ESTABLISHING THE VALIDITY OF READING-INTO-WRITING TEST TASKS FOR THE UK ACADEMIC CONTEXT

By

SATHENA HIU CHONG CHAN

A thesis submitted to the University of Bedfordshire, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

November 2013

ESTABLISHING THE VALIDITY OF

READING-INTO-WRITING TEST TASKS

FOR THE UK ACADEMIC CONTEXT

SATHENA HIU CHONG CHAN

ABSTRACT

The present study aimed to establish a test development and validation framework of reading-into-writing tests to improve the accountability of using the integrated task type to assess test takers' ability in Academic English. This study applied Weir's (2005) socio-cognitive framework to collect three components of test validity: context validity, cognitive validity and criterion-related validity of two common types of reading-into-writing test tasks (*essay task with multiple verbal inputs* and *essay task with multiple verbal and non-verbal inputs*). Through literature review and a series of pilot, a set of contextual and cognitive parameters that are useful to explicitly describe the features of the target academic writing tasks and the cognitive processes required to complete these tasks successfully was defined at the pilot phase of this study. A mixed-method approach was used in the main study to establish the context, cognitive and criterion-related validity of the reading-into-writing test tasks. First of all, for context validity, expert judgement and automated textual analysis were applied to examine the degree of correspondence of the contextual features (overall task setting and input text features) of the reading-into-writing test tasks to those of the target academic writing tasks. For cognitive validity, a cognitive process questionnaire was developed to assist participants to report the processes they employed on the two reading-into-writing test tasks and two real-life academic tasks. A total of 443

questionnaires from 219 participants were collected. The analysis of the cognitive validity included three stands: 1) the cognitive processes involved in real-life academic writing, 2) the extent to which these processes are elicited by the reading-into-writing test tasks, and 3) the underlying structure of the processes elicited by the reading-into-writing test tasks. A range of descriptive, inferential and factor analyses were performed on the questionnaire data. The participants' scores on these real-life academic and reading-into-writing test tasks were collected for correlational analyses to investigate the criterion-related validity of the test tasks. The findings of the study support the context, cognitive and criterion-related validity of the integrated reading-into-writing task type. In terms of context validity, the two reading-into-writing tasks largely resembled the overall task setting, the input text features and the linguistic complexity of the input texts of the real-life tasks in a number of important ways. Regarding cognitive validity, the results revealed 11 cognitive processes involved in 5 phases of real-life academic writing as well as the extent to which these processes were elicited by the test tasks. Both reading-into-writing test tasks were able to elicit from high-achieving and low-achieving participants most of these cognitive processes to a similar extent as the participants employed the processes on the real-life tasks. The medium-achieving participants tended to employ these processes more on the real-life tasks than on the test tasks. The results of explanatory factor analysis showed that both test tasks were largely able to elicit from the participants the same underlying cognitive processes as the real-life tasks did. Lastly, for criterion-related validity, the correlations between the two reading-into-writing test scores and academic performance reported in this study are apparently better than most previously reported figures in the literature. To the best of the researcher's knowledge, this study is the first study to validate two types of reading-into-writing test tasks in terms of three validity components. The results of the study proved with empirical evidence that reading-into-writing tests can successfully operationalise the appropriate contextual features of academic writing tasks and the cognitive processes required in real-life academic writing under test conditions, and the reading-into-writing test scores demonstrated a promising correlation to the target academic performance. The results have important implications for university

admissions officers and other stakeholders; in particular they demonstrate that the integrated reading-into-writing task type is a valid option when considering language teaching and testing for academic purposes. The study also puts forward a test framework with explicit contextual and cognitive parameters for language teachers, test developers and future researchers who intend to develop valid reading-into-writing test tasks for assessing academic writing ability and to conduct validity studies in such integrated task type.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

I owe my deepest gratitude to my PhD supervisors, Professor Cyril Weir and Dr Fumiyo Nakatsuhara for their excellent teaching, guidance, and constant support. I would like to thank them for believing in my potential three years ago and for shaping my apprenticeship as an academic during the entire PhD journey. I would also like to thank Professor Anthony Green for reading two chapters of the thesis and for his advice on statistical analysis. I would like extend my appreciation to the examiners, Mr Anthony Lilley OBE and Dr. Lynda Taylor for their insightful comments and advice on improving this thesis.

I am grateful to the Language Training and Testing Center (LTTC) for funding elements of the study. I would like to extend my thanks to all staff members of CRELLA, who provided valuable insights on my work. In particular, I would like to thank Dr John Field for many insightful discussions about cognitive processing of language use, and Nicola Latimer for proofreading the manuscript of the thesis.

I am also grateful to the lecturers in the Department of Language and Linguistics and Business School who allowed me to conduct the research during their classes, and to many students who participated in the study. Without their participation the research could not have been successfully completed.

Above all, I am thankful for my family, as always. I thank God every day for each of them. Last but not least, thank you - my husband.

1 INTRODUCTION

1.1 Background of the study

1.1.1 Global context

International students have come to the United Kingdom for higher education for centuries. However, the number has increased significantly in recent decades. According to the UK Council for International Students Affairs (UKCISA, 2012), international student numbers grew to 428,225 in 2010-2011. Among the population, 67,325 came from China (excluding those from Hong Kong and Macau), making it the top non-EU sender country. Vision 2020 (Bohm et al., 2004), a document prepared by the British Council, projected a continuous growth in the number.

UKCISA, which is the UK's national advisory body serving international students, stated in their 2011-2012 annual review (UKCISA, 2012) that language testing was one of the challenges for them and British institutions to cope with the largest changes to immigration rules for a decade. All international students who wish to study at tertiary level in the UK now need to provide proof of their English language ability with a qualification from a recognised English language test provider before they can be accepted onto a course. The requirements have increased demands for valid English for Academic Purposes (EAP) tests. EAP tests now have an important gate-keeping function to provide information about whether the prospective students have met the linguistic threshold necessary to cope with tertiary level education through the medium of English. It is essential that these language tests provide evidence of test takers' ability in language skills which resemble the skills people actually use in the real-life academic context.

Many post-hoc validation studies of high-stake EAP test results have been conducted to demonstrate the relationships between test scores and academic performance. However, the research findings do not seem to provide consistent evidence of the relationship between test scores and academic performance. Some studies found little or no significant relationship between high stakes language tests and academic results (e.g. Cotton & Conrow, 1998; Dooey, 1999). Others found low to moderate correlations between test scores and academic Grade Point Averages (GPA) (e.g. Davies & Criper, 1988; Hill, Storch, & Lynch, 1999; Kerstjens & Nery, 2000). Given that the relationship between test scores and academic performance is complex and subject to the effects of intervening variables, Weir (2005) argued that it is essential to define the construct clearly at a beginning stage of test development so that the test task will reflect the contextual features of the real-life tasks and elicit the language skills in a way that resembles how people employ them in a real-life academic context.

A good example of this in the last decade, following our improved understanding of test validity, is the reappearance of integrated reading-into-writing tasks which have regained popularity in standardised language tests. Reading-into-writing refers to 'a test that integrates reading with writing by having examinees read and respond to one or more source texts' (Weigle, 2004: 30). Common reading-into-writing tasks are summary tasks, essays from multiple sources, report writing from multiple sources, case studies, and literature reviews etc. For example, TOEFL underwent several major changes in the direction of integrated test tasks (see Cumming et al., 2004; Cumming et al., 2005). The reformed TOEFL iBT (ETS, 2013), which was introduced in 2006, added an integrated task which requires test takers to write an essay based on reading and listening input materials. The review exercise of Trinity's Integrated Skills in English (ISE) in 2010 affirmed the use of reading-into-writing tasks (Trinity College London, 2009, 2012). In addition, the future versions of the writing papers in some of the Cambridge English Language Assessment examinations will reintroduce a reading-into-writing summary task (Weir, Vidakovic, & Galaczi, 2013; Weir, 2013). Reading-into-writing tasks can also be found in more recent standardised tests such as

Pearson Test of English (PTE) Academic (Pearson, 2010), Language Training and Testing Center's (LTTC) General English Proficiency Test (GEPT) (LTTC, 2012) and EIKEN's Test of English for Academic Purposes (TEAP) (EIKEN, 2013). The revival of such integrated writing tasks seems to suggest that reading-into-writing is once again considered a better option to assess students' academic writing ability (See Weir at al 2013, Chapter 2, for an account of the history of the use of this task type).

While there is a widespread regaining popularity of integrated reading-into-writing task in standardised academic language tests, there seems to be insufficient empirical validity evidence of such test task format in the literature (Asencion, 2004; Esmaeili, 2002; Plakans, 2008, 2010; Weigle, 2004). Therefore, there is a need to collect validity evidence of the integrated reading-into-writing test tasks to assess academic writing skills.

1.1.2 Local context

Apart from providing test scores from a recognised language test prior to their admission onto a course, international students are typically asked to take an in-house language test for diagnostic purposes when they have joined the university. The University of Bedfordshire, which is one of the twenty largest recruiters of international students in the UK (UKCISA, 2012), requires all international students (30% of the total population) to take the Password Test (English Language Testing, 2013), which is a test of academic English knowledge in the form of 100 selected response items. The test is used to diagnose their language proficiency in order to ensure that the students can benefit maximally from the learning experience. Based on their test scores of the Password Test, students are then assigned to three levels of interventions: a drop-in service which offers immediate help and advice on Academic English, a more thorough independent consultation which helps students to set targets and recommends suitable tasks, and academic English classes to help improve students' English skills.

Researchers have argued that integration across reading and writing skills is essential for academic success (Carson, 2001; Carson & Leki, 1993; Flower, 1990; Grabe, 2001, 2003; Johns, 1993; Leki & Carson, 1994, 1997; Lenski &

Johns, 1997). L2 students, and arguably L1 students with less academic writing experience, would need training of academic writing activities which involve reading materials. The Password test has proven to be a valuable and reliable tool for its intended purpose of discriminating students effectively from the A2 to C1 Common European Framework (CEFR) level[1] (see Green, 2012). However, it assesses only language knowledge and therefore provides no information about the test taker's academic writing ability. For the university to offer more support in academic writing, an additional diagnostic test of academic language skills seems necessary. The language testing literature suggests that the integrated reading-into-writing format might be a more valid option to assess academic writing ability when compared to the independent writing-only format (Cumming et al., 2005, 2004; Plakans, 2008, 2010; Weigle, 2002, 2004; Weir et al., 2013, Chapter Three). However, there is considerably insufficient support for local universities like University of Bedfordshire which need to develop a valid academic writing test which involves reading materials. This research aims to unpack the specific contextual (i.e. task features) and cognitive (i.e. cognitive processes required to complete the tasks) parameters of valid academic reading-into-writing tests. The results of this study will assist both global international test developers and local universities in developing a valid test of academic writing ability.

1.2 Aims of the present research

The integrated reading-into-writing task type has the potential to satisfy the need for greater validity in the assessment of test takers' academic writing ability for both international and local EAP contexts. However, in order to achieve the validity, there is a need to collect validation evidence of the integrated reading-into-writing test task format in terms of different important components such as task features and cognitive processes elicited under the test conditions. The socio-cognitive framework (Weir, 2005) marks the first systematic attempt at providing language testing stakeholders, such as test developers, test takers and test score users (e.g. universities, teachers) with a

[1] The CEFR divides communicative proficiency into six levels arranged in three bands - Basic User (A1 and A2), Independent User (B1 and B2), Proficient User (C1 and C2).

coherent and accessible methodology for test development and validation. The framework conceptualises the test validation process by identifying different types of validity evidence that need to be collected at different stages, i.e. the a priori and a posteriori stages, of test development and validation (Geranpayeh & Taylor (eds), 2013: 27). The framework covers five components of test validity: (1) context validity, (2) cognitive validity, (3) scoring validity, (4) consequential validity and (5) criterion-related validity.

Context validity concerns the internal task features and linguistic demands of the test task, as well as the external social and cultural contexts in which the test task is used (for more detail, see Section 2.5.1). Cognitive validity concerns the cognitive processes elicited by the test task (for more detail, see Section 2.5.2). Linking directly to the context and cognitive validity components, scoring validity concerns the extent to which the task is objectively and reliably scored to produce reliable and valid decision-making indicators. Consequential validity addresses the social consequences of test score interpretation and the impact of the test on teaching and learning. Criterion-related validity is concerned with the extent to which test scores correlate with a suitable external criterion of performance (for more detail, see Section 2.5.3).

While the framework has been widely used in many test validation research projects, its application is currently limited to tests of four independent language skills: reading, writing, speaking and listening (e.g. Geranpayeh & Taylor (eds), 2012 - Examining Listening; Khalifa & Weir, 2009 - Examining Reading; Shaw & Weir, 2007 - Examining Writing; Taylor (ed), 2011 - Examing Speaking). This study aims to extend the application of the framework for integrated reading-into-writing tests in terms of context validity, cognitive validity and criterion-related validity. Consequential validity and scoring validity are beyond the scope of the study and were not investigated.

This study builds on the framework (Weir, 2005) to define the construct of academic writing which involves reading activities in terms of the contextual features of real-life academic writing tasks and the cognitive processes students used to complete these tasks. The findings will provide insights into

the target construct of a valid academic writing test in terms of precisely defined contextual features and cognitive processes. Based upon the findings, the study aims to investigate two components of the a priori evidence - contextual and cognitive validity of two different reading-into-writing test task types (*essay with multiple verbal inputs* and *essay with multiple verbal and non-verbal inputs*). The study also aims at the a posteriori evidence to explore the predictive power of the two reading-into-writing test task types. The findings of the thesis will provide empirical evidence of the validity of using integrated reading-into-writing tasks to assess academic writing ability.

1.3 Overview of the thesis

This study investigates validation evidence of the reading-into-writing tasks in terms of contextual validity, cognitive validity, and criterion-related validity. This introductory chapter has provided the global and local background of the study and the aims of research. This sub-section provides an overview of the thesis.

Chapter Two: Literature Review starts with a review of the nature of academic writing in relation to task types and cognitive processing. The chapter then discusses the dominant use of independent writing-only tasks in the past standardised language tests and the issues arising from such practice as discussed in the literature. After that, the chapter discusses whether integrated reading-into-writing tasks would be a more valid tool to assess academic writing ability. The chapter provides an overview of the definitions of reading-into-writing, and discusses the desirability of the task type, the use of such a task type in current standardised language tests, and the concerns and challenges of using reading-into-writing. The chapter then describes the socio-cognitive approach to test validation, and discusses in detail three major validity considerations, namely context validity, cognitive validity and criterion-related validity of EAP reading-into-writing tests. For each validity consideration, relevant studies in the literature concerning reading-into-writing are reviewed. The chapter ends with the research questions of the study.

Chapter Three: Methodology describes in detail the research methods of the present study. The chapter firstly describes the two real-life academic writing tasks and two reading-into-writing test tasks investigated in this study. The chapter then describes the qualitative and quantitative research methods used to investigate the contextual features of the real-life and reading-into-writing test tasks, the cognitive processes elicited by the real-life and reading-into-writing test tasks, and the correlations between real-life academic outcomes and test scores. Within each of these sub-sections, details such as participants, development of the research instruments, data collection procedures and methods of data analysis are presented.

Chapter Four: Establishing the Contextual Validity of Reading-into-Writing Tests to Assess Academic Writing Ability describes the findings of Research Question 1. Contextual features that are likely to influence the difficulty of a task were analysed using both automated tools and expert judgement. The chapter first presents and discusses the results regarding the overall task setting between the real-life writing tasks and reading-into-writing test tasks. The chapter then presents the results of the input texts features of real-life writing tasks and reading-into-writing test tasks. Results from expert judgement are then presented, followed by the results from automated textual analysis.

Chapter Five: Establishing the Cognitive Validity of *Reading-into-Writing Tests to Assess Academic Writing Ability* provides answers to Research Question 2. This chapter begins with the findings on the cognitive processes used by the participants in the real-life conditions. Significant differences of individual questionnaire items between the two real-life tasks are presented. After that, to define the target cognitive constructs, the chapter presents the results of the exploratory factor analysis (EFA) of the five academic phases elicited under the real-life conditions. The chapter then further compares the cognitive processes elicited by the two real-life tasks, and compares the cognitive processes employed by the high-achieving and low-achieving participants. The findings reveal the appropriate cognitive parameters for valid EAP writing tests. Subsequently, the chapter reports the cognitive processes elicited by the two reading-into-writing tests. The chapter makes

comparison of the cognitive processes elicited under test and real-life conditions in terms of the whole population and in proficiency groups, as well as comparison of the processes employed between high-achieving and low-achieving on each test task. The chapter then reports the results of EFA of the five academic phases elicited by each of the two reading-into-writing tests to discuss the underlying structure of the cognitive processes elicited by each of the task types. The findings reveal the extent to which reading-into-writing tests elicit the target cognitive processes from test takers.

Chapter Six: Establishing the criterion-related validity of Reading-into-Writing Tests to Assess Academic Writing Ability addresses the results of Research Question 3. The chapter provides details of the participants' proficiency level as measured by IELTS, and presents the participants' performances on the two reading-into-writing test tasks and the selected writing tasks in the real-life academic conditions. The chapter then presents the results from the correlational analyses between the two reading-into-writing test scores and the real-life scores to discuss the extent to which performances on reading-into-writing tests can predict test takers' ability to perform on real-life academic writing tasks.

Chapter Seven: Conclusions and Limitations summarises the salient findings of the three research questions in this study. The limitations of this study are discussed, followed by the contributions and implications of this study.

2 LITERATURE REVIEW

2.1 Introduction

This chapter is organised into six sections. Following this introduction (Section 2.1), Section 2.2 reviews the relevant literature to discuss the nature of academic writing in terms of task features and cognitive processing. Section 2.2.1 discusses the nature of academic writing tasks. The nature of the academic writing process is then discussed from different perspectives: knowledge transforming (Section 2.2.2), recursive multiple processes (Section 2.2.3), and integration of reading and writing skills (Section 2.2.4). Section 2.2.5 summarises the cognitive phases involved in academic writing which are most relevant to the context of this study.

Section 2.3 reviews the use of independent writing-only test tasks to assess academic writing ability. A review of the issues arising from such practice as discussed in the literature is provided (Section 2.3.1 to Section 2.3.3).

Section 2.4 considers whether integrated reading-into-writing would be a better tool to assess academic writing ability. Definitions of reading-into-writing are reviewed in Section 2.4.1. Section 2.4.2 reviews the desirability of such a task type. Section 2.4.3 reviews the current use of reading-into-writing tasks in standardised language tests. Section 2.4.4 reviews the major challenges as presented in the literature.

Section 2.5 shifts the focus to validation in language testing and discusses the approach taken by this study to establishing the validity of reading-into-writing test tasks to assess academic writing ability. Three major validity considerations in developing a valid academic writing test are discussed: context validity, cognitive validity and criterion-related validity (Section 2.5.1 to Section 2.5.3). The purpose is to derive some broad categories for

investigation in the present study. Section 2.5.1 proposes the contextual parameters which are most relevant to the discussion of the context validity of reading-into-writing test tasks. As the cognitive validity of reading-into-writing tests has received very little attention in the literature, the present study aims to fill this gap. Section 2.5.2 proposes the cognitive parameters which are most relevant to the discussion of the cognitive validity of reading-into-writing test tasks. Section 2.5.3 provides a brief review of the previous relevant studies of criterion-related validity.

This study, to the knowledge of the researcher, is the first study to establish a comprehensive validity argument for reading-into-writing as an academic writing test by examining three major validity components, i.e. cognitive, context and criterion-related validity. As stated in Chapter One, this study investigates the validity of reading-into-writing test tasks to assess academic writing ability using the socio-cognitive framework (Weir, 2005). The framework covers social, cognitive and evaluative (scoring) dimensions of language use and links these to the context and consequences of test use. Other frameworks developed during the 1990s, e.g. Bachman's (1990) Communicative Language Ability model and the Council of Europe's (2001) Common European Framework of Reference (CEFR), have undoubtedly addressed key issues of test development and validation from a theoretical perspective or by providing a set of reference level descriptions. However, when compared to these frameworks, the socio-cognitive framework is believed to have the following advantages:

(1) The socio-cognitive framework not only provides for theoretical consideration of test development and validation issues but can also be applied practically for critical analyses of test content across the proficiency spectrum.

(2) The socio-cognitive framework has direct relevance and value to operational language testing. The framework has provided a workable framework for the development and validation of large-scale language tests of four independent language skills: reading, writing, speaking and listening (e.g. Geranpayeh & Taylor (eds), 2013 - Examining Listening;

Khalifa & Weir, 2009 - Examining Reading; Shaw & Weir, 2007 - Examining Writing; Taylor (ed), 2011 - Examing Speaking).

(3) The cognitive dimension of the socio-cognitive framework addresses the current emphasis on the test taker (O'Sullivan, 2000) – it enables test developers to systematically define the target cognitive processes they aim to test and to monitor if these processes are elicited by the test task before the live testing event.

