr>
<tr><td></td><td></td><td colspan="3">EPRSC (mean difference=1.883, F=0.664, sig.<0.417 n.s.,</td><td colspan="3">partial η²=0.007, observed power=0.127)</td><td></td></tr>
<tr><td></td><td></td><td colspan="3">CERSC (mean difference=9.584, F=11.085, sig.<0.0015,</td><td colspan="3">partial η²=0.104, observed power=0.909)</td><td></td></tr>
<tr><td></td><td></td><td colspan="3">CPRSC (mean difference=8.786, F=12.772, sig.<0.0015,</td><td colspan="3">partial η²=0.119, observed power=0.943)</td><td></td></tr>
<tr><td></td><td></td><td colspan="7">**TOEFL effect on EPRSC in (1), (2)** [F=8.267, sig.<0.0055, partial η²=0.08, observed power=0.812]</td></tr>
<tr><td>VI</td><td>M</td><td colspan="3"><table><tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr><tr><td>0.864</td><td>3.609</td><td>0.009</td><td>0.136</td><td>0.858</td></tr></table>Box's M=30.323, F=0.934, sig.<0 570</td><td colspan="3"><table><tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr><tr><td>0.942</td><td>2.887</td><td>0.061</td><td>0.058</td><td>0.552</td></tr><tr><td>0.928</td><td>3.642</td><td>0.030</td><td>0.072</td><td>0 659</td></tr></table>Box's M=2.765, F=0.296, sig.<0.976
Note: effect of text type was not significant.</td><td colspan="2"><table><tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr><tr><td>0.878</td><td>6.054</td><td>0.002</td><td>0.122</td><td>0.899</td></tr></table>Box's M=15.994, F=1.71, sig.<0.081</td></tr>
<tr><td></td><td>U</td><td colspan="3"></td><td colspan="3"></td><td></td></tr>
<tr><td></td><td></td><td colspan="3">EERSC (mean difference=5.833, F=5.834, sig.<0.0185,</td><td colspan="3">partial η²=0.058, observed power=0.667)</td><td></td></tr>
<tr><td></td><td></td><td colspan="3">EPRSC (mean difference=2.303, F=1.029, sig.<0.313 n.s.,</td><td colspan="3">partial η²=0.011, observed power=0.171)</td><td></td></tr>
<tr><td></td><td></td><td colspan="3">CERSC (mean difference=8 616, F=8.785, sig.<0.0045,</td><td colspan="3">partial η²=0.085, observed power=0.835)</td><td></td></tr>
<tr><td></td><td></td><td colspan="3">CPRSC (mean difference=8.417, F=12.011, sig.<0.0015,</td><td colspan="3">partial η²=0.112, observed power=0.929)</td><td></td></tr>
<tr><td></td><td></td><td colspan="7">**TOEFL effect on EPRSC in (1), (2)** [F=6.797, sig.<0.0115, partial η²=0.065, observed power=0.733]</td></tr>
<tr><td></td><td></td><td colspan="7">**TXT*LANGORD interactive effect on CPRSC in (1), (3)** [F=4.056, sig.<0 0475, partial η²=0.041, observed power=0 513]</td></tr>
</table>

Note: 1. Models I-VI, see Figure 10.1; M=multivariate tests, U=univariate tests **2**. Effects from sources other than *text type* in the models, either as main or interactive effects, are not reported unless they're significant on RSC scores. All effects are from text type, otherwise stated clearly.

Appendix 33: Statistics of the effects of text type on summarization performances (HS)