This study focuses on three components of the framework – context validity, cognitive validity and criterion-related validity. Consequential validity and scoring validity are beyond the scope of the study and were not investigated in the study. Context validity and cognitive validity are a priori components of test validation to be obtained before the live testing event whereas criterion-related validity is a posteriori component. The a priori evidence of test validation enables test stakeholders to define and evaluate the nature and quality of a test inwardly whereas the a posteriori evidence relates to the score interpretation and test use outwardly at the target real-life context in which the test is located. This study will put slightly more emphasis on the two *a priori* validity components, i.e. context and cognitive validity than the a posteriori criterion-related validity component. This is because any results regarding criterion-related validity have to be supported by valid contextual and cognitive evidence. (See Section 2.5 for more detail on the socio-cognitive framework).

The chapter ends with the research questions of the study (Section 2.6).

2.2 The nature of academic writing

This section explores the nature of academic writing by discussing the types of writing tasks used in the real-life academic context (Section 2.2.1) and the nature of the academic writing process (Section 2.2.2 to Section 2.2.4). Based on the discussion, Section 2.2.5 summarises the cognitive phases involved in academic writing which are most relevant to the cognitive validity of academic writing tests.

2.2.1 Academic writing tasks: writing from reading sources

The best way to understand academic writing is perhaps to survey the types of writing that are required of students in the real-life academic context. Researchers have surveyed the tasks that are required of students in educational contexts across recent decades (e.g. Bridgeman & Carlson, 1983; Carson, 2001; Horowitz, 1986a, 1986b; Johns, 1993; Leki & Carson, 1994; Weir, 1983). Some of these studies were conducted with the purpose of test development, for instance, the Test of English as a Foreign Language (TOEFL) (Bridgeman & Carlson, 1983; Hale et al., 1996) and the Test of English for Educational Purposes (TEEP) (Weir, 1983). Although the research methods and task terminologies used vary from study to study, their findings have conclusively indicated that most academic writing tasks require students to write from reading sources.

By surveying teachers in 190 academic departments across undergraduate and postgraduate levels in Canada and USA, Bridgeman & Carlson (1983) found that *description and interpretation of non-verbal input* and *comparison and contrast plus taking a position* were the two task types perceived as the most typical by teachers. The two task types identified were adapted to the TOEFL Test of Written English and IELTS Academic Module Writing paper. However, their study was criticised for drawing entirely upon the perceptions of teachers rather than surveying actual tasks. Other studies which surveyed actual writing tasks and/or curriculum and syllabus documents also showed that reading plays a significant role in academic writing tasks. Hale et al (1996) analysed actual writing tasks assigned in 162 undergraduate and graduate courses in several disciplines at eight universities. They found that the most common real-life tasks across disciplines (social sciences' group and sciences' group) and levels (graduate and undergraduate) were *short tasks (i.e. writing tasks which require students to produce an output about half a page long), essays, summaries, reports with interpretation* and *research papers*. Similarly, based on an analysis of writing tasks in 38 faculties, Horowitz (1986a, 1986b) found that reading was essential in the most common academic writing task types. Common tasks he identified included *synthesis of multiple sources, connection of theory and data*, *report*, *research report* and *summary.* Among

these types, *synthesis of multiple sources* was most typical across the 38 faculties. More recently, Cooper & Bikowski (2007), with a pedagogical purpose for EAP, analysed 200 graduate course syllabi from 10 academic departments with follow-up interviews at one university. Their findings showed that library research papers and project reports were the most commonly assigned writing tasks across different disciplines, while reviews, proposals, case studies, and summaries were more common in the social sciences, humanities, and arts domains. Section 2.5.1 will further discuss the more specific contextual parameters which are important for valid academic writing tests.

The findings of these studies showed that most real-life academic writing tasks require students to write drawing upon external materials. In other words, students have to purposefully draw on a variety of external resources, such as textbooks, journal articles, websites, lecture notes, as well as internal resources from the writer's long-term memory, such as genre knowledge, linguistic resources, topic knowledge and strategic use knowledge, during the writing process. Many of the researchers, therefore, concluded that integration across reading and writing skills is essential for academic success (e.g. Carson, 2001; Grabe, 2001, 2003; Johns, 1993; Lenski & Johns, 1997). However, reading and writing have largely been regarded as two independent constructs in most language tests.

Apart from the nature of academic writing tasks, it is important to understand the nature of the academic writing process. The next three sub-sections (2.2.2 – 2.2.4) summarise from the literature three characteristics of the academic writing process: as knowledge transforming; as recursive multiple processes; and as integration of reading and writing skills.

2.2.2 Academic writing as knowledge transforming

This section discusses academic writing as a knowledge transforming process. Scardamalia & Bereiter (1987), from a pedagogical perspective, proposed the models of *knowledge telling* and *knowledge transforming.* The two models differentiate between the characteristics of the writing processes of novice elementary school writers at one end, and those which characterise more

advanced college, undergraduate and graduate writers at the other end of a continuum of writing expertise. Based on previous work (e.g. Lowenthal, 1980; Murray, 1978; Odell, 1980), they believed that the writing processes employed by expert writers involved transformation of knowledge, e.g. facts or opinions on a particular topic. They attempted to describe the writing processes employed by writers at contrasting levels of writing expertise in terms of (1) how knowledge is brought into the writing processes and (2) what happens to the knowledge during these processes. Their findings showed that novice writers (e.g. elementary school students) tended to use a 'knowledge telling' approach whereas advanced writers (e.g. undergraduates and graduates) tended to use a 'knowledge transforming' approach.

The knowledge telling approach to writing refers to a rather linear text generating process by 'telling' existing knowledge and information available from memory which have been automatically activated by the cues provided in the writing task (Scardamalia & Bereiter, 1987). When constructing a text, novice writers rely heavily on these automatic memory probes and they seldom engage in goal-directed planning, monitoring, and revising processes during the composition of the task. Scardamalia & Bereiter (1991) further argued that writers can be very 'skilful' in using the knowledge telling approach to produce coherent and well-formed texts, provided that the topic and genre are familiar to the writers.

On the other hand, advanced writers tend to have a high awareness of the conflicts between available resources and their writing goals. These resources can be internal resources retrieved from their own memory and/or external materials. Scardamalia & Bereiter (1991) found that advanced writers put explicit effort into establishing task representation (i.e. an initial understanding of the writing task) and setting writing goals. They proposed that these writers establish cognitive 'content problem space' and 'rhetorical problem space' to address the problems of 'what to write' and 'how to write'. During the writing process, advanced writers constantly evaluate the available resources against their goals and constraints. Scardamalia & Bereiter (1987) argued that such an approach to writing leads to knowledge transformation, which can be in the form of an enhanced understanding of the subject knowledge or well-

developed opinions about a particular topic. Critically, this process leads to the generation of novel ideas rather than the “retelling” of existing information. In contrast to the knowledge telling approach, the knowledge transforming approach is a complex problem-solving process.

Bereiter & Scardamalia's (1987) model distinguished the fundamental difference between the 'knowledge telling' writing approach typically employed by novice writers and the 'knowledge transforming' approach typically employed by advanced writers. It should be noted that the difference in their writing approaches is also arguably influenced by the type of tasks that they are typically asked to perform. However, the Bereiter & Scardamalia (1987) model did not account much for the interaction between task and processing. Based on their influential work in differentiating the two writing approaches at the two ends of a continuum of writing expertise (from novice writing by school writers or inexperienced L2 writers to expert writing by more experienced graduates), academic writing is widely regarded as a knowledge transforming process (Flower et al., 1990; Spivey, 1984, 1990, 1997; Weigle, 2002; Weir, Vidakovic, & Galaczi, 2013).

2.2.3 Academic writing as recursive multiple processes

The model of Scardamalia & Bereiter (1987, 1991) showed that advanced writers produce texts by employing a knowledge transforming approach. However, to develop valid academic writing tests, we need to understand more specifically the actual cognitive processes employed in real-life academic writing. This sub-section discusses the nature of academic writing as a set of recursive multiple processes. Before the discussion, it is useful to clarify some key terms: cognition, processing, strategy and metacognition, which are commonly used in process studies.

According to Field's (2004) 'Psycholinguistics: The key concepts', *cognition* refers to the faculty which permits a person to think and reason and the *process* involved in thought and reasoning. *Information processing*, which is an approach to analysing cognition developed by Donald Broadbent in the 1950s, refers to the flow of information through the mind when a task is performed. In contexts of communication, *cognitive processing* refers to the

processes / operations underlying (a) the four language skills; (b) the retrieval of lexical items (decoding); and (c) the construction of meaning and discourse level representation (p.224). The use of some low-level cognitive processes, e.g. decoding in reading, can be automatic, especially for skilled language users, and therefore may not be available to report.

Metacognition, is 'thinking about thinking'. It involves pre-planning which cognitive process, such as macro-planning, organising, monitoring and revising, to use and exercising control over the process or taking steps to ensure that its results are stored long term (Field, 2004: 61). Metacognition usually involves higher degree of awareness and is therefore more likely to be reportable.

A *strategy* is a compensatory technique '(a) to compensate for breakdowns in communication due to insufficient competence or to performance limitation and (b) to enhance the rhetorical effect of utterances (Canale, 1983: 339). In most second language studies, *cognitive process* and *strategy* were not made distinctive (e.g. Cohen, 1984; Purpura, 1997; Van Dijk & Kintsch, 1983). However, it is important to differentiate these terms and concepts for the discussion of this study. The cognitive validity of a language test concerns primarily whether the cognitive processes elicited by the tasks (through the specified task features) can represent reasonably the cognitive processing of skilled language users in real-life contexts.

In the writing literature, a considerable amount of research has been conducted in an attempt to establish the cognitive processes involved in writing and the internal and external variables that would impact on the writing processes (e.g. Field, 2004, 2011; Grabe & Kaplan, 1996; Hayes & Flower, 1983, 1980; Hayes, 1996; Kellogg, 1996; Shaw & Weir, 2007). Most of these studies investigated the writing processes in an educational or academic context. Findings which are relevant to the discussion of this study are reviewed below.

A highly influential model of writing was proposed by Hayes & Flower (1980). They investigated the writing processes of adult writers by an innovative use of 'think-aloud' protocols at the time. They proposed that writing is an extended, goal-directed, problem-solving exercise which involves multiple

recursions of *planning, translating* and *reviewing.* The model also explained that the writing processes interact with two other components: *task environment* and *writer's long-term memory*. Task environment is an external component which refers to task variables such as genre, topic and intended readership. Writer's long-term memory refers to the writer's internal content knowledge about the genre, topic and intended readership as well as rhetorical knowledge about how to write. Hayes & Flower's (1980) model challenged the perception of writing as a linear process and largely fixed the terminology of writing processes in the literature (Scardamalia & Bereiter, 1996). However, the model has been criticised that it does not explain how writing processes vary with different task types and how writing processes vary with memory constraints (Shaw & Weir, 2007).

An updated version of the model (Hayes, 1996) expanded the number of the internal and external components which may impact on the writing processes. Internal factors include working memory, long-term memory resources and the motivation of the writers whereas external factors include the physical environment of the task (e.g. the text read so far, the writing medium) and the social environment of the task (e.g. the audience, other texts read while writing). Regarding the writing processes, the revised model replaced the three major processes planning, translating and reviewing by more general process categories: *reflection, text production* and *text interpretation*. Planning was renamed as *reflection,* which involves problem-solving, decision making, and inferencing processes. Writers employ general problem-solving and decision skills in order to achieve writing goals. At the same time, writers make inferences about audience, writing content and so forth. Translation was replaced by *text production* which refers to a more active text producing process guided by cues from the writing plan or text produced so far. Reviewing was no longer a separate process but became part of *text interpretation, which involves reading and scanning graphics.* The new model attempted to describe the complex interactive nature of the writing processes within and among each process category and the relationship between the writing processes and different internal and external components. It emphasised the central role of working memory in writing. However, there is

apparently insufficient explanation about how various components interact with the writing processes, other than a general claim of the theoretical relationships among them. Another important contribution of Hayes's (1996) model was its acknowledgement of 'the role of reading in writing'. Hayes (1996) highlighted three major purposes of reading in writing: (1) reading to define the writing task, (2) reading source texts to obtain writing content, and (3) reading and evaluating text produced so far. However, the model did not explain in detail how the internal and external components impact on the writing processes and how the reading processes interact with the writing processes.

From a perspective of communicative language use, Grabe & Kaplan (1996) attempted to explain the cognitive processes involved in L2 writing. Their model is one of the few L2 writing models in the literature. The model consists of two major phases: the context of language use and a 'verbal working memory' unit. They proposed that goal setting, which is conducted within a task context (e.g. setting, task, text, topic), would activate three components in the 'verbal processing' unit, which are language competence, world knowledge and 'online processing assembly' (i.e. execution of the writing processes). In addition, they argued for the importance of metacognitive awareness and monitoring in the entire writing process. Although the model has drawn attention to the importance of goal setting and metacognitive processing in L2 writing, the model did not seem to distinguish adequately the differences between resources stored in long-term memory and the processes operated by working memory (Shaw & Weir, 2007: 35-36).

The above models are important in shaping the current understanding of the cognitive processes involved in writing. Although these models pointed out that cognitive processes are affected by test takers' characteristics and task and social factors, they do not seem to provide enough explanation of how the cognitive processes are influenced by these factors. More recent models, which build upon psycholinguistic theory, offer a clearer account of how writing processes are influenced by internal factors, when compared to the above models. Kellogg (1996, 1999, 2001) made a strong argument for the importance of working memory in writing. He proposed that the individual

processes of writing draw upon different components of working memory, rather than seeing working memory as a unitary facility. For example, planning and editing make use of spatial working memory, reading and translation make use of verbal working memory, whereas monitoring and interactions among these processes are coordinated by central executive working memory. Field (2004, 2011) proposed a model which accounts for the phases that a writer goes through when they produce a text (as a productive language skill). Field's model was based upon Kellogg's (1996) model and Levelt's (1989) model of speaking. He proposed that writing, as a productive skill, involves the phases of conceptualisation, organisation, encoding (grammatical, lexical, graphic), execution and monitoring (the phases will be further discussed in Section 2.2.5). Drawing upon information processing theory, Field (2004) explained how high proficiency and low proficiency writers tend to approach these phases differently (*High-proficiency writers* usually refers to writers with high proficiency of English whereas *expert writers* usually refers to writers with expertise in writing).

An important issue for language testing is identifying which phases and processes are relevant for test development and validity. Building upon Kellogg's and Field's models, Shaw & Weir (2007) considered five processes: (1) macro-planning, (2) organisation, (3) micro-planning, (4) translation, and (5) monitoring and revising to be most relevant to the discussion of the cognitive validity of writing tests. They argued that valid writing tests should elicit from test takers those core cognitive processes involved in real-life writing. They then evaluated how these processes have been elicited by the Cambridge English Language Assessment writing tests across different levels. Their approach to evaluating the cognitive validity of test tasks with a set of cognitive parameters has laid down principles for research in the field of language testing (the approach will be discussed further in Section 2.5).

The models of writing discussed above have shown that writing, including academic writing, is not a linear act, but involves a set of multiple recursive processes, such as planning (at macro and micro levels), organising, execution/translating, monitoring and revising. Although the writer may

employ particular processes at different phases, the processes are largely overlapping and looping back and forth. In addition, writing is not an isolated act but is influenced by internal variables such as working-memory capacity, long-term memory sources (e.g. linguistic, discourse and content knowledge), as well as external factors, such as task variables and other social variables. The selection of individual processes and the decision as to how to employ them during the writing process is based upon conscious planning, but execution is largely controlled by working memory. Some processes, such as planning and monitoring, are largely influenced by task variables (contextual parameters of academic writing tasks which will be discussed in detail in Section 2.5.1).

Despite their significance, the processes of integrating external reading materials into writing have largely been excluded from these models. Although some models (e.g. Hayes, 1996) have pointed out the essential role of reading as part of the writing process, the nature of these processes (e.g. what types of reading are involved in terms of current reading theories?) and how these processes interact with other processes are largely unclear. As presented in Section 2.2.1, real-life academic writing generally involves the use of reading materials, therefore it is deemed necessary to develop an adequate cognitive writing model which describes how the various inputs into the writing system are processed in order to output a coherent text (Wengelin et al., 2009). The next sub-section discusses the nature of academic writing as integration of reading and writing skills by reviewing relevant models of reading and discourse synthesis.

2.2.4 Academic writing as integration of reading and writing skills

Researchers from a variety of relevant fields (e.g. reading, writing, cognitive psychology) have become more interested in the relations between reading and writing since the 1980s. For the past three decades, research has been conducted to explore the relation between reading and writing (e.g. Tierney & Shanahan, 1991) (see Grabe, 2003 for a review of these studies). However, a cognitive model which accounts for the processes involved in writing with the use of reading sources has yet to be fully developed, especially in relation to

L2 contexts (Hirvela, 2004). This sub-section discusses academic writing as integration of reading and writing skills.

It is beyond the scope of the present thesis to review extensively the available models of reading per se (see Khalifa & Weir, 2009 for a thorough review of different reading models) because reading processing literature is much more well established than the reading-into-writing processing literature. This thesis regards reading-into-writing as a stand-alone construct which is distinct from the reading-only (comprehension) ability or writing-only ability. However, it is useful to summarise the major components of these reading models in order to shed further light on the integrated nature of academic writing. Whilst there is a rich body of research investigating models of reading, in common with writing, there is very limited discussion regarding how the reading processes interact with the writing processes when a writer writes based on reading materials. Field (2004, 2008, 2013) proposed that receptive skills (i.e., listening and reading) involve phases of input decoding, lexical search, parsing, meaning construction, and discourse construction. From the perspective of language testing, Khalifa & Weir (2009) expanded the model of reading to include the processes of, in an ascending order of cognitive demands, word recognition, lexical access, syntactic parsing, establishing propositional meaning, inferencing, building a mental model, creating a text level representation and creating an intertextual representation. Expert readers have high automaticity of the lower-level processes and hence can focus on the higher-level processes (Field, 2004). Khalifa & Weir (2009) provided a detailed account of how these reading processes are tested in standardised reading tests at different levels. However, there is little discussion in the literature about how these reading processes fit into a model of academic writing.

Despite a lack of comprehensive models of writing from sources, some research studies have investigated the 'unique' processes involved in writing from sources (i.e. the processes which are typically not involved in reading comprehension or writing-only tasks). Two important models proposed from this branch of work are related to summarising writing by van Dijk & Kintsch (1983) and synthesis writing by Spivey (Spivey, 1984, 1990, 1997, 2001;

Spivey & King, 1989). Kintsch & Van Dijk (1978) proposed that summarising writing involves three major processes: *deletion* of redundant propositions; *substitution* of a sequence of propositions by a more general one; and *selection* of the macroproposition of the text, or the *construction* of a macroproposition when one is not explicitly stated. In addition, Spivey and colleagues (e.g. Mathison & Spivey, 1993; Spivey & King, 1989; Spivey, 1990, 1997, 2001) conducted a series of studies to investigate the processes involved in different writing tasks which require the use of reading materials. This body of work is known as discourse synthesis which is defined as 'a process in which readers read multiple texts on a topic and synthesize them' (Spivey & King, 1989: 11). The findings of the studies showed that when writing from external reading materials, a writer transforms a new representation of meaning from multiple texts to their own text through three core processes: a) *selecting* relevant content from multiple texts, b) *organising* the content according to the writing goals and c) *connecting* the content from different sources and *generating* links between these ideas. The results of these studies indicate that reading-into-writing activities place higher cognitive demands on students than reading comprehension processes. The discourse synthesis processes, i.e. *selecting*, *organising*, *connecting* and *generating* proposed by Spivey & King (1989) will be investigated in the present study (These processes will be discussed again in Section 2.5.2.1).

In the L2 literature, Plakans (2008) studied 10 participants' writing processes on both reading-into-writing and writing-only tasks by the use of think-aloud protocols as well as pre-protocol and post-protocol interviews. Based on her findings, she proposed a reading-into-writing model. The results identified two stages of reading-into-writing: *preparing-to-write* and *write*. She argued that reading plays an important role at both stages in terms of 'reading and interacting with source texts' and 'using source texts'. However, the nature of the reading process (e.g. careful reading or search reading) and the interaction were largely unexplained. In a later publication, Plakans (2009) explored the role of the reading process in completing the task. She identified five processes used by the participants: (a) goal-setting for reading the source texts, (b) cognitive processing, (c) global strategies, (d) metacognitive strategies, and

(e) mining the source texts for use in writing. While the work provided useful insights into which reading processes writers employed when they wrote from reading materials, Plakans (2009) did not explain how these processes fit into the reading-into-writing model she proposed earlier (2008). In this study, in order to investigate the processes of a large number of writers on reading-into-writing tasks, questionnaire instead of think-aloud protocol will be used (further discussion regarding the research method is provided in Chapter Three). This study aims to fill the knowledge gap of how 'reading' processes interact with other writing processes during the completion of reading-into-writing tasks.

2.2.5 Cognitive phases involved in academic writing

As argued earlier, most real-life academic writing involves the use of reading materials. Therefore, when building a model of academic writing, it seems inaccurate and inadequate to consider academic writing only as a productive language skill. Academic writing might be more accurately understood as an integration of receptive (reading) and productive (writing) skills. Both receptive and productive language skills involve multiple cognitive phases, and each phase involves multiple processes. A major challenge for language testing is how to model these phases and processes under a test validation framework. A series of studies have identified the most appropriate cognitive processes for independent writing examinations (Shaw & Weir, 2007) and independent reading examinations (Khalifa & Weir, 2009) for adult users at intermediate level upwards, i.e. B2-C2 in terms of the CEFR (the proficiency levels university students need to be at). Test takers from an intermediate level upwards are presumed to possess high automaticity in low-level receptive phases, i.e. input decoding, lexical search, and parsing, and low-level productive phases, i.e. encoding. This study considers the following five higher-level cognitive phases to be most relevant to the discussion of the cognitive validity of academic writing tests: (1) conceptualisation, (2) meaning and discourse construction, (3) organising, (4) low-level monitoring and revising, and (5) high-level monitoring and revising.

Conceptualisation (Kellogg, 1996, Field, 2004, 2011) is the first phase of productive skills where the writer develops an initial mental representation of a writing task. Researchers have studied the processes involved in this initial phase of meaning construction. Writers create an initial understanding of the task situation through reading the task prompt. However, Flower (1990) argued that 'interpretation' rather than comprehension is more important at this phase. They refer to such a process as 'task representation', which is 'an interpretive process that translates the rhetorical situation – as the writer reads it – into an act of composing' (Flower, ibid: p.35). Another important process related to this phase is the process of planning (Hayes & Flower, 1980). Field (2004) and Shaw & Weir (2007) further distinguished the planning process conducted at macro- and micro- level. Planning conducted at this phase is largely at macro-level. Shaw & Weir (2007) defined macro-planning as a process of determining what is necessary for successful for task completion in terms of different aspects for consideration such as intended readership, genre, content, style.

Meaning and discourse construction is a higher-level phase from the model of receptive skills (Field, 2004, 2013). Meaning and discourse construction is a phase when the writer contextualises abstract meanings based on the contextual clues provided in the writing task and their own schematic resources (background knowledge) (Field, 2004, 2013) and integrates the information from different sources into a discourse representation (Brown & Yule, 1983). Kintsch & van Dijk (1983) argued that when writers summarise, they evaluate the relative importance of information from the reading materials, and evaluate whether the information fits into the macro- or micro-structure of their text. In Khalifa & Weir's (2009) model of reading, higher-level processes are important to meaning and discourse construction. These higher-level processes, including *establishing propositional meaning*, *inferencing*, *building a mental model*, *creating a text level representation* and *creating an intertextual representation*, seem to be relevant to this phase. According to Spivey (1990, 1997; Spivey & King, 1989), writers establish a discourse representation by (1) selecting information (which could be retrieved from long-term memory or selected from input texts) which is

relevant to the context, and (2) connecting the selected information from different sources to each other.

Organising is a phase where the writer 'provisionally organises the ideas, still in abstract form, in relation to the text as a whole and in relation to each other (Field, 2004, 329)'. Shaw & Weir (2007) explained that the process of organising employed when writing is to order the ideas to 'determine which are central to the goals of the text and which are of secondary importance (p.38)'. Spivey (1990, 1997; Spivey & King, 1989) argued that the process of organising is particularly challenging for writing which involves the use of reading materials.

Low-level monitoring and revising and **High-level monitoring and revising** are 'feedback' phases where the writer checks the quality of the text. After monitoring, a writer usually revises the unsatisfactory parts of the text (Field, 2004, 330). Monitoring and revising can focus on lower-level aspects of text quality such as accuracy or higher-level aspects such as argument and coherence. Monitoring and revising can be employed at any point of the writing process. They can be made mentally before the text has been composed, at the point of translating (i.e. at the current location of the text), or after the text has been translated (i.e., at a previous point in the text) (Field, 2004; Fitzgerald, 1987). Researchers (Field, 2004; Kellogg, 1996; Shaw & Weir, 2007) argued that monitoring and revising are highly demanding in terms of cognitive effect. Writers, especially L2 writers, tend to focus on one aspect of the text at a time due to short-term memory constraints. With attentional constraints, many writers would set aside high-level monitoring and revising to a later stage of the production.

This sub-section has identified some broad cognitive phases of academic writing which are useful for the discussion of cognitive validity of academic reading-into-writing tests. Individual writing phases involved at each phase are discussed again in Section 2.5.2 with more attention focused on how each process may help to distinguish unskilled writers from skilled writers. The next section shifts the attention to how academic writing is currently tested in standardised writing tests.

2.3 The use of independent writing-only tasks

As discussed above, real-life academic writing tasks almost always involve some external reading materials. Such an integrated task setting, however, seems to be under-represented in most writing tests. Horowitz (1986a, 1986b, 1991) argued that there is a fundamental discrepancy regarding the use of primary or secondary reading materials between real-life tasks and most writing test tasks. Studies have been conducted to review the task types used in writing assessments (see Weigle, 2002, Shaw & Weir, 2007; Weir et al., 2013). Their results showed that the independent writing-only task type has played a dominant role in most high-stakes language tests and university admission tests. The independent writing-only task type refers to tasks which do not require the use of reading sources. Test takers are expected to produce the text by drawing solely on their internal resources, e.g. background knowledge on the topic. Among different genres, the essay task (i.e. test takers write an essay in response to a point of view, problem, or an argument provided in a single line prompt) is found to be very common in writing assessments. IELTS Writing Task 2 would be a typical example of an independent writing-only task (see below for the task).

Task 2A

You should spend about 40 minutes on this task.

Write about the following topic.

> ***The first car appeared on British roads in 1888. By the year 2000 there may be as many as 29 million vehicles on British roads.***
>
> ***Alternative forms of transport should be encouraged and international laws introduced to control car ownership and use.***
>
> ***To what extent do you agree or disagree?***

Give reasons for your answer and include any relevant examples from your knowledge or experience.

Write at least 250 words.

(taken from IELTS sample paper http://www.cambridgeenglish.org/exams-and-qualifications/ielts/whats-in-the-test/)

As the independent writing-only task type has served in many high-stakes language tests, there is extensive discussion about the issues of using such a task type in the literature. Following an improved understanding of the nature of academic writing ability, researchers (Moore & Morton, 1999, 2005; Moore, Morton, & Price, 2010; Plakans, 2008; Shaw & Weir, 2007; Weigle, 2002) argued that writing-only tasks might not be the most suitable tool to assessing academic writing ability. The two key issues are related to task authenticity and test fairness, which are discussed below.

2.3.1 Unsatisfactory task situational authenticity (context validity)

Under the influence of the communicative testing approach, testing practitioners and researchers have become more aware of the importance of having test tasks which simulate reasonably the target language use context (Weigle, 2002; Weir, 1993; Shaw & Weir, 2007). Therefore, the design of the independent writing-only task type has been criticised for its lack of authenticity (see Cumming, 1997; Hamp-Lyons & Kroll, 1996; Lumley, 2005; Weigle, 2002, 2004). Writing tests, by nature, elicit sample performance from the test taker by using a very limited number of writing tasks. While it is impossible to simulate all real-life situations in any testing context, task authenticity is a fundamental concern of good language tests (Bachman & Palmer, 1996). Task authenticity can be achieved in terms of *situational* and *interactional* authenticity (Bachman & Palmer, 1996; Shaw & Weir, 2007). Shaw & Weir defined the situational authenticity as 'the contextual requirements of the tasks (2007: 9)' and the interactional authenticity as 'the cognitive activities of the test taker in performing the test task (ibid)'. In more recent frameworks of test validation (e.g. Weir, 2005), *situational* authenticity is part of context validity whereas *interactional* authenticity is part of cognitive validity (Test validation will be discussed in Section 2.5).

Situational authenticity (context validity) considers whether the test task itself is similar to the real-life tasks that the test takers are expected to encounter in the target language use context. As presented earlier, many studies which surveyed the writing demands in different academic contexts have concluded that academic writing is rarely done in isolation, but is overwhelmingly done

in response to source texts (e.g. Bridgeman & Carlson, 1983; Carson, 2001; Horowitz, 1986a, 1986b, 1991; Johns, 1981, 1993; Leki & Carson, 1994; Weir, 1983). Moore & Morton (1999, 2005) conducted one of the very few studies that compared test tasks and real-life tasks. They compared the IELTS Task 2 rubric with a corpus of 155 assignment tasks at both undergraduate and postgraduate levels across 79 academic departments in two Australian Universities. They made two specific observations regarding the discrepancy between the test tasks and real-life academic tasks in terms of the use of external sources and language functions.

First, the real-life academic tasks typically involved the use of primary sources (e.g. textbooks, journal articles, monographs) or secondary sources (e.g. a simple exhortation), either provided in tasks or collected by students. In contrast, they found that the test tasks did not engage test takers with external information. Test takers were required to process their prior knowledge while completing the task. Second, the real-life academic tasks usually involved more than a single language function. The most common functions were *evaluation, description, summarisation comparison, explanation* and *recommendation.* However, the predominant function identified in the test tasks was *evaluation*. The functions of s*ummarisation* and *description* were not identified in their sample of IELTS Task 2. Moore & Morton (1999, 2005), therefore, concluded that the typical essay test task seemed to represent the 'genre' but not specific contextual features of real-life academic essay tasks. Moore & Morton's studies have demonstrated the value of comparing the rubrics of actual test tasks to those of real-life tasks. This study will also compare the task features of reading-into-writing tests to those of real-life academic writing tasks. In addition to the rubric, this study will investigate the features of the reading materials of the reading-into-writing tests and the real-life tasks. The task analysis in Moore & Morton's studies was mainly conducted by the researchers. In this study, the tasks and the reading materials will be analysed by multiple sources (details are provided in Section 3.3).

2.3.2 Unsatisfactory task interactional authenticity (cognitive validity)

Considering interactional authenticity, independent writing-only tasks require the test takers to engage in 'writing from internal resources'. However, the process of writing from sources is considerably more cognitively demanding than the process of writing from internal sources (Plakans & Gebril, 2009).

Studies of task representation, discourse synthesis and summarising (e.g. Flower, 1990; Spivey & King, 1989; Spivey, 1990, 1997; van Dijk & Kintsch, 1983) reviewed previously showed that writing from sources required specific skills which may not be required when writers write from internal sources. Therefore, there is great concern about the construct being tested by the independent writing-only task type as it does not represent real-life academic writing processes. Another concern is that independent writing-only tasks would seem to encourage the 'knowledge telling' rather than 'knowledge transformation' approach to writing. Although knowledge telling is an approach typically employed by immature writers, Scardamalia & Bereiter (1987) argued that even advanced writers may use such an approach when they were asked to write on a 'knowledge telling' task, e.g. to produce a familiar genre which mainly involves recalling internal resources on a familiar topic.

2.3.3 Background knowledge effect (test fairness)

Another concern with the use of independent writing-only tasks is related to test fairness due to heavy topic effect. As argued previously, the independent writing-only task type requires the test taker to write drawing upon internal resources from their long-term memory. Weigle (2004) argued that students' performances are likely to be influenced by topic effect imposed by tasks which provide no input. This has inevitably led to test fairness issues when the topic of the writing task favours some test takers or is biased against others. Brown, Hilgers & Marsella (1991) conducted a large study to investigate the impact of topic on 3452 students' writing performance in a standardised language test. Ten topic prompt sets were used. The results showed that various topic prompts led to significant differences in the scores. The researchers thus concluded that writer's background knowledge on the topic

was a variable affecting the quality of his/her writing performance. Hughes (2003) argued that writing tasks in general language tests should not require the test taker to demonstrate specific topic knowledge. Douglas (2000) argued that an appropriate level of disciplinary topic knowledge should be part of the construct of the English for specific purposes (ESP) tests. However, the issue is less straightforward for English for academic purposes (EAP) tests. Although EAP tests are one type of ESP tests, the majority of EAP tests are not discipline-specific. The testing population of most large-scale EAP tests consists of test takers from a wide range of academic disciplines. In other words, these test takers do not share the same disciplinary background knowledge. Therefore, using independent writing-only tasks which require test takers to draw upon their background knowledge on the topic to assess their academic writing ability may not be the most appropriate method.

This section has discussed the use of independent writing-only tasks in writing assessments and the issues arising from the practice. The next section reviews and discusses the use of integrated reading-into-writing tasks.

2.4 Would integrated reading-into-writing be a better alternative?

Due to the above concerns about the use of independent writing-only tasks in high-stakes EAP tests, there has been a resurgence in the popularity of integrated reading-into-writing tasks over the past two decades. This sub-section reviews the literature regarding the use of reading-into-writing tasks. Definitions of reading-into-writing are reviewed in Section 2.4.1. Section 2.4.2 discusses the desirability of such a task type. Section 2.4.3 reviews the current use of reading-into-writing tasks in standardised language tests. Section 2.4.4 reviews the major challenges of using reading-into-writing tasks.

2.4.1 Definitions of reading-into-writing

Before the review of the use of integrated reading-into-writing tasks in writing assessments, it is useful to consider some definitions of reading-into-writing tasks (reading-to-write).

From a pedagogical perspective, Ascención Delaney (2008) defined reading-into-writing as 'instructional tasks that combine reading and writing for various education purposes' (p.140). Flower et al. (1990) defined reading-into-writing as 'the process of a person who reads a relevant book, an article, a letter, knowing he or she needs to write (p.6)'.

From a language testing perspective, Weigle (2004) defined reading-into-writing as 'a test that integrates reading with writing by having examinees read and respond to one or more source texts' (p.30). The term ‘integrated’ has been used by large-scale testing providers, e.g. English Testing Service (ETS) and Trinity College London, as a category to refer to their reading-into-writing tasks. As this study is primarily concerned with language testing, the terminology of 'reading-into-writing' rather than 'reading-to-write' is used throughout the thesis.

Reading-into-writing tasks are sometimes referred to as 'discourse synthesis' tasks due to the influential work conducted by Spivey (1984, 1990, 1997; Spivey & King, 1989) which has been reviewed previously. However, this study considers 'discourse synthesis' tasks to be a subordinate type of reading-into-writing task. In this study, reading-into-writing tasks refer to single tasks that require students to write a continuous text by drawing upon single or multiple reading materials which can be verbal, non-verbal or both. Students may or may not need to find additional reading materials on their own. Reading-into-writing tasks include, but are not limited to, summary tasks, response (argumentative) essays from multiple sources, report writing from multiple sources, case studies, and literature reviews.

2.4.2 Desirability of the reading-into-writing task type

Integrated reading-into-writing tasks can arguably fulfil validity considerations better than the dominantly used independent writing-only tasks do. This notion is well supported in the current research on writing assessment (Grabe & Stoller, 2002). Researchers in the field of language testing have argued for the use of this type of task in assessing academic writing abilities (e.g. Cumming et al., 2005; Cumming, Grant, Mulcahy-Ernt, & Powers, 2004; Hughes, 2003; Pollitt & Taylor, 2006; Weigle, 2002, 2004; Weir et al., 2013). This sub-

section summarises the four major reasons why reading-into-writing tasks might be more appropriate in academic writing tests.

2.4.2.1 Improved task authenticity (context validity)

As mentioned in Section 2.3, task authenticity is a fundamental consideration for the validity of any language test. A task is considered to be authentic if it represents the features of the task in the target language context and if it elicits processes which are similar to those the test takers have to use in the target language context. Many writing researchers have argued that as far as academic writing is concerned, writing an impromptu essay on a previously unseen topic is an inauthentic task. Therefore, reading-into-writing tasks are believed to be able to better represent the 'performance conditions' of real-life academic tasks (e.g. Carson, 2001; Hamp-Lyons & Kroll, 1996, 1997; Johns & Mayes, 1990; Johns, 1981; Leki & Carson, 1994; Plakans, 2008, 2010; Weigle, 2002). Specific contextual parameters that are important for academic reading-into-writing are addressed in Section 2.5.1.

2.4.2.2 Eliciting integration of skills (cognitive validity)

One may argue that the reading ability involved in performing writing tasks has been covered in reading tests. However, researchers reported no significant correlations between test takers' performances on writing-only and reading-into-writing abilities (e.g. Ascención Delaney, 2008; Yu, 2008). Studies have also found that there is only about 25% to 50% overlap between reading and writing ability (Grabe, 2003). Therefore, testing reading ability and writing ability separately would not sum up adequately the ability in completing writing tasks that involve integration of reading materials (Ascención Delaney, 2004, 2008).

Khalifa & Weir (2009) found that careful reading comprehension has been the focus of most standardised reading tests. Higher-level reading skills at the discourse and intertextual level and expeditious search reading skills have often been neglected. Weir et al (2013) further argued that higher-level intertextual reading can be best tested by reading-into-writing tasks. Other processes writers employ when they write from sources, as identified in

Spivey's discourse synthesis model, are important for academic writing, but these processes seem to have received little or no attention in most current writing tests.

Oller (1979) criticised the use of the discrete-point approach to language testing in earlier days. He argued that 'in any system where the parts interact to produce properties and qualities that do not exist in the parts separately, the whole is greater than the sum of its parts, organizational constraints themselves become crucial properties of the system which simply cannot be found in the parts separately' (p.212). His argument offers an important insight for the current discussion.

2.4.2.3 Providing equal access to subject knowledge (test fairness)

The provision of input reading materials in reading-into-writing tasks not only reflects real-life context, but also ensures equal access to subject knowledge among test takers. As indicated by different models of writing reviewed in Section 2.2.3, prior knowledge is one of the factors which contribute to the writing performance. A reading-into-writing test provides students with appropriate reading texts which may supply them with content ideas, which are necessary for task completion. In terms of test fairness, it is very challenging, if not impossible, to make sure that a topic is equally familiar to all test takers who take the same test. However, well designed reading-into-writing tasks would provide all test takers with equal access to the content which is sufficient for them to complete the task (Weir et al, 2013). The potential bias of the topic effect imposed on test takers would then be minimised, because even if a test taker is unfamiliar with the topic, she or he would not be disadvantaged.

2.4.2.4 Positive washback effect (consequential validity)

Washback broadly refers to the effect of a test on teaching and learning (see Green, 2007 for detailed a discussion of washback). Washback is part of the consequential validity of the socio-cognitive framework, although consequential validity covers broader social impact of test. The importance of positive washback in language testing is emphasised in the literature (Hughes,

2003). The literature generally regards reading-into-writing tests as having positive washback (Belcher & Hirvela, 2001; Campbell, 1990; Cumming et al., 2004; Esmaeili, 2002; Tierney & Shanahan, 1991; Weigle, 2004; Weir, 1983). Nevertheless, Wall & Horák (2006, 2008) identified contributing factors that shape the impact of an operational integrated test.

Based on the results of a student survey, Leki and Carson (1994, 1997) argued that students need practice of a range of 'more challenging literacy tasks that combine reading and writing'. In addition, Johns (1981, 1993; Lenski & Johns, 1997) made a very strong argument that L2 students need to be exposed to a range of academic tasks, so they can understand the demands of academic tasks and develop necessary corresponding skills for their undergraduate or postgraduate studies, e.g. search for relevant materials, careful comprehension, read for main ideas, build intertextual representations, summarise reading materials, express own interpretation. His points of view have largely shaped the recent development of EAP literacy with a focus on reading-writing relations.

In addition, writing tasks that involve integration of reading materials are regarded as having good pedagogical value for literacy development. Researchers generally regard reading-writing relations as mutually supportive in terms of literacy development (Grabe, 2003). For example, Leki (1993) argued that summary writing can improve reading comprehension skills. More specifically, Armbruster, Anderson & Ostertag (1987) found that summarising instruction can facilitate higher-level reading skills, e.g. identifying the macrostructure of a text. Some argued that reading-into-writing tasks can promote the development of 'critical literacy' through high-level processes of integrating existing texts to create texts of their own (Flower et al., 1990). Others have found that reading-into-writing tasks can lead to effective content learning (e.g. McCarthy & Leinhardt, 1998; Perfetti, Britt, & Georgi, 1995; Wiley & Voss, 1999).

On the other hand, Wall & Horak (2006, 2008) investigated the factors which were likely to play an important role in shaping the impact of an operationalised integrated test (the TOEFL iBT) on teaching and learning.

They argued that teachers' good awareness of the nature of integrated tasks (as compared to the more widely used independent writing-only tasks) were crucial for positive washback to happen in classrooms. However, they found that while most teachers involved in the studies had a good awareness of the nature of the integrated tasks, their understanding of the tasks was not perfect. They found that the gap in the understanding was partly due to the degree of explicitness present in the explanations of the test, e.g. the test web site and the official test preparation materials and other commercial test preparation materials. Another factor was the availability of resources needed to design courses to help students to cope with the demands of the integrated task type.

2.4.3 Use of reading-into-writing tasks in standardised writing tests

Integrated reading-into-writing tasks are increasing in popularity and either replacing or complementing writing-only independent tasks used in assessing academic writing (Gebril & Plakans 2009, Weigle 2002, 2004). The use of reading-into-writing tasks in language tests can be traced back as far as the early 1930s. According to Weir et al's (2013) book which reviews language testing in the past century, reading-into-writing tasks were used in large-scale language tests as early as 1931. *Summary* was apparently the earliest integrated reading-into-writing task type used in large-scale language tests (e.g. a summary task was added to the English Literature paper of the Certificate of Proficiency (CPE) in English in 1936) (CPE was used for academic admission purposes). It was chosen because summary writing was similar to what people did in a lot of real-life occupational contexts at that time, which is still true at present (for details of the use of summary in CPE, see Weir et al 2013: 128). However, the integrated task type fell out of favour in the 1970s when the testing of separate language skills was preferred.