<table>
<tr><td>Model*</td><td></td><td colspan="3">Dependent variable(s)</td></tr>
<tr><td></td><td></td><td>(1) EEHS, EPHS, CEHS & CPHS</td><td>(2) EEHS & EPHS</td><td>(3) CEHS & CPHS</td></tr>
<tr><td>I</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.875</td><td>3.378</td><td>0.012</td><td>0.125</td><td>0.832</td></tr>
</table>
Box's M=12.376, F=1.183, sig.<0.2975</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.959</td><td>2.095</td><td>0.129</td><td>0.041</td><td>0.421</td></tr>
</table>
Box's M=3.938, F=1.284, sig.<0.2785</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.883</td><td>6.426</td><td>0.002</td><td>0.117</td><td>0.895</td></tr>
</table>
Box's M=3.297, F=1.074, sig.<0.358</td></tr>
<tr><td></td><td>U</td><td colspan="3">EEHS (mean difference=0.748, F=3.271, sig.<0.0745, n.s., partial η²=0.032, observed power=0.443)
EPHS (mean difference=0.614, F=2.410, sig.<0.124, n.s., partial η²=0.024, observed power=0.336)
CEHS (mean difference=1.662, F=12.091, sig.<0.0015, partial η²=0.110, observed power=0.931)
CPHS (mean difference=1.139, F=7.087, sig.<0.0095, partial η²=0.067, observed power=0.750)</td></tr>
<tr><td>II</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.862</td><td>3.725</td><td>0.007</td><td>0.138</td><td>0.871</td></tr>
</table>
Box's M=27.381, F=0.843, sig.<0.7115</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.961</td><td>1.933</td><td>0.150</td><td>0.039</td><td>0.392</td></tr>
</table>
Box's M=9.850, F=1.053, sig.<0.395</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.869</td><td>7.142</td><td>0.001</td><td>0.131</td><td>0.925</td></tr>
<tr><td>0.940</td><td>3.036</td><td>0.053</td><td>0.060</td><td>0.575</td></tr>
</table>
Box's M=7.616, F=0.814, sig.<0.603
Note: the TXT*PRESMODE interactive effect was approaching significance level (sig.<0.0535)</td></tr>
<tr><td></td><td>U</td><td colspan="3">EEHS (mean difference=0.712, F=2.931, sig.<0.0905, n.s., partial η²=0.030, observed power=0.396)
EPHS (mean difference=0.613, F=2.332, sig.<0.1305, n.s., partial η²=0.024, observed power=0.327)
CEHS (mean difference=1.730, F=13.586, sig.<0.0005, partial η²=0.124, observed power=0.954)
CPHS (mean difference=1.175, F=7.447, sig.<0.0085, partial η²=0.072, observed power=0.771)
TXT*PRESMODE interactive effect on CEHS in (1), (3) [F=6.037, sig.<0.0165, partial η²=0.059, observed power=0.682, textA (computer=9.538, paper=10.407, textC (computer=8.962, paper=7.524)]</td></tr>
<tr><td>III</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.875</td><td>3.309</td><td>0.014</td><td>0.125</td><td>0.823</td></tr>
</table>
Box's M=31.678, F=0.975, sig.<0.5045</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.956</td><td>2.163</td><td>0.121</td><td>0.044</td><td>0.433</td></tr>
</table>
Box's M=6.376, F=0.682, sig.<0.726</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.883</td><td>6.275</td><td>0.003</td><td>0.117</td><td>0.887</td></tr>
<tr><td>0.929</td><td>3.647</td><td>0.030</td><td>0.071</td><td>0.660</td></tr>
</table>
Box's M=10.839, F=1.159, sig.<0.317
Note: the TXT*LANGORD interactive effect (sig.<0.0305).</td></tr>
<tr><td></td><td>U</td><td colspan="3">EEHS (mean difference=0.754, F=3.255, sig.<0.0745, n.s., partial η²=0.033, observed power=0.431)
EPHS (mean difference=0.639, F=2.661, sig.<0.1065 n.s., partial η²=0.027, observed power=0.365)
CEHS (mean difference=1.631, F=11.684, sig.<0.0015, partial η²=0.109, observed power=0.923)</td></tr>
</table>