The communicative approach to learning gained popularity during the early eighties, so that the integrated reading-into-writing task type was again used in high-stakes writing tests. For example, in Certificate in Advanced English (CAE) (the test has been renamed as *Cambridge English: Advanced*) Paper 2 (Writing), Part 1 requires test takers to integrate a range of reading inputs, e.g. newspapers/magazines, letters, reports. However, the CAE Writing paper was

revised in 2008 so that the length of the examination was reduced from two hours to 90 minutes. The number of words to be written and the reading materials to be read were substantially reduced. The Test of English for Educational Purposes (TEEP) was developed in 1980s based on an extensive research study on the language problems of International students in the UK (Weir, 1983). Task 1 of the TEEP requires test takers to produce a summary of about 200 words based on one passage. The TEEP was redeveloped during 1991 and 2001. The current format of integrated writing task of the TEEP requires test takers to use their own ideas as well as ideas retrieved from the reading and listening materials (University of Reading, 2013). The original IELTS writing tasks developed in 1989 required test takers to write upon the reading materials of the reading section (for details of the development of IELTS, see Davies, 2008 Chapter 5). Task 1 required test takers to transfer and repossess non-verbal information, e.g. diagrams, tables, charts (15 minutes) whereas Task 2 required test takers to draw on information from a variety of the reading materials they read previously in the reading section in addition to using their own experience to present an argument or solve a problem (30 minutes). However, in 1995, IELTS decided to drop the integrative nature of the writing tasks, particularly Task 2, by removing the thematic link between the Reading and Writing Modules. This was largely due to concerns of muddied measurement (see Charge & Taylor, 1997). This concern will be discussed further in Section 2.4.4.1. Post-1995 IELTS Writing Task 2 has become an independent writing-only task where test takers are required to write upon their own experience to present an argument or solve a problem. As a result, the independent writing-only 'essay' task was used as the dominant task in most large-scale writing tests during the 20th century. Those integrated reading-into-writing tasks developed in the 1980s were either completely dropped from the test (e.g. IELTS Task 2) or downscaled to involve substantially fewer reading materials (e.g. CAE Writing Paper 1).

The use of reading-into-writing tasks in large-scale writing examinations did not come back to place until more recently. Large-scale language tests, not only those in the U.K. but worldwide, have once again shown interests in incorporating different types of integrated tasks into their writing paper. For

example, Pearson's PTE Academic and Trinity College London's ISE exams in the U.K., LTTC's General English Proficiency Test (GEPT) in Taiwan, EIKEN's TEAP[2] in Japan, and Georgia State Test of English Proficiency (GSTEP) and ETS's TOEFL iBT in the U.S..

Pearson launched a computer-based test for Academic English which is called PTE Academic in 2009. Part 1 of the Writing Section is an integrated task which requires test takers to write a one-sentence summary (not more than 30 words) of a passage after reading a text in ten minutes (Perason, 2010) (see Appendix 2.1.1 for an example of the task). Trinity College London's Integrated Skills in English (ISE) exam III is a level-specific examination targeting at the level of CEFR C1. The exam includes an integrated reading-into-writing task. Task 1 involves writing an article upon multiple verbal and non-verbal reading materials (see Appendix 2.1.2 an example of ISE III reading-into-writing task) (Trinity, 2013). GEPT Advanced developed by LTTC in Taiwan is another level-specific test at C1. The writing paper of GEPT Advanced includes two integrated reading-into-writing tasks. Task 1 requires test takers to summarise the main ideas from multiple verbal materials and express own opinions whereas Task 2 requires test takers to summarise the main ideas from multiple non-verbal materials and provide solutions (see Appendix 3.1.3 for an example of GEPT Advanced Task 1) (LTTC, 2013). In the U.S., following an extensive revision exercise during the early 21st century, Education Testing Service (ETS) added an integrated writing task which requires test takers to write upon reading and listening to materials towards TOEFL (the test was renamed as TOEFL iBT). Georgia State Test of English Proficiency (GSTEP) is a university admission test. The integrated reading-into-writing task requires test takers to write upon two passages which are also used in the Short Answer Section (Weigle, 2002). Table 2.1 provides a summary of the above mentioned reading-into-writing tasks used in current large-scale language tests for academic use.

As documented in Table 2.1, there is apparently a revival of the integrated task type. A range of reading-into-writing tasks are being widely used in

[2] TEAP has not been operationalised and the detailed test information is not yet available.

standardised language tests for academic purposes. As described previously, the integrated reading-into-writing task type has been subject to different attitudes in the past, i.e. the integrated task type was dropped from language tests and used again. One reason is perhaps a lack of thorough construct arguments in the literature concerning reading-into-writing test tasks.

Table 2.1 A summary of the current uses of reading-into-writing test tasks for academic purposes

Test	Task	Task description	Input format	Output	Time	Marking criteria	CEFR level
Cambridge English: Advanced	Part 1 – Q1	Write an article/a report/ a proposal/a letter	Verbal: up to 150 words	Report/ proposal /letter (180-220 words)	1 hr 20 mins (including another writing task)	• Content • Organisation and Cohesion • Appropriacy or Register and Format • Range • Target reader	C1
GEPT Advanced	Task 1	Summarising main ideas from verbal input and expressing opinions	Verbal: 2 texts (about 400 words each)	Essay (at least 250 words) • Introduction • Summarise the main points • State own viewpoint • Conclusion	60 mins	• Relevance and adequacy • Coherence and organization • Lexical use • Grammatical use	C1
	Task 2	Summarising main ideas from non-verbal input and providing solutions	Non-verbal: 2 graph/table/chart/diagram	Letter/report (at least 250 words) • Introduction • Summarise the main findings • Discuss possible reasons • Make suggestions	45 mins		

Georgia State Test of English Proficiency (GSTEP)	Integrated Reading and Writing	Write an essay responding to a prompt about the reading passages	Verbal: 2 argumentative texts (300-500 words)	Essay	45 mins	• Content • Organization • Language • Range • Complexity • Language • Accuracy	N/A
IELTS – Academic	Task 1	Describe some visual information	Non-verbal: 2 graph/table/chart/diagram	Description (150 words)	20 mins	• Task achievement • Coherence and Cohesion • Lexical resource • Grammatical range and accuracy	N/A
Integrated Skills in English (ISE) Exam III	Integrated reading into writing task	Write a report or article based on multiple verbal and non-verbal materials	Verbal: a passage and a shorter text, perhaps in bullet points Non-verbal: chart or table or diagram	Report / Article (about 300 words) • Summarise information from the source texts • Give your own opinion	2 hrs and 30 mins (including another two writing tasks)	• Task fulfilment • Accuracy and range	C1
PTE Academic	Summarise Written Text	After reading a text, write a one-sentence summary of the passage	Verbal: a text up to 300 words	Summary (not more than 30 words)	10 mins	• Content • Form • Grammar • Vocabulary • Spelling	N/A
TOEFL iBT - Writing	Integrated Writing Task - Read/Listen/ Write	Write essay responses based on reading and listening tasks	Verbal: • a reading text (230-300 words) • a listening text (230-300 words)	Summary (150-225 words) • summarise the main points in the listening passage • explain how these relate to the key points of the reading passage	50 mins (including another writing task)	• Content (accuracy and completeness) • Appropriate use of language and sentence structure	N/A

Some research has been done on test tasks which involve non-verbal materials, e.g. IELTS Writing Task 2 (Bridges, 2010), GEPT Advanced Writing Task 2 (Yu and Lin, forthcoming), a single verbal material, e.g. PTE Academic Writing Part 1 (See Chan, 2010). Researchers such as Plakans (2009, 2010) has investigated reading-into-writing test tasks which involved multiple reading materials, but not in a testing context. Therefore, to narrow the research gap, this thesis aims to investigate reading-into-writing test tasks which involve multiple verbal inputs, e.g. GEPT Advanced Writing Task 2, and multiple verbal and non-verbal inputs.

2.4.4 Concerns and challenges of using the reading-into-writing task type

While there has been a resurgent interest in using reading-into-writing in large-scale writing assessments, it is important not to forget that the use of integrated reading-into-writing tasks does not offer a simple solution to the problems found with impromptu writing tasks (Plakans, 2008). Additionally, there are some unique challenges attached to the integrated task type. This section will address some of the concerns of using reading-into-writing tasks to assess academic writing ability raised in the literature.

2.4.4.1 Muddied measurement?

First of all, some researchers have questioned the possibility of a 'muddied measurement' (Weir, 2005: 101) due to the confusing effects of reading and writing abilities on the reading-into-writing performance (Alderson, Clapham, & Wall, 1995). Their concerns are understandable when reading and writing used to be understood largely as two mutually exclusive constructs. However, based on current improved understanding of the reading-writing relationship and the nature of academic writing ability (as reviewed in Section 2.2), it is felt that reading-into-writing tasks actually measure integrated language skills which involve, but are not limited to, high level reading skills of creating a global representation at a text or intertextual level and knowledge transforming writing skills in an authentic manner. Weir et al (2013) argued that no independent task type can possibly assess such integrated language skills. Their recommendation of reintroducing a reading-into-writing summary task

in Cambridge examinations (See also Khalifa & Weir, 2009 Chapter 8) will take place in the 2013 version of the writing paper in CPE.

Nevertheless, it is a common concern that poor performances on any reading-into-writing could be a result of poor basic comprehension skills (Cumming et al, 2004). It is, therefore, important to control the level of the reading texts so that the target test takers should be able to comprehend the reading texts, an issue which is going to be discussed next.

2.4.4.2 Appropriate input

As discussed previously, while the provision of reading texts can reduce the potential bias of background knowledge imposed on test takers by providing an equal starting point, the task design may unintentionally hinder students from completing the task if they cannot comprehend the texts. Therefore, the level of the reading input texts is critical to the effectiveness of any reading-into-writing task. There is no doubt that the difficulty of the reading materials should be set at an appropriate level in terms of cognitive and linguistic demands, but the discussion should be based on a clear understanding of the purpose of the test. This study considers the purpose of reading-into-writing tasks to be assessing academic writing skills which involves high-level intertextual reading skills, as explained in Section 2.2.5. The use of reading-into-writing tasks to assess basic reading comprehension skills is beyond the focus of this study.

There is a rich literature on how to identify the difficulty of a reading text/test. This body of research will be revealed in Section 2.5.1.2. One recent study of such was conducted by Wu (2012) who investigated the issue of the comparison of level-based test batteries by comparing two reading tests - GEPT and Cambridge Examinations- at two CEFR levels - B1 and B2. Among other results, she found that 'the Cambridge B2 level tests were significantly more difficult than the GEPT counterpart in terms of test takers' performance and cognitive demands, but the Cambridge texts were significantly less complex than the GEPT in terms of contextual features' (206).

Test takers in her study performed significantly better on GEPT test (higher cognitive demands on reading skills) than the Cambridge test (higher linguistic demands on comprehending the texts) at the same CEFR B2 level. This finding is particularly relevant to our discussion for reading-into-writing task because it raises an interesting question of the role of linguistic demands and cognitive demands in determining the level of a test. There does not seem to be much discussion regarding this issue in the literature. This study aims to shed light on the issue by analysing the source texts students read while completing their written tasks in the real-life and test contexts. The findings of this study will hopefully reveal useful cognitive and contextual parameters to differentiate different levels of reading-into-writing test.

2.4.4.3 Extensive copying of the input materials

Another problem that arises in reading-into-writing tasks is that the provision of reading materials seems to have led to significant lifting of the input materials by students (Shi, 2004). The earlier mentioned study by Cumming et al. (2005) found that writers at the middle-range of proficiency tended to use more phrases verbatim from source texts than did their more or less proficient counterparts. Students in Yu's (2008) study honestly admitted that they preferred the use of English summary task as a means to assess their academic writing ability because they could 'copy directly from or refer to the source texts without necessarily fully understanding the copied text or the whole text' (2008: 538). Some researchers are concerned with the negative impact of providing writers with reading materials. For example, Lewkowicz (1997) argued that the provision of reading materials would restrict the development of ideas as students tend to rely heavily on the source texts in terms of ideas and language.

While the above concerns are potential challenges, the problem of inappropriate lifting of sources and too much reliance on the source texts also exists in the real-life academic context, usually addressed as plagiarism. One obvious solution is that the test tasks should reflect plagiarism rules as they apply in the real-life academic context.

From a test development perspective, clear task instructions which warn test takers about the inappropriate use of source texts would help to reduce such behaviour. For example, Weir et al (2013) suggested that a task should state clearly the permitted amount of direct copying from the reading input texts, e.g. no more than 3 words of continuous text. However, it is also important to pay attention to the cognitive demands when setting reading-into-writing tasks. Any good reading-into-writing task should demand a language and content transformation from test takers. It is vital for test developers to be able to specify how test takers are expected to interact with the source texts.

2.4.4.4 Appropriate marking scheme

Another challenge of using reading-into-writing tasks in large-scale writing assessments relates to a seeming lack of appropriate marking schemes. Recent studies have found that texts produced across the two task types varied significantly. For example, Cumming et al (2005) compared the features of texts produced from six trial TOEFL iBT tasks, which included two independent essays, two reading-into-writing essays and two listening-into-writing essays. The results showed significant linguistic differences between the independent essays and the integrated essays, with regards to the aspects of lexical sophistication, syntactic complexity, argument structure, voice in source evidence, and message in source evidence.

However, it is not uncommon for the same marking scheme to be used for both independent writing-only and integrated reading-into-writing tasks. Reading-into-writing tasks target a range of integrated skills of reading and writing which are not likely to be assessed by independent reading comprehension tasks or independent writing-only tasks. There is a lack of discussion in the literature regarding of the qualities of successful reading-into-writing performances (Weir et al, 2013). It is essential to discuss how different levels of the reading-into-writing skills can be addressed in the marking scheme (the cognitive validity of reading-into-writing will be discussed in detail in Section 2.5.2).

The issue of significant lifting of source texts discussed in the previous section also needs to be addressed in the marking scheme so that any inappropriate behaviour can be penalised and reflected in the test scores. However, marking reading-into-writing scripts can be difficult. The Educational Testing Service conducted a series of validation studies while they were revising the old TOEFL test to develop the new integrated TOEFL iBT test. One of the studies (Lee, Kantor, & Mollaun, 2002) found that raters in the trial had difficulty identifying copied language versus students' own wording in the scripts.

2.5 Validation in language testing (The socio-cognitive approach)

Traditionally, validity was seen as an issue of 'whether a test really measures what it is supposed to measure' (Cronbach, 1988; Lado, 1961). Test validity was addressed by individual enquiries of content, construct, concurrent and predictive validities separately until the 1980s. Messick (1989) argued that test validity should be seen as a unified concept which integrates considerations of content, criteria and consequences into a *framework*. His view on unitary validity broke new ground for current understanding of validity (Weir, 2005). The challenges of building a coherent validity argument are to define which evidence is needed to demonstrate different validity arguments of a test, and how and when to obtain such evidence, and from where. Instead of discussing test validity abstractly, the field of language testing has moved towards an argument-based approach in test validation in which guidance for collecting validity evidence against a set of well-defined criteria is provided (For details see Fulcher, 2010; Kane, 2012). The criteria are set based upon the target real-life contexts. As it is impossible to replicate the entire target real-life contexts fully under test conditions, the concept of test validity is a matter of degree of representativeness (relevancy and range). The more a test task can represent the target tasks in the real-life contexts, the more valid the test is. In addition, instead of building validity arguments of the quality of a test itself, current understanding of validity focuses more on the quality of inferences made upon test scores. In other words, the inferences of test scores need to be supported by evidence of different validity components.

Sharing a similar view that each individual validity component combines with others to collectively demonstrate test validity, Bachman and Palmer (1996; Bachman, 1990) argued that construct validity arguments should be built by defining the underlying trait of particular abilities or skills hypothesised on the basis of a theory of language ability. Following this notion, test developers started to make claims regarding which underlying language abilities and skills their tests are meant to measure, and what samples of language skills and structures are represented in the content of the tests. Bachman and Palmer's (1996) approach undoubtedly improved people's awareness of test construct

from a theoretical point of view. However, Hughes (1989, 2003) argued that such an approach to building validity arguments itself would not necessarily demonstrate the validity of a language test because additional empirical research is required to confirm whether (1) such language abilities or skills exist, (2) these abilities and skills can be measured, and (3) they are indeed measured in a particular test.

Following the argument-based approach to collecting evidence of test validity, Weir (1988) proposed that evidence of individual validity components should be collected **before** the test event as well as **after** the test has been administered. He stressed the need for describing the construct that a test attempts to measure at the a priori stage of test development and then evaluating how well the construct is operationalised in the test at the a posteriori stage. Additionally, Weir (1988) argued that test construct can be better defined by the cognitive processing involved in language use in real-life.

Weir (2005) made a notable attempt to develop an evidence-driven socio-cognitive validation framework which integrates the considerations of the underlying **cognitive** ability, the **context** of language use and the process of **scoring** operationalised in the language tests, and the **criterion-related** validity of the tests. **Consequential** validity (Messick, 1989) was also incorporated in the framework. The framework allows test developers and researchers to conduct systematic analyses of test input and output, from both psycho- and socio-linguistic perspectives. Unlike previous frameworks which focused on a uniform construct (content) validity, the socio-cognitive framework unpacked the abstract 'construct' of a test in terms of the cognitive and context components in order to provide strong evidence of construct validity. Language examination boards, such as Cambridge English Language Assessment in the UK, Language Training and Testing Center (LTTC) in Taiwan, Eiken Foundation of Japan (formerly STEP) in Japan, have used the framework to revisit the extent to which these major validity components have been operationalised in their tests of the four skills, i.e. reading, writing, listening and speaking. This socio-cognitive validation approach is believed to have led to improvements in test design and has reframed effective validity arguments (i.e. coherent evidence of different validity components supporting

the interpretation of test scores) regarding the use and interpretation of test scores (see Shaw & Weir (2007) - Examining Writing; Khalifa & Weir (2009) - Examining Reading; Taylor (ed) (2011) - Examining Speaking; and Geranpayeh and Taylor (ed) (2013) - Examining Listening).

While the socio-cognitive framework has undoubtedly made a noticeable contribution to validation in language testing since 2005, its current application limits to language tests which assess the four skills separately. As reviewed in Section 2.4.3, the use of integrated reading-into-writing tasks has seemingly become a feature of 'new' or revised writing assessments. Therefore, it is necessary to extend the application of the framework to the design and validation of the integrated reading-into-writing tests to assess academic writing ability.

This study aims to establish validity evidence of EAP reading-into-writing tests in terms of three components of the socio-cognitive framework (Weir, 2005): context, cognitive and criterion-related validity.

- context validity - the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample (Weir, 2005, p.19)
- cognitive validity - the extent to which the chosen task 'represents the cognitive processing involved in writing contexts beyond the test itself' (Shaw & Weir, 2007, p.34)
- criterion-related validity - the extent to which test scores correlate with a suitable external criterion of performance with established properties (Weir, 2005, p.35)

Context validity and cognitive validity are the most important components of the construct validity. Criterion-related validity is essential because if the test does not predict academic performance then its validity is in doubt. As the first large-scale validation study of reading-into-writing test tasks to assess academic writing ability, it is necessary to conduct a check on the value of using these tasks to predict performance in the target academic contexts. Based upon relevant literature, Section 2.5.1 and Section 2.5.2 propose a set of

contextual parameters and cognitive parameters to be investigated in this study, and Section 2.5.3 reviews previous studies on the criterion-related validity of academic writing tests.

2.5.1 Context validity consideration for academic writing tests

Context validity addresses the appropriateness of the linguistic and content demands set in a writing test task in comparison with the contextual parameters of the writing tasks in the target language use context (Weir, 2005; Shaw & Weir, 2007). Bachman and Palmer (1996: 44) defined target language use domain as '*a set of specific language use tasks that the test taker is likely to encounter beyond the test itself, and to which we want our inferences about language ability to generalise*' (1996: 44). The importance of developing test tasks which are representative of the target language use domain is well perceived in the literature (Bachman, 1990; Bachman & Palmer, 1996; Weigle, 2002; Weir 1983, 1990 and 2005). As test takers only respond to one or two writing tasks in most writing tests, researchers in the field are concerned about how the contextual features of these chosen test tasks may reflect real-life tasks in the target language use domain, and how the contextual features of these tasks should be controlled in order to simulate appropriate performance conditions at different levels. This sub-section discusses the contextual parameters which are most relevant to an academic reading-into-writing context.

In the last decade, a few important publications in the writing testing literature addressed the issue of task variables from slightly different perspectives. Four of the important pieces of work are discussed briefly below.

Hughes (2003), in his text book about language testing for language teachers, set the 'minimum requirement' of writing task design by identifying the task variables need to be specified in a writing test. The most essential task variables include *operations, types of text, addressees, length of texts, topics, dialect and style*.