<table>
<tr><td></td><td></td><td>CPHS (mean difference=1.126, F=7.271, sig.<0.0085, partial η²=0.070, observed power=0.761)
LANGORD main effect on EPHS in (1) approaching significance level [F=3.804, sig.<0.0545, partial η²=0.038, observed power=0.488]
TXT*LANGORD interactive effect on CPHS in (1), (3) [F=6.878, sig.<0.0105, partial η²=0.067, observed power=0.738; textA: E/C=10.26, C/E=11.179, textC: E/C=10.229, C/E=8.957]</td></tr>
<tr><td>IV</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.915</td><td>2.173</td><td>0.078</td><td>0.085</td><td>0.621</td></tr>
</table>
Box's M=12.376, F=1.183, sig.<0.2975
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.978</td><td>1.090</td><td>0.340</td><td>0.022</td><td>0.236</td></tr>
</table>
Box's M=3.938, F=1.284, sig.<0.278
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.920</td><td>4.197</td><td>0.018</td><td>0.080</td><td>0.725</td></tr>
</table>
Box's M=3.297, F=1.074, sig.<0.358</td></tr>
<tr><td></td><td>U</td><td>EEHS (mean difference=0.569, F=1.721, sig.<0.1935 n.s., partial η²=0.017, observed power=0.255)
EPHS (mean difference=0.456, F=1.207, sig.<0.2755 n.s., partial η²=0.012, observed power=0.193)
CEHS (mean difference=1.408, F=7.976, sig.<0.0065, partial η²=0.076, observed power=0.798)
CPHS (mean difference=0.934, F=4.359, sig.<0.0395, partial η²=0.043, observed power=0.543)</td></tr>
<tr><td>V</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.899</td><td>2.595</td><td>0.0415</td><td>0.101</td><td>0.709</td></tr>
</table>
Box's M=27.381, F=0.843, sig.<0.711
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.980</td><td>0.942</td><td>0.394</td><td>0.020</td><td>0.209</td></tr>
</table>
Box's M=9.850, F=1.053, sig.<0.395
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.903</td><td>5.022</td><td>0.008</td><td>0.097</td><td>0.804</td></tr>
</table>
Box's M=7.616, F=0.814, sig.<0.603</td></tr>
<tr><td></td><td>U</td><td>EEHS (mean difference=0.529, F=1.455, sig.<0.2315, partial η²=0.015, observed power=0.223)
EPHS (mean difference=0.441, F=1.084, sig.<0.3005 n.s., partial η²=0.011, observed power=0.178)
CEHS (mean difference=1.553, F=9.775, sig.<0.0025, partial η²=0.093, observed power=0.872)
CPHS (mean difference=0.978, F=4.633, sig.<0.0345, partial η²=0.047, observed power=0.568)
PRESMODE*TXT interactive effect on CEHS in (1), (3) [F=4.188, sig.<0.0435, partial η²=0.042, observed power=0.526; textA: computer=9.568, paper=10.217, textC: computer=9.017, paper=7.663]</td></tr>
<tr><td>VI</td><td>M</td><td>
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.915</td><td>2.138</td><td>0.082</td><td>0.085</td><td>0.612</td></tr>
</table>
Box's M=31.678, F=0.975, sig.<0.504
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.976</td><td>1.142</td><td>0.324</td><td>0.024</td><td>0.246</td></tr>
</table>
Box's M=6.376, F=0.682, sig.<0.726
<table>
<tr><th>Wilks'Λ value</th><th>F</th><th>Sig.</th><th>Partial η²</th><th>Observed power</th></tr>
<tr><td>0.920</td><td>4.112</td><td>0.019</td><td>0.080</td><td>0.716</td></tr>
<tr><td>0.930</td><td>3.553</td><td>0.033</td><td>0.070</td><td>0.647</td></tr>
</table>
Box's M=10.839, F=1.159, sig.<0.317
Note: TXT*LANGORD interactive effect (sig.<0.0335)</td></tr>
<tr><td></td><td>U</td><td>EEHS (mean difference=0.573, F=1.712, sig.<0.1945 n.s., partial η²=0.018, observed power=0.254)
EPHS (mean difference=0.484, F=1.384, sig.<0.2425 n.s., partial η²=0.014, observed power=0.214)
CEHS (mean difference=1.382, F=7.709, sig.<0.0075, partial η²=0.075, observed power=0.785)
CPHS (mean difference=0.934, F=4.563, sig.<0.0355, partial η²=0.046, observed power=0.561)
LANGORD effect on EPHS in (1), (2) [F=3.825, sig.<0.0535, approaching significance level, partial η²=0.039, observed power=0.490]
TXT*LANGORD interactive effect on CPHS in (1), (3) [F=6.696, sig.<0.0115, partial η²=0.066, observed power=0.726; textA: E/C=10.18, C/E=11.078, textC: E/C=10.322, C/E=9.069]</td></tr>
</table>

Note: 1. Models A-F, see Figure X; M=multivariate tests, U=univariate tests **2.** Effects from sources other than *text type* in the models, either as main or interactive effects, are not reported unless they're significant on HS scores. All effects are from text type, otherwise stated clearly

Appendix 34: Statistics of the effects of text type on EEHS and CEHS and pairwise comparisons

Model*	Source	Dependent variable							
		EEHS				CEHS			
		F	Sig.	Partial η^2	Observed power	F	Sig.	Partial η^2	Observed power
II	TXT	3.844	0.024	0.048	0.690	10.788	0.000	0.125	0.989
	PRESMODE	0.704	0.403	0.005	0.133	0.835	0.362	0 005	0.148
	TXT*PRESMODE	0.642	0.528	0.008	0.156	3.370	0.037	0.043	0.628
	▲	▲textB/textC=1.109, sig.<0.0205				textA/textC=1.730, sig.<0.0005 textB/textC=1.869, sig.<0.0005			
III	TXT	4.075	0.019	0.051	0.717	9.631	0.000	0.113	0.980
	LANGORD	0.024	0.877	0.000	0.053	1.486	0.225	0.010	0.228
	TXT*LANGORD	0.080	0.923	0.001	0.062	0.677	0.510	0.009	0.162
		textB/textC=1.147, sig.<0.0165				textA/textC=1.631, sig.<0.0015 textB/textC=1.807, sig.<0.0005			
IV	TXT	4.376	0.014	0.054	0.749	8.305	0.000	0.099	0.960
	RDGTOEFL	2.319	0.130	0.015	0.328	6.817	0.010	0.043	0.737
		textB/textC=1.176, sig.<0.0115				textA/textC=1.370, sig.<0.0105 textB/textC=1.724, sig.<0.0005			
V	TXT	4.052	0.019	0.052	0.714	9.196	0.000	0.110	0.975
	RDGTOEFL	1.742	0.189	0.012	0.259	4.425	0.037	0.029	0.552
	PRESMODE	0.793	0.375	0.005	0.143	0.361	0.549	0.002	0.092
	TXT*PRESMODE	0.414	0.662	0.006	0.116	2.361	0.098	0.031	0.472
		textB/textC=1.140, sig.<0.0155				textA/textC=1.468, sig.<0.0055 textB/textC=1.801, sig.<0.0005			
VI	TXT	4.267	0.016	0.054	0.738	8.032	0.000	0.097	0.954
	RDGTOEFL	2.272	0.134	0.015	0.322	6.713	0.011	0.043	0.730
	LANGORD	0.007	0.935	0.000	0.051	1.130	0.289	0.008	0.184
	TXT*LANGORD	0.118	0.889	0.002	0.068	0.855	0.427	0.011	0.195
		textB/textC=1.182, sig.<0.0125				textA/textC=1.340, sig.<0.0125 textB/textC=1.712, sig.<0.0005			

Notes ▲: Pairwise comparisons of the estimated marginal mean differences. In both EEHS and CEHS, it is always textA<textB, textA>textC, textB>textC (i.e., textB>textA>textC).
*Design I was already analysed in the univariate analyses of variances (see Table 10.1)