Weigle's (2002) Assessing Writing, as part of the Cambridge University Press series on language testing, targets a more specialised readership. Building on

the work of Purves, Soter, Takala, & Vahapassi (1984) and Hale et al. (1996), Weigle (ibid) proposed a longer list of the most essential task variables. Apart from *subject matter, genre, rhetorical task, pattern of exposition, specification (of audience, role, tone, and style), length, time allowed*, she added *stimulus material, prompt wording, choice of prompts, transcription mode* and *scoring criteria*.

Shaw & Weir (2007)'s book regarding research and practice in assessing second language writing relates particularly to large-scale writing tests, especially Cambridge examinations. Building on Weir's (2005) socio-cognitive framework, they proposed a more systematic approach to address task variables (context validity) in terms of two major components: *setting* (task and administration) and *linguistic demands* (task input and output). Setting concerns test task as well as administration. Administration refers to the physical testing conditions, logistics and security issues, which will not be discussed in this study. Their list of the task setting variables includes *response format, purpose, knowledge of criteria, weighting, text length, time constraints* and *writer-reader relationship*. Apart from task setting, Shaw & Weir (2007) argued for the importance of operationalising the linguistic demands of task input and output when developing a writing test task. Parameters of linguistic demands include *lexical resources, structural resources, discourse mode, functional resources* and *content knowledge*. Their notion of controlling the demands of task input and output imposed on test takers is particularly useful for reading-into-writing tests. Another contribution of their framework is that they attempted to explain how changes in each of these task variables may influence the processes used by the test takers (cognitive validity), and in turn have an impact on their performance to be scored.

Douglas (2000), from the perspective of Languages for Specific Purposes (LSP) testing, stressed the importance of comparing task characteristics of the target real-life and test conditions. In his framework, dimensions of comparison include rubric, input, expected response, interaction between input and response, and assessment. Task rubric includes *objective, procedures for responding, structure, time allotment* and *evaluation.* Regarding task input and expected response, Douglas (2000) explained thoroughly in his book what

parameters need to be specified in order to provide test takers with an 'authentic' context in the test condition. Task input consists of prompt and input data. A valid task prompt should specify the *features of context* by providing information regarding *setting, participants, purpose, form, tone, language, norms* and *genre*, and identify the *problem* for the test takers. Test developers also need to specify the features of the input data in terms of format, *vehicle of delivery (e.g. written or spoken), length* and *level of authenticity*. In addition, for expected response, test developers need to specify the *format, type, response content (language and knowledge)* and *level of authenticity*. The most insightful category in this framework for our discussion here is the *interaction between input and response* which includes the *scope* and *directness* of how test takers should interact with the reading materials. The last category is assessment which includes *construct definition, criteria for correctness* and *rating procedures*. While Douglas's (2000) framework of task characteristics was the most detailed among those discussed, some of the categories are not strictly task variables, e.g. criteria for correctness and rating procedures are typically regarded as part of the scoring validity. Nevertheless, his list of identifiable task characteristics is certainly important to be taken into consideration while developing a reading-into-writing test that aims to be indicative of the target LSP context.

Building upon their work, this study attempts to identify the contextual parameters which are relevant to academic writing which involves integration of reading materials in terms of 1) overall task setting (productive demands), and 2) input text features (receptive demands). To the knowledge of the researcher, there is seemingly a lack of discussion regarding the contextual parameters that are useful to evaluate the context validity of reading-into-writing tests in the literature. The following Section 2.5.1.1 and 2.5.1.2 review the contextual parameters which have been established in the literature to evaluate the context validity of reading tests and writing tests. A summary of the selected contextual parameters to be analysed in this study is provided in Section 2.5.1.3.

2.5.1.1 Overall task setting (receptive demands)

Contextual parameters regarding the overall task setting need to be specified on a reading-into-writing task to assess academic writing include, but are not limited to, purpose of the task, topic domain, genre, cognitive demands, language functions, discourse mode, intended reader, and knowledge of criteria.

Purpose

Shaw & Weir (2007) stressed that it is vital for a writing task to present test takers with a 'clear, precise and unequivocal' purpose which 'goes beyond a ritual display of knowledge for assessment' (ibid: 71). Douglas (2000) similarly argued that it is important to provide test takers with a purpose for interacting with a context, which he called 'the identification of the problem'. Presenting a clear and precise purpose in the rubrics would arguably be more essential for successful completion of a reading-into-writing task. Research has shown that writers plan how to make use of the reading materials based on how they perceive the task purpose (Flower et al, 1990).

Topic domain

Topic is one of the major variables shown to have significant impact on writing performance (Clapham, 1996; Douglas, 2000; Read, 1990). In the widely adopted CEFR writing grid, topics are categorised into personal and daily life, social, academic and professional domains (Council of Europe, 2001). Tests of the lowest proficient level typically involve topics from personal and daily life, medium level from the social domain whereas the highest level typically involves topics from the academic or professional domains. Previous studies have shown that writers in general perform better with a familiar topic than an unfamiliar one (e.g. Clapham, 1996). Alderson (2000: 69) argued that 'topic (un)familiarity cannot be compensated for by easy vocabulary: both difficult vocabulary and low familiarity reduce comprehension, but texts with easy vocabulary do not become easier if more unfamiliar topics are used, and vice versa'. Urquhart & Weir (1998:143) suggested that text content that test takers are sufficiently familiar with can

activate schemata to employ appropriate skills and strategies to comprehend the text. Therefore, a topic set for a task has a direct impact on the demands of content knowledge and lexical knowledge imposed on test takers.

Genre

Genre is defined as 'the expected form and communicative function of the written product; for example, a letter, an essay, or a laboratory report (Weigle, 2002: 62). Johns et al (2006) argued that a genre should not be thought of only as a type of text (e.g. letter) or as a situation (e.g. meeting a business client); a genre involves both, and is the result of interaction between both. There is 'a fluid relationship between text and context, writer's purposes, voice, and occasion' (Johns et al 2006: 235). Essay seems to be the commonly assessed genre in most large-scale writing assessment. However, as presented in Section 2.2.1, essay is only one of the many genres that are required of students in real-life academic writing contexts.

Cognitive demands

Weigle (2002) is one of the few researchers who regarded the level of cognitive demands imposed on test takers as a task variable. According to Purves et al. (1984), the three major cognitive levels are (i) reproducing facts/ideas, e.g. copying from sources, recalling from long-term memory; (ii) organising / reorganising information, e.g. retelling information, summary; and (iii) generating new ideas through processing the given ideas, e.g. applying, analysing, synthesizing, evaluating. This notion seemingly echoes Scarmadalia & Bereiter's (1987) distinction between the knowledge telling strategies used typically by novice writers and the knowledge transforming strategies typically used by expert writers. Douglas (2000) argued that the cognitive demands of the interaction between input and output should be analysed in terms of the scope of the reading needed to be processed and how directly writers should draw upon the material. In other words, a task with more linguistically challenging input texts would not necessarily be more difficult than a task with easier input texts, if the former requires copying a few factual details and the latter requires a thorough evaluation of the given ideas. According to Fitzgerald & Shanahan's (2000) model of the development of reading and

writing, a student's ability to handle interaction between reading and writing varies across different L1 proficiency stages from lower-level processing of *mastering reading and writing as two separate skills* to higher-level processing of *handling multiple viewpoints in reading and writing*, and *constructing and reconstructing knowledge through reading and writing*. Therefore, it is important to consider the cognitive demands of the reading-into-writing tasks in terms of the expected interaction between input and output texts.

Language functions

Apart from genre, researchers are also concerned about the specific language functions test takers are expected to perform, e.g. discuss, express opinions, justify. Douglas (2000) regarded the functions which test takers are expected to perform on the task, e.g. express himself clearly, edit writing, as construct definition. He argued that the functions tested need to be identified based on context-based research and consultations with subject specialists. Shaw & Weir (2007) saw the possible relationship between the functions tested (what they called functional resources) and the can-do statements in CEFR (Council of Europe, 2001) which aims to identify what learners are able to do across levels. As described earlier, Moore & Morton (1999, 2005) identified some discrepancy between the language functions required by real-life academic tasks and those required by test tasks.

Intended reader

In order to provide an 'authentic' communication context, the audience of the test takers' text needs to be specified. Shaw & Weir (2007) argued that it is important to specify the relationship between reader and writer. The provision of information about intended reader is even more important for reading-into-writing tasks.

Time constraint

Time constraint is one major distinction between real-life and testing conditions. Real-life tasks typically allow more time for completion than a test task does. There is comparatively little discussion in the literature regarding how the time constraint on reading-into-writing tasks influences

test takers' cognitive processes. Field (2004) argued that the difference between skilled writers and unskilled writers is that the former can employ more processes with greater automaticity than the latter. Another question for reading-into-writing tests is the proportion of time allocated for processing the reading materials. Most reading-into-writing tests tend to allow test takers to decide upon the allocation of time themselves.

Knowledge of criteria

Knowledge of criteria is another contextual parameter which often receives insufficient attention. Shaw & Weir (2007) argued that test takers should be fully aware of which criteria are to be used in the marking. Douglas (2000) commented in a similar way that evaluation criteria tend not to be well-specified in task rubrics. As discussed previously, one challenge of current reading-into-writing tests is the lack of specific marking criteria. Knowledge of marking criteria is important because it affects test takers' monitoring and revising processes where differences between expert and novice writers are expected to emerge. In other words, provided with sufficient knowledge of marking criteria, skilled writers tend to monitor and revise their text based on the criteria whereas unskilled writers tend not to do so.

2.5.1.2 Input text features (receptive demands)

In the context of reading-into-writing tests, it is vital to control the receptive and productive demands imposed on test takers through specifying the level of the input provided for them and the output they are expected to produce. As the variety in nature and amount of input material is infinite, it is important to investigate how variation in input would influence the quality of output (Weigle, 2002). However, without an advanced model of reading-into-writing ability in the literature, this is a very challenging task.

Input format

The possibilities of input format provided in reading-into-writing tests include single or multiple verbal / non-verbal inputs (as presented in Table 2.1). There has been surprisingly little research on the effect of different input formats on the writing process. Weigle (1994, 1999) compared two common types of non-

verbal input: table / chart and graph. The findings indicated that the table/chart prompt (making and defending a choice based on information presented in a table or chart) tends to elicit traditional five-paragraph essays, while the graph prompt (describing trends in a graph and make predictions based on the information presented in the graph) would elicit several rhetorical angles. While the rhetorical functions of the task, i.e. make and defend a choice vs. describe trends, also certainly contribute to the differences found in the output, the findings insightfully suggested the need to investigate the variable of input data format. For example, it is important to understand if multiple inputs encourage intertextual processing and result in more complicated rhetorical organisations.

Discourse mode

Discourse mode can be broadly defined as narrative, descriptive, expository and argumentative. Generally speaking, narrative texts recount an event or a series of related events. Descriptive texts describe a person, place or thing using sensory details. Expository texts tend to give information about or an explanation of an issue, subject, method or idea. Argumentative texts typically involve a course of reasoning.

Concreteness/Abstractness of ideas

It is believed that a text with more abstract ideas is harder to understand than a text which contains more concrete ideas, e.g. description of real objects, events or activities. Alderson (2005) argued that the more concrete, conceivable and interesting the ideas are, the more readable the text is. However, concreteness of ideas in a text might be difficult to assess. Although automated textual analysis tools, e.g. Coh-metrix, claim to be able to assess such quality, Wu (2012) found that indices relating to text abstractness and cohesion were of limited use, based on the feedback provided by the expert judges in her study.

Explicitness of textual organisation

The importance of textual and intertextual reading processes to produce an integrated representation of multiple texts has been well established in the literature (For example, see Britt & Sommer, 2004; Hartmann, 1995; Perfetti,

Rouet, & Britt, 1999; Perfetti, 1997; Spivey, 1997). These intertextual processes play a significant role in the cognitive processing that takes place in knowledge transforming writing tasks (Scardamalia & Bereiter, 1987). Therefore, the textual organisation of the source texts would directly impact the complexity of a reading-into-writing task. The reading inputs of reading-into-writing are usually adapted from authentic texts. The textual structure of the test task input text is usually simplified during such procedures to suit the level of the test. However, more evidence is needed to discuss the most appropriate way to control the textual organisation of reading input materials.

Cultural specificity

Hughes (1989, 2003) argued that in language testing the subject areas should be as "neutral" as possible to avoid bias being imposed on particular groups of test takers. Apart from subject knowledge discussed previously, cultural specificity of a text also has an impact on its level of complexity (Sasaki, 2000).

Linguistic complexity of input texts (lexical complexity, syntactic complexity and degree of coherence)

In the reading testing literature, there have been many studies on the complexity of the reading texts through investigating a wide range of textual parameters (See a summary of these studies in Wu (2012). Important studies include Bachman, Davidson, Ryan & Choi's (1995) test comparability study, Alderson, Figueras, Kujper, Nold & Takala's (2006) study, Enright et al.'s (2000) study on TOEFL, and Khalifa & Weir's study (2009) on Cambridge reading examinations. In Bachman et al's (1995) study, textual variables such as the nature of the text, length, vocabulary, grammar, cohesion, distribution of new information, type of information, topic, genre, rhetorical organisation and illocutionary acts were identified. Enright et al (2000), investigating TOEFL reading, identified three groups of salient textual features: grammatical/discourse, pragmatic/rhetorical and linguistic variables. Alderson et al (2004) included text source, authenticity, discourse type, domain, topic, nature of content, text length, vocabulary and grammar as relevant features for text analysis. Khalifa & Weir (2009) summarised the contextual features

proposed in the literature and established a subset of contextual parameters which can effectively distinguish between levels of proficiency in the Cambridge reading examinations.

Some of the above textual parameters, e.g. topic and discourse mode, have already been covered previously (see Section 2.5.1.1). The discussion here focuses on the parameters which impact on the linguistic complexity of the input reading materials. Following the practice of previous studies, to make the discussion more effective, the parameters are divided into three aspects: lexical complexity, syntactic complexity and degree of coherence (Green et al., 2012; Weir et al., 2013 Appendix B; Wu, 2012).

The approach to analysing the lexical and syntactic complexity of a text is reasonably well established in the literature with a list of commonly used lexical indices (e.g. Alderson, 2000; Enright et al., 2000; Khalifa & Weir, 2009; Urquhart & Weir, 1998). **Word frequency, word length and type-token ratio (TTR)** are the most commonly used lexical indices. Word frequency measures the proportion of vocabulary of a text against different word lists such as British National Corpus (The British National Corpus, 2007) and the Academic wordlist (Coxhead, 2000). Word length measures the number of letters or syllables a word contains, as shorter words tend to be easier to read. Type-token ratio measures the number of different words in a text. A higher ratio indicates a higher degree of lexical variation and thus suggests increased text difficulty. As the type-token ratio decreases, words are repeated many times within the text, which should increase the ease and speed of text processing. This measure is particularly useful when texts of similiar length are compared.

Text length, average sentence length, syntactic similarity and **readability formulas** are syntactic-related indices commonly used to reflect the complexity of a text. Text length measures the total length of a text. The longer the text, the more difficult it tends to be. A longer text would require more information processing of word recognition, sentence parsing and propositional encoding (Grabe, 2009). Average sentence length is often used to estimate syntactical complexity because short sentences tend to be

syntactically simpler than long sentences. Syntactic similarity measures how syntactically similar the sentences of a text are. It is easier to process a text with more syntactically similar sentences than with more syntactically different sentences due to a syntactic parsing effect. Readability formulas, such as Flesch Reading Ease and Flesch-Kincaid Grade Level, measure the relative numbers of syllables, words and sentences found in a text. They are widely used to measure text difficulty, especially in the States. Despite the widespread application of the readability formulas, researchers (e.g. Green et al., 2012; Weir 2012; Weir, et al. 2013 Appendix 2) argued that these measures (i.e. Flesch Reading Ease and Flesch-Kincaid Grade Level) seem to come up with results which are closely aligned to individual measures of average syllables, words and sentences and therefore results from individual measures might be more useful for the purpose of development of test materials.

In addition to the lexical and syntactic complexity of a text, **degree of cohesion** is also deemed important to determine the difficulty level of a text. Generally speaking coherent texts tend to be easier to comprehend than less coherent texts (Beck, McKeown, Sinatra, & Loxterman, 1991). Goldman and Rakestraw (2000) showed that cohesive devices which contribute positively to establishing textual coherence would also help readers to connect ideas. However, the effects of cohesion may not be a totally independent indicator because cohesion interacts with the readers' familiarity of the topic and their own proficiency level (Alderson, 2000: 68).

2.5.1.3 Summary

Based on Section 2.5.1.1 and Section 2.5.1.2, the contextual features that are most relevant in describing different levels of reading-into-writing test tasks are compiled in Table 2.2.

Table 2.2 Contextual parameters proposed to be analysed for the context validity of reading-into-writing test tasks

Overall task setting (productive demands)	Input text features (receptive demands)

• Purpose
• Topic domain
• Genre
• Cognitive demands
• Language functions to perform
• Intended audience
• Knowledge of criteria
	• Input format
	• Verbal input genre
	• Non-verbal input genre
	• Discourse mode
	• Concreteness of ideas
	• Textual organisation
	• Cultural specificity
	• Linguistic complexity
	  • Lexical complexity
	  • Syntactic complexity
	  • Degree of cohesion

The traditional method of contextual analysis used to be expert judgement with a check-list questionnaire and/or focus-group interview. Recently, with the aid of automated text analysis, e.g. Cohmetrix, VocabProfile, researchers are able to employ a more systematic and objective methodology for establishing the complexity of a text. A series of studies were conducted to evaluate the features of texts used in different large-scale tests by using automated textual analysis (See Green, Unaldi, & Weir (2010) for the IELTS Academic Reading test; Green et al. (2012) for CAE reading texts, Weir (2012) for the Test of English for Academic Purposes (TEAP) writing scripts, Wu (2012) for GEPT Reading test). Although these studies were conducted for reading tests (with the exception of the Japanese TEAP reading into writing test), the methodology and findings can be adapted to suit the reading-into-writing context. This study will investigate the above contextual parameters by using both qualitative expert judgement and quantitative automated textual analysis. The procedures will be discussed in detail in Chapter Three.

2.5.2 Cognitive validity considerations for academic writing tests

A comparison between the contextual parameters set in real-life academic writing tasks and those set in EAP reading-into-writing test tasks is important but it would not complete the validation of the task type on its own. Bachman (1990) pointed out that the deficit of many validation studies is that they 'examine only the products of the test taking process, the test scores, and

provide no means for investigating the processes of test taking themselves' (269). The evidence collected for the context validity can only indicate if the characteristics of the test tasks are appropriate so that a set of desirable cognitive processes may arguably be elicited from the test-takers. However, the actual cognitive processes employed by test-takers in response to the task prompt might vary according to different internal and external factors, as indicated by models of writing reviewed in Section 2.2. Hale et al's (1996) account of the limitations of their study may further illustrate why contextual parameters on their own are insufficient. Initially, they planned to analyse the writing tasks by four subcategories of cognitive demands (i.e. retrieve/ organise/relate, apply, analyse/synthesise, and evaluate) but they later found it 'unclear from the wording of the assignment alone [contextual parameters]' (1996: 44) which cognitive demands might apply. Their solution was to categorise the cognitive demands into two board categories: lower- or higher-level. Some other researchers also attempted to investigate the cognitive demands of academic writing tasks by analysing the rhetorical functions presented in the wording of the assignment. For example, Moore & Morton (2005) revealed that evaluation, description and summarisation are the three most frequently incorporated rhetorical functions. Their findings revealed, to some extent, what cognitive demands are required of students in order to complete academic writing tasks. However, the cognitive processes test takers actually engage in while completing the writing tasks may not be the same as those they are expected to employ to complete the tasks by the task rubrics/syllabi/teachers. Therefore, effort needs to be made to demonstrate the actual cognitive processes employed by the test takers at different proficiency levels, even though this is 'rarely an easy matter' (Shaw & Weir, 2007: 35).

The second aim of the present study is to investigate the degree of correspondence in cognitive parameters, between EAP reading-into-writing test tasks and real-life academic writing tasks. In order to investigate the cognitive validity of a test task, we need to consider the nature of cognitive process involved in academic writing. By reviewing relevant models of writing, reading and discourse synthesis in the literature in Section 2.2, this study proposes that the following five cognitive phases that a writer is likely to go

through when writing from reading sources are most relevant to the discussion of the cognitive validity of academic writing tests. The five phases are: a) conceptualisation, b) meaning and discourse construction, c) organising, d) low-level monitoring and revising, and e) high-level monitoring and revising (See Section 2.2.5). Based on the literature, the following subsection decomposes the cognitive processes that are involved at each of these phases of academic writing which requires integration of reading materials.

2.5.2.1 Cognitive parameters

The cognitive parameters which are most relevant to the discussion of the cognitive validity of academic reading-into-writing tests are proposed below. The selection of the processes was based upon the models of writing, reading, and discourse synthesis reviewed in Section 2.2. Shaw & Weir (2007) argued that when identifying cognitive parameters to be examined in a test, it is important to demonstrate how writers at different levels would employ these cognitive processes with 'educationally significant differences' (p.142). Therefore, the processes discussed below are those which have been established in the literature as characterising the writing process employed by a skilled writer which distinguishes them from the processes employed by a less skilled writer.

In addition, the selection of the processes was made with reference to, in particular, Shaw & Weir's (2007) validation framework for writing tests and Khalifa & Weir's (2000) validation framework for reading tests. Although the two frameworks treat reading and writing separately with few claims of the interaction between the reading and writing process, they are particularly useful for the present study because (1) the work was a synthesis of earlier reading models and writing models as well as a more recent model by Field (2004) which is rooted in the information-processing principles of psycholinguistics; (2) the frameworks target the context of L2 reading and writing assessments; and (3) the frameworks have demonstrated their practicality for test validation purposes. The primary objective of this study is to identify cognitive parameters that are useful for the design and validation of

reading-into-writing tests. Table 2.3 below presents the proposed cognitive parameters to be analysed in this study.

Table 2.3 Cognitive parameters proposed to be analysed for the cognitive validity of reading-into-writing test tasks

Cognitive phases	Cognitive processes
Conceptualisation	Task representation
	Macro-planning
Meaning and discourse construction	High-level reading processes
	Connecting and generating
Organising	Organising
Low-level monitoring and revising	Low-level editing
High-level monitoring and revising	High-level editing

Task representation

This is a process whereby writers create an initial understanding of the rhetorical situation of the task for themselves (Flower et al 1990). Writing usually starts with the process of task representation. Writers tend to create a task representation by reading the task instruction (which usually contains information about the overall purpose of the test/assignment, structure of the test, time constraints, scoring criteria, word length) and task prompt (usually contains information about the topic of the task, genre and intended reader, rhetorical functions expected, e.g. describe, discuss, and details about input data, e.g. number of texts).

Task representation is an important process in writing. Grabe & Kaplan (1996) argued that writers set goals (planning) based on their understanding of the task. When writers approach the same task, they may choose to employ different processes based on the task representation they created and hence produce very different products. Flower et al (1990) found that undergraduates created task representation differently for the same reading-into-writing tasks in real-life academic contexts. The students had a different understanding of the same task in terms of primary sources of ideas, features of the text, organisational structure of the text, and strategies to use. The findings showed that students with more academic writing experience tended

to create a more accurate task representation than those with less academic writing experience. Ruiz-Funes (2001) studied the processes of how fourteen advanced-level L2 students produced an essay to discuss a literary text. Similar to Flower et al's study, Ruiz-Funes found that the writers created task representation of the same tasks differently. Her findings also revealed that a more cognitively complex representation of rhetorical style did not always lead to a text with the most complex textual structure.

Plakans (2010), as a follow-up to Plakans' study (2008), investigated ten undergraduates' task representations on reading-into-writing argumentative essays through think-aloud protocols. The findings revealed that the group of students who did not have much academic writing experience believed that all source texts needed close understanding. In contrast, the group of students who were more experienced with academic writing regarded the source texts as a tool to generate ideas and tended to read the source texts quickly.

Scardamalia & Paris (1985) investigated how advanced and novice writers may create task representation differently by analysing writers' retrospective recalls of the text they produced. They argued that the information beyond the content of the text reported in the recall protocols may reflect the task representation created by the writer. Their findings showed that almost 50% of the recall protocols from advanced writers reveal information other than the actual content of the text they produced. In contrast, the recall protocols of weaker writers mentioned primarily the text content. Although their method of investigating writers' task representation was indirect, their study reflected, at least to some extent, that when advanced writers approach a task they tended to generate a more complete and complex understanding of the task by considering issues such as writing goals, gist of the text (what to write), organisational structure of the text (how to write). Task representation, is therefore, an important process which can distinguish skilled writers from unskilled writers.

Macro-planning

Following Hayes & Flower's (1983) model which proposed that planning involves generating, organising and goal setting, Field (2004) divided planning into: macro-planning, organisation and micro-planning to explain the different purposes of the planning. Macro-planning is a process in which writers set goals, identify possible constraints, and decide where to gather ideas for task completion (Shaw & Weir, 2007). Writers typically consider the different issues, such as content, purpose of writing, target readership and genre during this process.

Scardamalia & Bereiter (1987) argued that unskilled writers who use the knowledge telling approach do not seem to employ planning or goal setting at the macro level because they write a text following a process where they retrieve relevant ideas from long-term memory. On the other hand, skilled writers who use the knowledge transforming approach tend to put explicit effort into macro-planning. Field (2004) shared a similar notion that skilled writers pay a lot more attention to planning than do unskilled writers. Eysenck & Keane (2005) further argued that it is the planning process which helps to differentiate skilled from unskilled writers. Similar results that L2 writers tend to plan less than L1 writers were reported by Hyland (2002). L2 writers are also found to encounter more difficulty in setting goals. Researchers have been attempting to reveal what writers actually plan by using think-aloud protocols. The results of Burtis, Bereiter, Scardamalia & Tetroe's (1983) study showed that the planning protocols of immature writers closely resemble the ideas presented in the text they produced. In contrast, the planning protocols of advanced writers consist of 'provisional ideas, goal statements, comments, and problem-solving attempts' (Burtis et al, ibid: 154). The process of macro-planning is believed to be influenced by contextual features of the task (Grabe & Kaplan, 1996; Shaw & Weir, 2007). It is necessary for test developers to know how the contextual setting in reading-into-writing tasks impacts on the task takers' macro-planning process.

Higher-level reading processes

On a reading-into-writing task, an important activity is to 'read' the input data. *Reading input data* is usually used in reading-into-writing studies as an umbrella term to reveal the occurrence (i.e. location and length) of input data reading processes during task completion. It is straightforward to identify when and for how long a writer reads the input data, yet it is much more important to investigate which reading processes are involved in reading-into-writing tasks. However, such empirical evidence is seemingly very limited in the literature.

Khalifa & Weir (2009), based on Weir (1983) and Urquhart & Weir (1998), identified two major types of reading: careful and expeditious reading which students perform in real-life conditions. Careful reading involves comprehension of every part of the whole text while expeditious reading means processing texts selectively, quickly and efficiently to access desired information from a text. Both careful and expeditious reading can be processed at global or local levels. Careful local reading involves primarily lower-level processes, such as *decoding at the word or phrase levels* and *establishing propositional meaning at the sentence level*. Careful global reading, on the other hand, is used to handle the majority information in the text(s), and thus involves the use of higher-level processes such as *linking propositions in building a mental model, inferencing, building a mental model, creating a text level representation* and *creating an intertextual representation.* Independent reading tests tend to test lower-level reading processes because a majority of the items focus on the microlinguistic level (for details see Weir (1990) for IELTS and TEEP; Urquhart & Weir (1998) for ELTS). Academic reading-into-writing test tasks should aim to elicit high-level reading processes from test takers, especially *creating a text level representation* and *creating an intertextual representation.*

Expeditious reading includes skimming, search reading and scanning. These processes are similar to the selecting process in the discourse synthesis model (Spivey, 1884, 1990, 1997; Spivey & King, 1989). Selecting is a process whereby writers select relevant ideas to put in the new text from the source

texts or from their own prior knowledge based on a set of criteria perceived as appropriate. Spivey (1991) argued that selecting plays an important role in constructing meaning because the meaning constructed by the writers is blueprinted by what they have selected from internal and / or external sources. Scardamalia & Bereiter (1987) argued that advanced and immature writers select ideas in significantly different ways. Immature writers 'select' idea units by running a test of appropriateness, a process which is similar to what Gomulicki (1956) described as an unconscious process of ranking elements according to importance when recalling knowledge. In contrast, advanced writers select content which is relevant to the task with conscious cognitive effort. Their selecting process is guided by explicit sets of criteria regarding the writing goal, appropriateness for intended reader, structure of text, available linguistic resources.

Previous studies have investigated what reading processes are tested by independent reading tests. Their results revealed that independent reading tests tend to be targeted at measuring careful local reading at the clause and sentence level rather than careful global reading, and rarely at expeditious forms of reading (Urquhart & Weir 1998, Khalifa & Weir 2009; Moore, Morton, & Price, 2010). However, there is seemingly a lack of discussion in the literature regarding the types of higher-level reading processes being elicited by academic reading-into-writing test tasks.

Connecting and generating

Connecting is a process in which writers generate links between ideas or new meaning by connecting ideas in the source texts with their own knowledge (Spivey, 1984, 1990, 1997). Barlett (1932) argued that a person cannot understand anything unless he or she 'connects something that is given with something other than itself' (p. 227). Spivey (1991) argued that writers *generate* new meaning by *connecting* content they *select* from source texts with knowledge they *retrieve* from memory, which can be world knowledge (Seifert, Robertson, & Black, 1985), topic knowledge (Pearson, Hansen, & Gordon, 1979) and/or schema of discourse knowledge (Rumelhart, 1975). The new 'meaning' generated can take the form of inferences of missing details

(Kintsch, 1974) or connections between present ideas from source texts and ideas stored in mind (Seifert, Robertson, & Black, 1985).

When writers compose from sources, they actually construct meaning from two sources: ideas selected from source texts and ideas retrieved from their prior knowledge. According to Scardamalia & Bereiter's (1987) model, the writing process of knowledge telling writers is a rather linear and straightforward process, from identifying probes, retrieving ideas which are relevant to the task from long-term memory, to putting down the ideas. Although this meaning construction process itself can be repeated many times, the primary purpose of knowledge telling writers is text generation, e.g. to generate enough ideas. They tend not to connect ideas provided in the source texts with their own prior knowledge deeply. In other words, neither the ideas presented in the source texts nor prior knowledge are likely to be reconstructed. Therefore, it is not uncommon for large chunks of verbatim copying to be found in unskilled writers' texts (for details see Cumming et al. 2005).

In contrast, regarding the knowledge transforming approach to writing, skilled writers constantly connect ideas selected from source texts with those retrieved from memory when they are solving the problem of what to write. In order to sort out what to write, skilled writers connect ideas from both sources to find out which ideas are most relevant or appropriate to the task context. However, these ideas from both sources are possibly repetitive or in a different order of importance according to the writers' goals. The process of organising (which will be discussed next) is usually activated to solve the problem of how to express these connected ideas. As a result of these processes, links between ideas or new representations of existing knowledge are always generated.

Despite the important role of connecting and generating in academic writing, such processes are largely underrepresented in the majority of academic writing tests.

Organising

This is a process in which writers organise the ideas to be put into the next text. Writers may order the ideas based on an evaluation of the relevance or

importance. They may also identify the relationship between different ideas, and to the overall text. Spivey (1991) argued that when writers read the source texts on a reading-into-writing task, they not only try to comprehend the ideas but also to organise relationships between these ideas for the text they are about to produce. Field (2004) argued that there is an abstract provisional organisation of ideas in the writers' mind. For example, if a task requires the writers to describe an event, most writers would have a sequential structure in mind. If the task asks writers to compare and contrast, most writers would have an advantages-vs-disadvantages structure in mind. These structures generated in writers' minds may or may not be the same as the source texts. Therefore, when organising ideas, some writers may retain a similar global structure as one presented in a single source text (Spivey, 1984) or generate a new structure in order to incorporate different idea chunks from multi-source texts (Spivey, 1991).

Scardamalia & Bereiter (1987) argued that immature writers who adopt the knowledge telling approach rely largely on a natural flow of writing without devoting much effort to organising the ideas. As knowledge telling writers aim at a smooth text generation, they are likely to put down the ideas in the same order as they were retrieved from memory (i.e. put down what appeared in mind first) or as they were presented in the source texts. In other words, this is similar to a process of 'dumping all (the) knowledge at once', which is identified as a strategy employed by writing-disabled students whose ability to organise is disrupted (cited in Cherkes-Julkowski, Sharp, & Stolzenberg, 1997:179). Johns & Mayes's (1990) study revealed that less proficient L2 writers tend to copy without organising the ideas they have selected from the source texts.

On the other hand, advanced writers approach a writing task as a problem-solving exercise in which they have to solve three major problems of 'what to write' created in *content problem space* as well as 'how to write it' and 'whom to write to' in *rhetorical problem space* (Graham, 2006: 460). When goals and constraints in one problem space interact with the other, a process of organising is activated in order to settle these goals (e.g. to compare A and B), constraints (e.g. lack of a structure of comparison) and resources (e.g. ideas

retrieved from memory and/or selected from source texts). Most advanced writers would 'transform' available ideas into their own text by explicitly ordering them, identifying relationships between them and/or determining which are central to their writing goals and so on.

It is interesting to note that Scardamalia & Bereiter (1987) warned that a cohesive text can also be produced by a mere knowledge telling process by 'skilled' immature writers whenever there is a smooth retrieval of ideas from memory by probes identified from the task prompt. Shaw & Weir (2007) argued that organising is an important process as it is closely related to the organisational requirements and assessment employed in most large-scale writing tests, e.g. FCE, CAE and CPE.

Micro-planning and translating

Field (2004) argued that writers conduct planning and organising not only at a macro level of text production, but also at the sentence and paragraph level. For micro-planning, writers plan for the part of the text that is about to be produced. At the paragraph level, writers plan for the goal of a particular paragraph and possibly align it with the previously formed macro-plans (e.g. writing purpose, genre, readership and overall structure) as well as the text produced so far. At the sentence level, writers plan for the structure of an upcoming sentence. Field (2004:329) argued that the actual text is produced based upon these micro-plans rather than macro-plans. Translating is the process whereby a writer's internal ideas are translated into linguistic forms. Shaw & Weir (2007) argued that the language translated needs to be not just lexically and syntactically appropriate but functionally appropriate as well. Field (2005) further pointed out that the cognitive demand of translating for L2 writers may be so high that the execution of other processes is hindered. Micro-planning and translating are two important phases in writing. However, when compared to other processes, micro-planning and translating seem to be more difficult to be reported reliably unless through directed verbal protocols (Field, 2004). Previous studies which investigated these two processes usually focused on these individual processes solely under experimental settings (see

Kellogg, 1994 for a review). These two processes will not be investigated in this study (for more explanation, see Chapter Five).

Monitoring and revising (editing)

Although monitoring appears to receive less attention in previous models, Field (2004: 330) argued that writers actually monitor at different levels of their text production at different stages of the process. Nevertheless, he suggested that most writers are likely to 'monitor at only one level at a time, either the sentence, the paragraph, the text so far' or the completed first draft. Therefore, it is likely that lower-level features such as accuracy of spelling, punctuation and syntax are monitored during the text production process while higher-level features such as development of arguments are usually monitored at a post-production stage. Field (2004) found that unskilled writers encounter difficulties in assessing rhetorical impact and locating possible areas of revision. In contrast, monitoring plays an important role in skilled writers' processing. Field (2005) further argued that many L2 writers do not monitor because they may not be able to assess their plans during translation due to the additional cognitive demands of spelling, syntax and lexical retrieval.

The process of revising is highly connected to the monitoring process as writers identify areas which need revising through the monitoring process. Although all areas identified will not necessarily be revised, it is very unlikely that revising occurs without monitoring. Generally speaking, there are two levels of revising. The higher-level relates to aspects such as meaning and coherence whereas the lower-level relates to accuracy or range of grammar, vocabulary and sentence structure. In the EAP context, writers are also concerned with plagiarism. In other words, writers will revise unsatisfactory parts where quotations are poorly made or where sentences are copied directly from source texts.

Many studies have compared the revising process employed by skilled and unskilled writers and found that skilled writers are better at revising than less skilled (Graham & Harris, 2000, 1996; Severinson Eklundh & Kollberg, 2003). Flower & Hayes (1980) found that fifteen per cent of the protocols made by skilled writers related to revising. Perl (1979) found that writers who followed

the knowledge transforming model made revisions about goals and main ideas. With the aid of key-stroke logging technology, systematic analysis of revisions is now possible. Severinson Eklundh & Kollberg (2003) used it to investigate how texts were constructed by skilled writers (a group of 10 university students) by tracking all the writers' revisions of the texts throughout the writing process. They found that these writers conducted revisions of different aspects: *creating a new global content unit, revising the contrast, revising for consistency, revising for coherence, revising for clarity, explicitness, or emphasis, revising to eliminate repetition, revising to emphasize text structure.*

In contrast, unskilled writers tend to devote little attention to revising. They usually revise at lower level, e.g. correcting errors and making small changes in wording (Fitzgerald, 1987; MacArther, Graham, & Harris, 2004). Perl (1979) found that some unskilled writers would make revisions at the sentence level, but the revisions tend to focus on '*the order of adding an introductory sentence, adding a conclusion, providing additional descriptive information, and inserting missing information'* (155).

Scardamalia & Bereiter (1987) further showed that based on data collected from a 6-week group instruction on writing, it was possible to increase the number of revisions made by immature writers, but not to improve the level of the revisions they made. It seems likely that it is the level of revising rather than the number of revisions that distinguishes unskilled writers from skilled writers.

Therefore, for the purpose of better distinguishing the processes employed by writers at different proficiency levels, it is important to investigate the level of the monitoring and revising processes elicited by the reading-into-writing test tasks.

2.5.2.2 Summary

A glossary of the processes to be analysed in this study is provided in Table 2.4.

Table 2.4 Working definitions of the cognitive processes

Cognitive processes	Working definitions
Task representation	• Create an initial understanding of the task (e.g. the overall purpose of the test/assignment, structure of the test, time constraints, scoring criteria, word length, topic, genre and intended reader, rhetorical functions to perform)
Macro-planning	• Plan for writing goals, content and organisation of the text. • Identify major constraints (genre, readership, language resources)
Higher-level reading	• Careful reading to create textual and / or intertextual representations • Search reading (e.g. select ideas which are relevant to the task context to put in the new text from the source texts based on a set of criteria perceived as appropriate)
Organising	• Organise the ideas to put in the next text (e.g. prioritize ideas in terms of relevance or importance, re-order, re-combine, delete, categorise, create new structure)
Connecting and generating	• Generate links between ideas or new meaning by connecting ideas/discourse features provided in the source texts with their own knowledge.
Micro-planning*	• Plan for the part of the text that is about to be produced
Translating*	• Translate abstract ideas into linguistic forms
Monitoring and Revising (editing)	• Higher-level: meaning and coherence, impact of reader • Lower-level: accuracy or range of grammar, vocabulary and sentence structure, plagiarism

*The processes of micro-planning and translating are not analysed in this study.

2.5.3 Criterion-related validity considerations

2.5.3.1 What is criterion-related validity?

Criterion-related validity, unlike contextual and cognitive validity, is a form of external evidence for the validity of a test obtained post hoc using statistical procedures. It is defined as 'a predominantly quantitative and a posteriori concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance with established properties' (Weir, 2005: 35). The comparison between test scores on the test to be validated and the external criterion of performance may be either concurrent or predictive in nature. Concurrent validity is usually examined by 'comparing scores from a given test with some other measure of the same ability of the test takers taken at the same time as the test' (Shaw & Weir, 2007: 229), whereas predictive validity involves comparing the test scores with an external measure of the same candidate's performance some time later, after s/he has taken the test. The most commonly used external measures include test scores, rating by teachers, test takers' self-assessment, and real-life academic results (for examle, see Bachman et al., 1995; Davies & Criper, 1988; Weir, Chan, & Nakatsuhara, 2013; Wu, 2012).

According to the test linking literature, this body of research has four major purposes, a) to link parallel test forms; b) to link tasks of the same construct but different format, lengths or levels, c) to generate unidirectional prediction from one task to a different task; and d) to link different tasks via rating using a common scale (see Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Linn, 1993; Mislevy, 1992). This study focuses on the third purpose, i.e. to investigate to what extent the reading-into-writing test scores relate to real-life academic scores. Nevertheless, establishing relationships between test scores and real-life performances is never an easy task. Previous studies, which will be reviewed in Section 2.5.3.2 below, showed contradictory findings regarding the correlations between test scores and real-life scores. One may question the value of establishing such criterion-related validity between test scores and real-life scores. However, for language tests which are used with a gate keeping function in the admission of students to university, it is necessary for test developers to demonstrate to what extent test scores relate to real-life

academic scores. Users of the test scores can then decide how to interpret the test scores as an indication of whether the test taker possesses a certain level of proficiency in English sufficient to cope with the language demands in the real-life academic context. The value of validating tests against suitable external criteria has been advocated by professional testing bodies such as the Association of Language Testers in Europe (ALTE) (see ALTE, 1998). Clapham (1988) urged that 'however difficult this [to check a test's external validity] may be, we clearly have to make the external criteria as valid and reliable as possible' (p.49).

As explained previously, the present study will put slightly more emphasis on establishing the a priori context and cognitive evidence of reading-into-writing tests than establishing the a posteriori criterion-related validity of reading-into-writing tests. This is because any results of the criterion-related validity of a test need to be built upon valid context and cognitive parameters in the performances instantiated by the test itself. Therefore, the discussion below will be shorter than the previous discussion on contextual parameters (Section 2.5.1) and cognitive parameters (Section 2.5.2).

2.5.3.2 Previous studies on criterion-related validity

This sub-section reviews the relevant studies on criterion-related validity in the literature and highlights a few issues related to research nature and sample size, results of correlations between overall test scores and academic outcomes, and results of correlations between writing test scores and academic outcomes.