Appendix 35: Statistics of the effects of text type on the lengths of summaries

Model		Effect	Wilks'Λ	F	Sig.	Partial η^2	Observed power
I	M	TXT	0.903	3.840	0.005	0.050	0.894
		Box's M=8.598, F=1.405, sig.<0.208					
	U	Significant effects on *Chinese* summary length only (F=4.974, sig.<0.0085, partial η^2=0.063) Pairwise comparisons: textA/textB (mean difference=74.502, sig.<0.0105) textA/textC (mean difference=62.655, sig.<0.0525, approaching significance) textB/textC (mean difference=-11.847, n.s.)					
II	M	TXT	0.890	4.332	0.002	0.057	0.930
		PRESMODE	0.927	5.684	0.004	0.073	0.857
		TXT*PRESMODE	0.962	1.421	0.227	0.019	0.440
		Box's M=23.060, F=1.483, sig.<0.102					
	U	Significant effects on *Chinese* summary length only (F=5.985, sig.<0.0035, partial η^2=0.076) Pairwise comparisons: textA/textB (mean difference=76.521, sig.<0.0055) textA/textC (mean difference=68.989, sig.<0.0195) For details of the significant main univariate effects of PRESMODE on *Chinese* summary length (F=11.447, sig.<0.0015, partial η^2=0.073), please see Chapter 9.					
III	M	TXT	0.907	3.621	0.007	0.048	0.873
		LANGORD	0.911	6.992	0.001	0.089	0.922
		TXT*LANGORD	0.945	2.081	0.083	0.028	0.616
		Box's M=14.107, F=0.907, sig.<0.556					
	U	• Significant univariate effects of TXT on *Chinese* summary length only (F=5.088, sig.<0.0075, partial η^2=0.066). Pairwise comparisons: textA/textB (mean difference=75.779, sig.<0.0085) textA/textC (mean difference=60.389, sig.<0.0605, approaching sig.) • Significant univariate effects of LANGORD on *English* summary length only (F=10.374, sig.<0.0025, partial η^2=0.0.067) • Significant interactive effects of TXT*LANGORD on *Chinese* summary length only (F=3.575, sig.<0.0305, partial η^2=0.047)					
IV	M	TXT	0.917	3.222	0.013	0.043	0.826
		TOEFL-R	0.974	1.926	0.149	0.026	0.395
		Box's M=8.296, F=1.355, sig.<0.229					
	U	Significant effects on *Chinese* summary length only (F=3.373, sig.<0.0375, partial η^2=0.044) Pairwise comparisons: textA/textB (mean difference=63.741, sig.<0.0425)					
V	M	TXT	0.902	3.765	0.005	0.050	0.887
		PRESMODE	0.927	5.586	0.005	0.073	0.851
		TOEFL-R	0.987	0.959	0.386	0.013	0.214
		TXT*PRESMODE	0.967	1.191	0.315	0.016	0.372
		Box's M=22.294, F=1.433, sig.<0.122					
	U	• Significant effects of TXT on *Chinese* summary length only (F=4.59, sig.<0.0125, partial η^2=0.06) Pairwise comparisons: textA/textB (mean difference=69.863, sig.<0.0175) textA/textC (mean difference=63.933, sig.<0.0455) textB/textC (mean difference=-5.931, n.s.) Significant effects of PRESMODE on *Chinese* summary length only (F=11.144, sig.<0.0015, partial η^2=0.072)					
VI	M	TXT	0.920	3.031	0.018	0.041	0.800
		LANGORD	0.919	6.285	0.002	0.081	0.891
		TOEFL	0.973	1.955	0.145	0.027	0.400
		TXT*LANGORD	0.944	2.092	0.082	0.029	0.619
		Box's M=14.501, F=0.932, sig.<0.527					
	U	• Significant effects of TXT on *Chinese* summary length only (F=3.496, sig.<0.0335, partial η^2=0.047) Pairwise comparisons: textA/textB (mean difference=65.388, sig.<0.0335) • Significant effects of LANGORD on *English* summary length only (F=9.327, sig.<0.0035, partial η^2=0.061) • Significant interaction effect of TXT and LANGORD on *Chinese* summary length only (F=3.572, sig.<0.0315, partial η^2=0.048).					

Keys: I-VI models see Figure 10.1; M=multivariate tests, U=univariate tests;
Note: The analyses above excluded both the univariate and multivariate outliers.