Research nature and sample size

Previous studies of criterion-related validity between test scores and academic outcomes are predominantly quantitative. Most studies measure academic achievement by Grade Point Averages (GPA) (e.g. Allwright & Banerjee, 1997; Archibald, 2001; Avdi, 2011; Cotton & Conrow, 1998; Dooey & Oliver, 2002; Dooey, 1999; Feast, 2002; Green, 2005; Humphreys et al., 2012; Read & Hayes, 2003). Some used coursework grades (e.g. Ingram & Bayliss, 2007; Ushioda & Harsch, 2011). The sample size of the criterion-related validity studies varies widely from 17 (Read & Hayes, 2003) to thousands (Cho &

Bridgeman, 2012). However, most studies had a sample size less than a hundred.

Results of correlations between overall test scores and academic outcomes

The findings of the correlations between overall test scores and academic outcomes in the literature are contradictory. Some studies showed no significant correlations between the two measurements. Cotton & Conrow (1998), who studied 33 students, found no significant correlations between the participants' IELTS bands and the language difficulties they experienced in their course work. Ingram & Bayliss (2007) studied 28 non-native students' language behaviour by students' self-assessment, interview, teachers' rating and researcher's observation. They found that while the students were generally able to produce the target language, there was no significant relationship between their IELTS scores and course-related task scores. Other studies such as Kerstjens & Nery (2000) shared similar results. They investigated the correlations between 113 first-year international students' overall IELTS bands and their GPA, and reported a non significant correlation coefficient of 0.15. Dooey (1999) investigated the university admission threshold level of IELTS 6.0 in particular. The results showed no evidence that students who did not meet such level are more likely to fail.

Some others found significant, but usually low, correlations between the two measurements. Criper & Davies (1988), one of the earliest studies, found a correlation of 0.3 between participants' English Language Testing Service (ELTS) scores and their GPA. Two decades later, Feast (2002) similarly found a significant but weak regression coefficient of 0.39 for between IELTS scores and GPA, with 101 international students in an Australian university. Cho and Bridgeman (2012), in their study of 2594 students, also found a weak correlation between TOEFL iBT and GPA. Their study was one of the largest scale studies of this kind. Yen & Kuzma (2009) investigated 77 undergraduates at a British University and found that their IELTS scores correlated significantly with their first semester GPA 0.46 and their second semester GPA at 0.25.

Results of correlations between writing test scores and academic outcomes

On the other hand, some studies compared the predictive power of the individual macro-skills, i.e. reading, listening, speaking and listening. Generally speaking, receptive reading and listening test scores tend to have better correlations to academic outcomes than productive speaking and writing test scores. Writing test scores tend to have lower correlations to academic outcomes; that could also be because speaking and writing tests usually use a more limited point-scale than reading and listening tests. A limited range of scores has been seen as one of the problems of criterion-related validity studies (Ingram & Bayliss, 2007).

Similar to the results of studies into the predictive power of overall test scores, the results of studies comparing writing test scores and academic outcomes are inconclusive. Cotton & Conrow (1998) found no significant correlation between 17 students' IELTS writing scores and their GPA. Humphreys et al. (2010), who studied 51 students from different disciplines, found no significant correlation between the participants' IELTS Writing test scores and their academic grades in either first or second term. Ingram & Bayliss (2007) investigated the relationships between students' IELTS writing scores and their self-rating regarding the difficulty they experienced in completing essays and reports. The findings revealed no correlation between the two measurements. However, studies with a larger sample size tend to report significant results. The earlier mentioned Kerstjens & Nery's (2000) study found a significant correlation of 0.25 between IELTS writing scores and GPA. Ushioda & Harsch (2011) found that IELTS Writing explained over 33% of the variance in academic coursework grades. While these results comparing writing test scores and academic outcomes are inconclusive, they have suggested the range of correlations found between writing tests and academic outcomes.

Implications from previous studies

While the above mentioned studies have provided insights into the relationships between test scores and academic outcomes, the use of GPA as a

measurement of academic outcomes has been criticised. Many researchers argue that final academic success arguably depends on a range of non-linguistic factors rather than solely on language proficiency (Davies & Criper, 1988; Ingram & Bayliss, 2007; Weir, 2005). Affective factors/motivation, learning strategy and social-cultural factors are the most commonly discussed factors in the literature (See for example, Cotton & Conrow, 1998; Ingram & Bayliss, 2007; Kerstjen & Nery, 2000). It is beyond the scope of this study to discuss these non-linguistic factors. However, the limitations of using GPA as the measurement of academic outcomes mean that it should be avoided. This study therefore will select more relevant external measurements of test-takers' real-life writing performance than using their GPA.

Secondly, previous studies tended to focus on the overall test scores, which are reasonable for their own research purposes, e.g. establishing the general degree of accuracy against external (final) measurements. However, information regarding individual papers is more useful for the purpose of test improvement. As reviewed above, only a limited number of studies investigated the correlations between writing test scores and academic outcomes. And none of them investigated the predictive power of writing test tasks which involve multiple verbal inputs or multiple verbal and non-verbal inputs.

While reading-into-writing tasks have been in use for a long time (see Section 2.3), the criterion-related validity evidence of reading-into-writing tests is scarce in the literature. The present research aims to examine the extent to which reading-into-writing tasks can provide predictive information about students' writing ability in a real-life academic context and if reading-into-writing tests would provide any new information which has not been revealed by writing-only tests.

2.6 Research Questions

In light of the literature reviewed in this chapter, the three main research questions to be addressed in this study are:

1. What are the contextual characteristics of the academic writing tasks that students would normally encounter in real life? To what extent do the reading-into-writing test tasks resemble these contextual features under test conditions?

2. What are the cognitive processes that students would normally employ to complete the real-life academic writing tasks? To what extent do the reading-into-writing test tasks elicit these cognitive processes from test takers?

3. To what extent can performances on the reading-into-writing tests predict test takers' ability to perform on real-life academic writing tasks?

Building on Weir's (2005) socio-cognitive framework, the present study attempts to establish the construct validity of integrated reading-into-writing tests by investigating their context, cognitive and criterion-related validity. This study first of all investigates the contextual features of real-life academic writing tasks and the cognitive processing employed to complete these tasks, in order to define the qualities of a valid task in assessing academic writing ability. It then examines the contextual features of two types of reading-into-writing test tasks (*essay with multiple verbal inputs* and *essay with multiple verbal and non-verbal inputs*) and the cognitive processes elicited by them. The results collected from both the real-life and test conditions are compared to provide empirical evidence of the contextual and cognitive validity of reading-into-writing tests in assessing academic writing ability.

In other words, RQ1 investigates whether the characteristics of the reading-into-writing tasks are an adequate and comprehensive representation of those that would be normally encountered in the real-life context. RQ2 investigates

whether the cognitive processes required to complete the reading-into-writing test tasks sufficiently resemble the cognitive processes which a test taker would normally employ in non-test conditions, i.e. are they construct relevant (Messick, 1989). Is the range of processes elicited by the test tasks sufficiently comprehensive to be considered representative of real-world behaviour, i.e. not just a small subset of those which might then give rise to fears about construct under-representation? Finally, comparisons are made between performances on the real-life tasks and the reading-into-writing tests to find out the extent to which performances on reading-into-writing tests predict test takers' ability to perform on real-life academic writing tasks (RQ3).

3 METHODOLOGY

3.1 Introduction

Firstly this chapter describes how the two academic writing tasks and the two reading-into-writing test tasks were selected as being representative of real-life academic and test tasks in this study, and describes the basic features of these tasks (Section 3.2). The chapter then describes the research design of this study with respect to the investigation of the three components of language test validity: context validity (Section 3.3), cognitive validity (Section 3.4) and criterion-related validity (Section 3.5). Each of these research design sub-sections presents the details of participants, data collection methods and instruments, data collection procedures and methods of data analysis (See Table 3.1 for an overview of the study). A summary of the chapter is provided in Section 3.6.

Table 3.1 Overview of the study

Focus	Data collection in both real-life and test conditions	Data analysis
Context validity (RQ1)	• Investigated the overall task setting by **expert judgement** (n=10) on 7 categories, i.e. *purpose*, *topic domain*, *genre*, *cognitive demands*, *rhetorical functions*, *intended reader* and *knowledge of criteria* • Investigated the input text features by **expert judgement** (n=2) on 7 categories, i.e. *input format*, *verbal input genre*, *non-verbal input*, *discourse mode*, *concreteness*, *textual organisation* and *cultural specificity*, and **automated textual analysis** (17 indices measuring lexical complexity and syntactic complexity and degree of cohesion)	• Descriptive analyses of the expert judgement results • Non parametric independent sample tests to compare the automated textual indices between a) real-life input texts and Test Task A (with multiple verbal inputs) input texts, and, b) between real-life input texts and Test Task B (with multiple verbal and non-verbal inputs) input texts
Cognitive validity (RQ2)	• Investigated the cognitive processing elicited in both the real-life and test conditions by a **Cognitive Process Questionnaire** (a total of 443 questionnaires were collected from 219 participants – 70 on real-life task A, 73 on real-life task B, 160 on reading-into-writing Test Task A, 140 on reading-into-writing Test Task B)	• Exploratory factor analyses to show the underlying structure of the cognitive processes elicited in each condition • Non parametric independent sample tests to investigate whether the cognitive parameters distinguished the processes employed by high-achieving and low-achieving participants in each condition. • Non parametric related-sample tests to investigate whether the cognitive processes elicited by the each of the two reading-into-writing tests resembled the processes employed by the test takers (whole population and in groups of proficiency) in the real-life conditions
Predictive validity (RQ3)	• Collected the participants' performances on real-life task A, real-life task B, two other real-life tasks (i.e. question-and-answer test and case-study exam) and the two reading-into-writing test tasks	• Descriptive analyses of the performances on each task. • Analyses of the correlation between performances on the real-life tasks and Test Task A, and between the real-life tasks and Test Task B

3.2 Reading-into-writing tasks

3.2.1 Real-life academic writing tasks

The first step to understanding what makes a reading-into-writing task valid for assessing academic writing was to investigate what tasks are used in the target real-life academic context (Hamp-Lyons & Kroll, 1996). This sub-section describes the real-life academic writing tasks used in this study. Based on the literature review in the last chapter, it is now clear that academic writing tasks almost always involve the use of reading materials as well as the writer's internal resources, and that high-level knowledge transformation processes play an important role in academic success (See Section 2.2). It is beyond the scope of this thesis to conduct a comprehensive survey of academic tasks as previous studies have done (See Section 2.2.1 for a review of these studies). Nevertheless, for the purpose of this empirical validation study, it was first of all necessary to sample the real-life academic writing tasks within the local context of the study.

To the researcher's knowledge, the present thesis is the first EAP reading-into-writing validation study covering both contextual and cognitive validity as well as predictive power. It was felt appropriate that the present study limits its scope to a single British university and to a single discipline. As mentioned in Section 1.1.2, University of Bedfordshire (UoB) is one of the twenty largest recruiters of international students (UKCISA, 2012). The Business discipline was chosen for the purposes of investigation because Business and Administrative Studies contained the highest proportion of international students in the UK (36%) (UKCISA, 2012). In addition, the Business School at the UoB is the largest Business School in the region with six departments, 4000 students and over 100 teaching and research staff (University of Bedfordshire, 2013).

Sampling the predominant real-life tasks started with a survey of the writing task types assigned to students at the Business School of UoB. Eight module handbooks were collected from the University's Business School. The specifications and assessment plan of each module were assessed according to the following criteria:

- **Enrolment rate of the module**

  The Business School offers students a choice of different selective modules to complete their programme. In order to guarantee an adequate sampling of the number of students of each module, modules which had an enrolment rate of less than 50% of the population were not considered.
- **Individual writing assignments only**

  Group/collective assignments were sometimes used in the Business School but this thesis focuses on individual writing tasks only. These tasks were therefore not considered.
- **The type of input source involved**

  The type of input source involved in each task was examined. All tasks involved multiple input types.
- **The type of response text (genre) required**

  Essay and report were apparently the most common genres assigned in the modules. Repeated genres were not considered.

As a result, two writing tasks were selected from two different modules to represent the predominant real-life tasks in this study (see Appendix 3.1.1 and 3.1.2 for the two real-life tasks). Table 3.2 describes the basic features of these two real-life tasks. To define the target constructs of a valid academic writing task, the two real-life tasks were analysed in terms of their contextual and cognitive parameters as well as students' performances (See Section 3.3, 3.4 and 3.5 for details of the research design of each component).

Table 3.2 Basic features of these two real-life tasks

Real-life tasks	Genre	Task instructions	Input	Output
A	Essay	Write an essay on a given topic - Summarise salient issues - Discuss the issues with justified personal views	• Verbal (a stimulus article) • Non-verbal • Students are expected to used other input texts of their choice	5000 words
B	Report	Write a report to forecast the business of a company - Describe the data - Discuss and justify ways of analysis - Make recommendations	• Verbal • Non-verbal (a numeric dataset) • Students are expected to used other input texts of their choice	2000 words

A small corpus of the real-life input texts

In the literature, there are a few studies investigating the features of texts students have to read in an academic context (e.g. Green, Unaldi, & Weir, 2010; Weir, Hawkey, Green, Unaldi, & Devi, 2009). However, there is seemingly insufficient discussion regarding the difficulty level or features of the texts students read in order to complete writing tasks in an academic context. A reasonable start was to build a small corpus of real-life input texts by sampling from the texts students read to complete the two selected real-life tasks by the following steps:

1. 100 student scripts were collected from each of the selected real-life tasks, totalling 200 students' scripts.
2. The bibliography of each script was examined.
3. The ten most cited source texts from each of the two real-life tasks were identified.
4. Three extracts (from the beginning, middle and end of each text) were obtained from each of the twenty selected source texts. Each extract was about 500 words long.

5. A small corpus of real-life input texts, containing 60 extracts from the 20 most cited source texts, was computed.

3.2.2 Reading-into-writing test tasks

In order to investigate how closely the contextual and cognitive parameters of real-life academic writing tasks are being represented in the reading-into-writing tests, a range of reading-into-writing test tasks from current large-scale language tests were collected (See Section 2.4.3 for a review of these reading-into-writing test tasks). This sub-section describes the reading-into-writing test tasks used in this study. The tasks collected included IELTS Academic Task 1 (with a non-verbal input), PTE Academic – Summarise Written text (with a single short verbal input), TOEFL iBT Integrated Writing (with verbal and listening inputs), GEPT Advanced Writing Task 1 (with multiple verbal inputs) and Task 2 (with multiple non-verbal inputs), and Trinity Integrated Skills in English (ISE) Level IV Task 1 (with multiple verbal and non-verbal inputs) and Task 2 (with multiple verbal inputs). The suitability of these tasks for this study was assessed based on:

- **Purpose of the test**
  Reading-into-writing test tasks which are not for academic purposes were excluded.
- **Literature gap / use of multiple sources**
  As reviewed in Section 2.3, reading-into-writing involving multiple input sources have largely been neglected in the literature. Preference was then given to tasks which involve multiple sources.
- **Access to authentic test tasks**
  As the purpose of this thesis is to establish the construct validity of reading-into-writing tests, it was felt that using real test tasks, i.e. not practice tasks, would be more suitable.
- **Access to standardised scoring**
  As RQ3 involves analysis of test scores, access to standardised scoring was necessary.

As a result, the GEPT Advanced Writing Task 1 (hereafter Test Task A) was selected because this task type (i.e. summarising multiple verbal source texts

and giving personal opinions and justifications) has not received much attention in the literature and met the above criteria (for Test Task A, see Appendix 3.1.3). To the knowledge of the researcher, the cognitive processes elicited by such a task type have not been investigated under authentic test conditions. In addition, the researcher was able to obtain agreement from the examination board to supply authentic test tasks and provide standardised scoring of the scripts. Ten testlets of Test Task A were used in the contextual analysis of this study. Two passages were collected from each of the ten testlets of the test, resulting in a collection of twenty Test Task input texts.

Based on the results of sampling the real-life tasks in the context of this study, tasks with multiple verbal and non-verbal inputs were common in the target academic writing context. However, when this study was conducted, there seemed to be no publicly available standardised EAP reading-into-writing test tasks incorporating such a task type. Therefore, the second test task (hereafter Test Task B) for the study was chosen from a newly developed University in-house diagnostic test (for Test Task B, see Appendix 3.1.4). As mentioned in Section 1.1.2, in the local context of this research, there was a need to assess newly arrived international students' academic writing abilities by a valid diagnostic test at the University of Bedfordshire. The reading-into-writing task which requires the use of multiple verbal and non-verbal inputs was developed by the Centre for Research in English Language Learning and Assessment (CRELLA) to be used in the University's diagnostic test[3] (The researcher was part of the test development team). As Test Task B was a newly developed test, at the time of the study, only one operationalised testlet was available.

The basic features of the two selected reading-into-writing test tasks are provided in Table 3.3.

[3] Details of the test development and test specifications are available in a separate construct document (CRELLA, forthcoming)

Table 3.3 Basic features of Test Task A and Test Task B

Test task	Test instructions	Time allowance	Input	Output	Function	Level
A	• Write a comparative essay summarising the main ideas from the verbal inputs and stating own viewpoints	60 minutes	2 articles without non-verbal input	At least 250 words	Criterion-referenced level specific test	CEFR C1
B	• Summarise the main ideas from both verbal input and non-verbal inputs and express opinions	60 minutes	2 articles with a non-verbal input each	180-200 words	University in-house diagnostic test	CEFR B2

Section 2.3 has described the procedures of selecting the two real-life academic writing tasks and two reading-into-writing test tasks to be used in this study. These four tasks will be used in the investigation of the three validation components described in Sections 3.3, 3.4 and 3.5 below.

3.3 Research design: context validity

Context validity of a writing task addresses the particular performance conditions, the setting under which it is to be performed (such as purpose of the task, time available, length, specified addressee, known marking criteria as well as the linguistic demands inherent in the successful performance of the task) (Shaw & Weir, 2007: 64). This sub-section describes how contextual features of the real-life and reading-into-writing test tasks were analysed in this study.

3.3.1 Participants

Ten judges with experience in language testing and/or language teaching were recruited to form an expert panel to evaluate the contextual features of the real-life tasks and test tasks. Figure 3.1 presents the profile of the judges' experiences.

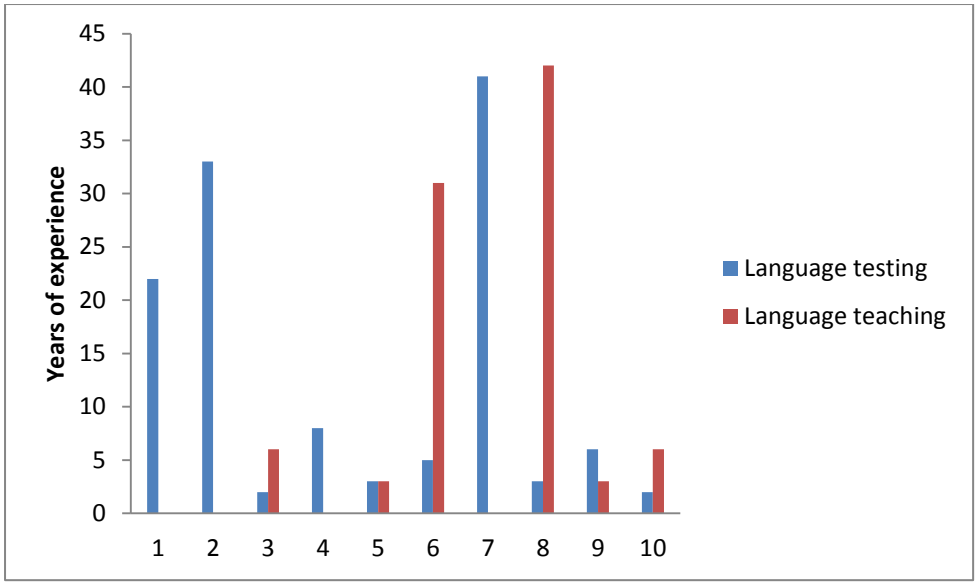


Figure 3.1The profile of the judges' experiences

3.3.2 Data collection methods and instruments

3.3.2.1 Contextual parameter proforma for expert judgement

A Contextual Parameter Proforma was developed to address the task variables reviewed in Section 2.5.1. The proforma aimed to facilitate an operationalisable analysis of the contextual variables of the reading-into-writing tasks by a group of judges. Two pilot studies informed the development of the proforma.