Appendix 36: Multiple comparisons of summarization performances between the three groups of text difficulty judgements

Multiple Comparisons

Scheffe

Dependent Variable	(I) txt difficulty in understanding	(J) txt difficulty in understanding	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
EERSC	Low	Medium	7.616*	2.7749	.027	.716	14.515
		High	9.024*	3.5661	.045	.158	17.891
	Medium	Low	-7.616*	2.7749	.027	-14.515	-.716
		High	1.409	3.1257	.904	-6.363	9.181
	High	Low	-9.024*	3.5661	.045	-17.891	-.158
		Medium	-1.409	3.1257	.904	-9.181	6.363
EPRSC	Low	Medium	4.429	2.7121	.268	-2.314	11.173
		High	5.146	3.4855	.340	-3.521	13.812
	Medium	Low	-4.429	2.7121	.268	-11.173	2.314
		High	.716	3.0551	.973	-6.880	8.312
	High	Low	-5.146	3.4855	.340	-13.812	3.521
		Medium	-.716	3.0551	.973	-8.312	6.880
CERSC	Low	Medium	7.579	3.3764	.086	-.817	15.974
		High	12.789*	4.3392	.016	2.000	23.578
	Medium	Low	-7.579	3.3764	.086	-15.974	.817
		High	5.210	3.8033	.395	-4.246	14.667
	High	Low	-12.789*	4.3392	.016	-23.578	-2.000
		Medium	-5.210	3.8033	.395	-14.667	4.246
CPRSC	Low	Medium	6.081	2.8723	.112	-1.061	13.223
		High	12.284*	3.6913	.005	3.106	21.463
	Medium	Low	-6.081	2.8723	.112	-13.223	1.061
		High	6.203	3.2355	.165	-1.841	14.248
	High	Low	-12.284*	3.6913	.005	-21.463	-3.106
		Medium	-6.203	3.2355	.165	-14.248	1.841
EEHS	Low	Medium	.150	.4149	.936	-.875	1.176
		High	.616	.4904	.456	-.596	1.829
	Medium	Low	-.150	.4149	.936	-1.176	.875
		High	.466	.4108	.527	-.550	1.481
	High	Low	-.616	.4904	.456	-1.829	.596
		Medium	-.466	.4108	.527	-1.481	.550
EPHS	Low	Medium	.889	.4767	.182	-.297	2.074
		High	.821	.6126	.411	-.702	2.344
	Medium	Low	-.889	.4767	.182	-2.074	.297
		High	-.067	.5370	.992	-1.403	1.268
	High	Low	-.821	.6126	.411	-2.344	.702
		Medium	.067	.5370	.992	-1.268	1.403
CEHS	Low	Medium	.691	.4585	.324	-.443	1.824
		High	1.870*	.5419	.003	.530	3.209
	Medium	Low	-.691	.4585	.324	-1.824	.443
		High	1.179*	.4540	.037	.057	2.301
	High	Low	-1.870*	.5419	.003	-3.209	-.530
		Medium	-1.179*	.4540	.037	-2.301	-.057
CPHS	Low	Medium	.840	.5149	.269	-.441	2.120
		High	1.503	.6617	.081	-.142	3.149
	Medium	Low	-.840	.5149	.269	-2.120	.441
		High	.664	.5800	.522	-.778	2.106
	High	Low	-1.503	.6617	.081	-3.149	.142
		Medium	-.664	.5800	.522	-2.106	.778
ESL	Low	Medium	14.38	15.310	.644	-23.47	52.22
		High	-1.45	18.096	.997	-46.18	43.29
	Medium	Low	-14.38	15.310	.644	-52.22	23.47
		High	-15.82	15.159	.581	-53.30	21.65
	High	Low	1.45	18.096	.997	-43.29	46.18
		Medium	15.82	15.159	.581	-21.65	53.30
CSL	Low	Medium	23.18	29.815	.740	-50.54	96.86
		High	73.39	35.240	.118	-13.72	160.50
	Medium	Low	-23.18	29.815	.740	-96.86	50.54
		High	50.23	29.521	.238	-22.75	123.20
	High	Low	-73.39	35.240	.118	-160.50	13.72
		Medium	-50.23	29.521	.238	-123.20	22.75

* The mean difference is significant at the .05 level.