First pilot

The first version of the proforma was drafted based on Shaw & Weir's (2007) contextual proforma for writing tests and Wu's (2012) contextual proforma for reading tests. A range of categories which are appropriate to the context of this study were chosen. Other materials such as Chapter Two of Weigle's Assessing Writing (2002) and the CEFR grid for writing tasks (ALTE, 2011) were consulted to develop the choice under each category. The draft proforma in this study included 10 categories of contextual feature: *genre, purpose, topic domain, rhetorical task, cognitive demands, cultural specificity, input text abstractness, writer-reader relationships, language functions* and *knowledge of criteria*. It was piloted with 10 judges (7 language testing experts, 1 Business subject lecturer, 2 language teachers). They were

instructed to use the proforma to analyse one real-life task and one test task individually. Verbal feedback regarding the effectiveness of the proforma and their experience of the analysis was collected. Changes were made to address the following concerns raised in the pilot:

(1) Regarding *genre, topic domain, rhetorical task* and *language functions,* there was confusion about which part of the tasks these categories should be applied to, e.g. the prompt, the input texts, or the output text. It was decided that the categories would be divided into two sections: *overall task setting* (considering the prompt and the output) and *input texts features*. New categories were added to address the contextual features of the input texts. As a result, the second version of the proforma in this study included a total of 14 categories. Categories 1 to 7 address the *overall task setting* which included *genre, purpose, topic domain, cognitive demands, clarity of intended reader, knowledge of criteria* and *language functions to perform*. Categories 8 to 14 address variables of the *task input demands*, which cannot be effectively analysed by automated textual tools (The procedures of automated textual analysis are described in Section 3.3.3). They were, namely, *input format, verbal input genre, non-verbal input type, discourse mode, concreteness of ideas, explicitness of textual organisation* and *cultural specificity*.

(2) Regarding the *writer-reader relationship*, respondents were asked to identify the relationship between the writer and intended reader(s). However, it was felt that asking respondents to rate the clarity of the intended reader presented in the task could be more effective. In the second version, respondents were asked to rate how clearly the intended reader was presented in the task by using a 5-point Likert scale.

(3) Regarding the *language functions*, some functions were felt to be too similar, e.g. *justifying* and *reasoning*. *Reasoning* was kept. Some respondents expressed confusion between *summarising* and *synthesising*. An explanation of synthesising was added. The respondents also identified additional language functions, such as *predicting, citing sources* and *illustrating visuals*.

(4) Respondents had difficulty in analysing the *topic domain* of the task. In most cases, respondents felt that one task fell into two or even three domains.

They also felt that this category was more subjective, i.e. involving more personal perception, than the others. It was decided that instead of choosing one topic domain, the judges would be asked to rate the degree to which the task falls into each domain.

(5) After completing the analysis individually, the respondents were asked to discuss their responses with the whole group. While the discussion was engaging and insightful, it might not be the most effective way as the discussion was sometimes dominated by a few respondents. It was decided that, in the main study, the judges would be asked to share and discuss their individual responses in pairs and to fill in a separate proforma for their agreed responses.

(6) The judges recommended that the entire expert judgment exercise should not take more than three hours to maintain the effectiveness of the judgement. As each real-life task had 10 selected input texts (the selection procedures was presented in Section 3.2), it was felt that it would not be feasible for the ten judges to analyse all input texts within a three-hour slot. Considering the fact that the input texts will also be analysed by automated textual tools (see Section 3.3.2.2), a solution generated by the feedback discussion of the first pilot was that, in the context of this study, it was deemed advantageous to have a separate meeting to analyse the input texts with only two of the judges.

Second pilot

The second version of the proforma was piloted with 2 judges. They were asked to analyse all four tasks (2 real-life tasks and 2 test tasks). While the issues raised in the first pilot were resolved, it was felt that the provision of a glossary of the analysis categories (see Appendix 3.2) would facilitate the researcher's verbal explanation. Recommendations of analysing the contextual features of reading-into-writing tasks for future research will be provided in Chapter Seven.

The finalised *Contextual Parameter Proforma* (see Table 3.4) consisted of two parts: *overall task setting* and *input texts*. Items 1 to 7 address the overall task setting which includes *genre, purpose, topic domain, cognitive demands,*

intended reader, *knowledge of criteria* and *language functions to perform*. Items 8 to 14 address input text features, which cannot be effectively analysed by automated textual analysis tools. The variables include *input format*, *genre*, *discourse mode*, *concreteness of ideas*, *concreteness of textual organisation* and *cultural specificity.*

Table 3.4 Contextual Parameter Proforma

Part 1 - Overall task setting					
1. Purpose	1 Unclear	2	3	4	5 Clear
2. Topic Domain (Please circle a rating for each domain)	Personal Not at all Definitely 1 2 3 4 5	Social Not at all Definitely 1 2 3 4 5	Academic Not at all Definitely 1 2 3 4 5	Professional Not at all Definitely 1 2 3 4 5	
3. Genre	Essay	Report	Case Study	Summary	Others (Please specify):
4. Cognitive demands	1. Telling personal experience / viewpoints	2. Summarising / organising given ideas	3. Transforming given ideas into new representations		
5. Language functions to perform (you may choose more than 1)	Classifying	Citing sources	Describing	Defining	Evaluating
	Persuading	Predicting	Recommending	Reasoning	Summarising
	Synthesising (to combine different (parts of) texts to form a new text with own interpretations)	Expressing personal views	Illustrating visuals	Others (Please specify):	
6. Clarity of intended reader	1 Unclear	2	3	4	5 Clear
7. Knowledge of criteria	1 Unclear	2	3	4	5 Clear

Part 2 - Input text features						
8. Input format	Single verbal	Single non-verbal	Multiple verbal	Multiple non-verbal	Multiple verbal and multiple non-verbal	
	Others (Please specify):					
9. Verbal input genre	Book Chapter	Journal article	News / Magazine article	Proposal	Report	Review
	Others (Please specify):					
10. Non-verbal input	Table	Graph	Diagram	Picture		
11. Discourse mode (Consider the primary purpose of the text)	Narrative	Descriptive	Expeditionary	Argumentative		
12. Concreteness of ideas	1 Abstract	2	3	4	5 Concrete	
13. Explicitness of textual organisation	1 Inexplicit	2	3	4	5 Explicit	
14. Cultural specificity	1 Neutral	2	3	4	5 Specific	

3.3.2.2 Automated textual analysis tools

In addition to expert judgement, automated textual analyses were performed to analyse a range of textual features of the input texts in this study. Automated textual analysis has been regarded as a more systematic and efficient way to assess textual features than the more traditional expert judgement method, especially when a large number of texts are involved. Many researchers have used automated textual analytic tools to evaluate the features of different types of texts such as L1 students' scripts (e.g. Crossley & McNamara, 2010), L2 students' scripts (e.g. Crossley & McNamara, 2012), reading materials (e.g. Crossley, Louwerse, McCarthy, & McNamara, 2007; Green, 2012), undergraduate reading texts (e.g. Green et al., 2010), reading texts in language tests (e.g. Green et al., 2012; Wu, 2012), and L2 test takers' scripts (e.g. Weir, 2012)

In this study, two automated textual analysis tools were chosen - CohMetrix version 2.1 (Graesser, McNamara, Louwerse, & Cai, 2004) and VocabProfile version 3 (Cobb, 2003). CohMetrix was used in all the above-mentioned textual analysis studies. It is one of the most popular textual analysis tools in the literature. It is freely assessable on the Internet and it produces a very comprehensive list of about 60 textual indices. More importantly, CohMetrix was designed to explore attributes of cognitive language use. Graesser, McNamara & Kulikowich (2011) argued that CohMetrix's automated indices measure 'deep-level factors of textual coherence and processing' (223). VocabProfile (Cobb, 2003) is another popular textual analysis tool which provides a profile of texts in terms of different vocabulary frequency bands based on BNC (The British National Corpus, 2007) (e.g. the most frequent 1000 words) and different types of vocabulary (e.g. academic words based on Coxhead, 2000). The tool has been used to assess the difficulty level of reading texts in many studies.

Both tools have been used in the testing literature. For instance, Green et al (2010) compared IELTS reading texts and undergraduate texts at British universities, Green (2012b) investigated reading texts targeted at different levels of the Common European Framework of Reference for Languages (Council of Europe, 2001), Green et al (2012) investigated the features of reading texts in CAE, and Wu (2012) compared Cambridge Main Suite and GEPT Taiwan examinations at the B1 and B2 levels. Weir (2012) investigated features of the test takers' scripts of the TEAP test in Japan.

While CohMetrix and VocabProfile allow researchers to automate a large number of textual indices in an objective and reliable way, the results have to be interpreted with caution. Researchers have argued that not all indices produced are equally useful or interpretable. Green et al (2012) criticised the fact some of the indices seem to overlap and Green (2012b) attempted to identify those indices which are helpful to distinguish texts between adjacent CEFR levels.

It is seemingly important for individual researchers to establish which of the indices are helpful in their context of study. Green et al. (2012) showed that 17

CohMetrix and 2 VocabProfile indices meaningfully exhibited significant differences in the reading texts across three levels of Cambridge examinations: FCE (B2), CAE (C1) and CPE (C2). Based on 25 CohMetrix indices and 6 VocabProfile indices, Wu (2012) compared the features of reading texts between the GEPT and Cambridge examinations at B1 and B2 levels. Weir (2012) found that 12 CohMetrix indices were useful in establishing criterial differences in the L2 test takers' scripts rated at the A2 and B1 levels. Green et al. (2010) compared the features of undergraduate texts and IELTS reading texts by 19 CohMetrix and 5 VocabProfile indices.

Drawing upon previous studies, especially those looking at reading texts (e.g. Green et al., 2010; Green et al., 2012 and Wu, 2012), the usefulness of the all CohMetrix and VocabProfile indices were examined by the researcher in a pilot analysis. 30% of the real-life input texts were analysed in the pilot analysis. Based on the results of the pilot analysis, 13 CohMetrix and 4 VocabProfile indices were selected to analyse the features of the input texts and draw comparisons between the real-life and reading-into-writing test tasks (See Table 3.5 for a glossary of the selected indices). The selection of the indices in this study was similar to the previous studies. However, it was considered more appropriate to categorise the selected indices in terms of lexical complexity, syntactic complexity and degree of cohesion for the context of this study (i.e. reading-into-writing tasks for academic purposes), rather than categories such as vocabulary, grammar, readability, cohesion and text abstractness used in Green et al's (2010) study. The list of the deleted indices and reasons for deletion are presented in Table 3.6.

Table 3.5 Selected automated textual indices

Contextual index	Definition (Extracted from the official documents of the two tools)	Automated analysis tool
Lexical		
High frequency words (K1)	The ratio of words which appear in the first most frequent 1000 BNC (2001) wordlist to the total number of words per text	VocabProfile
High frequency words (K2)	The ratio of words which appear in the second most frequent 1000 BNC (2001) wordlist to the total number of words per text	VocabProfile
Academic words	The ratio of words which appear in the Academic Wordlist (Coxhead, 1998) to the total number of words per text	VocabProfile
Low frequency words (Offlist)	The ratio of words that do not appear in either the most frequent 15000 BNC wordlist to the total number of words per text	VocabProfile
Log frequent content words	The log frequency of all content words in the text	Cohm 46
Average syllables per word	The mean number of syllables per content word, a ratio measure	Cohm 38
Type-token ratio (content words)	The number of unique words divided by the number of tokens of these words	Cohm 44
Syntactic		
Average words per sentence	The mean number of words per sentence	Cohm 37
Sentence syntax similarity	The proportion of intersection syntactic tress nodes between all sentences	Cohm 56
Mean number of modifiers per noun-phrase	The mean number of modifiers per noun-phrase	Cohm 41
Mean number of words before the main verb	The mean number of words before the main verb of the main clause in sentences	Cohm 43
Logical operator incidence	The incidence of logical operations (i.e. connectives), such as *and*, *or*, *not*, *if*, *then*, etc	Cohm 26
Cohesion		
Adjacent overlap argument	The proportion of adjacent sentences that share one or more arguments (i.e. noun, pronoun, noun-phrase) or has a similar morphological stem as a noun	Cohm 16
Adjacent overlap stem	The proportion of adjacent sentences that share one or more word stems	Cohm 17
Adjacent overlap content word	The proportion of content words in adjacent sentences that share common content words	Cohm 58
Proportion of adjacent anaphor references	The proportion of anaphor references between adjacent sentences	Cohm 18
Adjacent semantic similarity (LSA)	The measure of conceptual similarity between adjacent sentences	Cohm 27

Table 3.6 List of deleted CohMetrix indices and reasons for deletion

	Reasons	CohMetrix indices deleted
1	They overlapped with other indices – they showed very similar or no difference in results of other indices. For instance, 40 (Flesch-Kincaid grade level) and 39 (Flesch reading ease) are already covered by 38 (Average syllables per word and 37 (Average words per sentence). Celex measures are covered by word frequency measures.	• Cohm 24 Number of conditional expressional, incidence score • Cohm 25 Number of negations, incidence score • Cohm 34 Number of sentences • Cohm 39 Flesch Reading Ease • Cohm 40 Flesch-Kincaid Grade Level[4] • Cohm 55 Sentence syntax similarity adjacent • Cohm 57 All within paragraphs
2	They were difficult to interpret in terms of text complexity. For instance, 23 (ratio of pronouns to noun phrase) is affected by text type.	• Cohm 12 Incidence of negative additive connections • Cohm 13 Incidence of negative temporal connections • Cohm 14 Incidence of negative causal connections • Cohm 23 Ratio of pronouns to noun phrase • Cohm 30 Personal pronoun incidence score • Cohm 36 Average sentences per paragraph
3	They were not applicable to the data in this study due to the sampling procedures of the real-life texts.	• Cohm 33 Number of paragraphs • Cohm 35 Number of words
4	They were not useful or effective in determining the complexity of a text because a) the index produced a score which is difficult to interpret; b) insufficient explanation was provided in the CohMetrix menu; and/or c) results obtained contradict human judgment.	• Cohm 7 Incidence of causal verbs, links and particles • Cohm 8 Ratio of causal particles to causal verbs • Cohm 9 Incidence of positive additive connectives • Cohm 10 Incidence of positive temporal connectives • Cohm 11 Incidence of positive causal connectives • Cohm 15 Incidence of all connectives • Cohm 19 Argument Overlap, all distances • Cohm 20 Stem overlap all distances unweighted

[4] Most previous studies reported a scale of 0-16 but Coh-metrix provides the scale of 0-12.

		• Cohm 21 Anaphor reference, all distances • Cohm 22 Noun phrase incidence score (per thousand words) • Cohm 28 LSA all sentences combination mean • Cohm 29 LSA paragraph to paragraph mean • Cohm 31 Mean hyponymy values of nouns • Cohm 32 Mean hyponym value of verbs • Cohm 42 Higher level constituents per word • Cohm 45 Celex, raw, mean for content words • Cohm 47 Celex raw minimum in sentence for content words • Cohm 48 Celex, logarithm, minimum in sentence for content words (0-6) • Cohm 49 Concreteness, mean for content words • Cohm 51 Incidence of negative logical connectives • Cohm 52 Ratio of intentional particles to intentional content • Cohm 53 Incidence of intentional actions, events and particles • Cohm 54 Mean of tense and aspect repetition scores • Cohm 59 Mean of location and motion ratio scores • Cohm 60 Concreteness

3.3.3 Data collection procedures

Regarding expert judgement of the contextual features between the real-life academic writing tasks and the reading-into-writing test tasks, the judges were trained prior to the panel meeting with the adapted Familiarisation and Specification procedures (Council of Europe, 2009) as below:

1. The researcher explained the Contextual Parameter Proforma (Part 1 - overall task setting only) to the judges. An explanation sheet of the analytical categories was provided (See Appendix 3.2). The judges sought clarification of any unclear points.
2. The judges were grouped into pairs.
3. The judges were assigned to analyse one of the four tasks (two real-life and two reading-into-writing test tasks) individually and filled in the Proforma. The order of the tasks assigned to each pair was counter-balanced. After they had completed the individual analyses, they discussed their responses in pairs. They were asked to record the reasons for any disagreement. They then filled in another Proforma to record their agreed responses. The judges handed in their responses (both individual and pair) to the researcher.
4. The judges analysed the other three tasks one by one following Step 3.
5. The judges completed the Feedback Evaluation Questionnaire (See Appendix 3.3).

As explained earlier, a second expert judgement meeting focusing on the input texts was conducted. The input text analysis involved 1) ten sample texts for each of the two real-life tasks, and 2) two passages from ten testlets of Test Task A and two passages from one testlet of Test Task B, totalling 20 real-life input texts and 22 test task input texts in all. Following the same procedures detailed above, the two judges used Part 2 (*Input text features*) of the Contextual Parameter Proforma (See Table 3.4) to analyse all the corresponding input texts from the four tasks (two real-life and two reading-into-writing tasks). They analysed each input text individually and then in pairs.

The input texts were also analysed by automated textual tools. 60 extracts from the 20 real-life input texts (see Section 3.2.1), 20 passages from 10 testlets of Test Task A, and two passages from one testlet of Test Task B were analysed using the 13 CohMetrix and 4 VocabProfile indices (see Table 3.5) by the researcher.

3.3.4 Method of data analysis

The purpose of the contextual analysis in this study was to investigate what features are important in determining the level of a reading-into-writing task, so that the results can be used to inform decision making in task development as well as *a priori* validation procedure.

The expert judgement responses on the 14 categories regarding the overall task setting and input text features were reported. For the classification categories (i.e. Items 3-5, 8-11), results of the percentage of each option were presented. For the rating categories (i.e. Items 1-2, 6-7, 12-14), the mean and standard deviation on the five-point Likert scale were presented. Descriptive comparisons were made between the real-life tasks and the test tasks. Descriptive statistics instead of inferential statistics were used due to a small sample size. Graphic presentation of the data was provided for further illustration where necessary (see Section 4.2 for the results of the overall task setting and Section 4.3.1 for the results of the input text features by expert judgement).

Regarding the automated textual analyses of the input text features, the mean and standard deviation of the 17 selected indices were obtained. Results were compared for the each of the two test tasks with the real-life textual indices. As the comparisons involved non-normally distributed data, non-parametric Mann-Whitney tests were performed, where appropriate, for inferential statistics analyses between the conditions (for results of the input text features by automated textual analyses, see Section 4.3.2).

3.4 Research design: cognitive validity

Cognitive validity (Glaser, 1991) addresses the extent to which a test elicits from test takers cognitive processes that correspond to the processes which they would normally employ in a real-life context. The cognitive processes elicited by the real-life tasks and the reading-into-writing test tasks in this study were investigated through a self-report Writing Process Questionnaire to identify the target processes of academic writing and to investigate the processes elicited by the reading-into-writing test tasks.

3.4.1 Participants

As argued in Chapter Two, a homogenous sampling would be suitable for the context of this study. According to UK Council for International Students Affair (UKCISA, 2012), Chinese students formed the largest international population studying in the UK. 219 Chinese students studying on a full-time collaborative undergraduate programme at the Business School, the University of Bedfordshire, were recruited. Their English proficiency was estimated to be between CEFR B2 and C1 (For more details, see Chapter Six). They were pursuing one of four majors: Business Administration, Advertising and Marketing Communications, Human Resource Management, Marketing and Accounting (For the profile of the participants, see Tables 3.7, 3.8 and 3.9).

Table 3.7 Overview of the gender proportion

Gender	Frequency	Per cent
Male	110	50.2
Female	109	49.8
Total	219	100.0

Table 3.8 Participants' majors

Majors at the Business Department	Frequency	Per cent
Business Administration	43	19.6
Advertising and Marketing Communications	37	6.9
Human Resource Management	48	21.9
Marketing	50	22.8
Accounting	41	19.7
Total	219	100.0

Table 3.9 Participants' IELTS scores

	IELTS Reading	IELTS Writing
Mean	5.86	5.52
Std. Deviation	0.60	0.50
Minimum	4.5	4.5
Maximum	7.5	7

3.4.2 Data collection methods and instruments

Many studies on cognitive processing in the writing and reading-into-writing literature use the method of think-aloud, which involves participants thinking aloud (i.e. describing the cognitive processes) as they are completing a task (J. R. Hayes & Flower, 1980, 1983; Plakans, 2008, 2010; Spivey, 1990; Spivey & King, 1989). This method allows researchers to conduct an online investigation of the cognitive processes employed by the participants. Think-aloud protocols can provide comprehensive, in-depth information about the cognitive processes employed by the participants, if they are well-trained (J. R. Hayes & Flower, 1983). Despite its usefulness in showing the participants' cognitive processes online, it is not suitable for the context of this study which involves a large number of L2 participants in both real-life academic and testing conditions. As think-aloud is a very time-consuming method, it is usually used in studies with a small number of participants. Think-aloud has been criticised because of the reactivity and disruption imposed on the actual cognitive processes, especially with L2 participants (Smagorinsky, 1994; Stratman & Hamp-Lyons, 1994). In addition, the method, by its very nature, is unsuitable for use in authentic test conditions.

Many other studies used questionnaire as a non-intrusive method to investigate the cognitive processes employed by the participants during writing or reading-into-writing tasks (e.g. Esmaeili, 2002; Weir, O'Sullivan, Jin, & Bax, 2007). The method can report the cognitive processes employed by a large number of participants in different conditions in a systematic and efficient way. Nevertheless, Purpura (1998) pointed out three concerns regarding the use of cognitive process questionnaires. He suggested that researchers should (1) use human information processing theory as a basis for the questionnaire construct, (2) examine the psychometric characteristics of their instruments before

relating them to performance, and (3) use statistical techniques to assess the underlying construct validity of the questionnaire (ibid., 113). With reference to these concerns, this research aims to develop a Writing Process Questionnaire which would serve as a reliable instrument in test validation procedures in the future. The procedures of developing the questionnaire are reported below.

3.4.2.1 Writing Process Questionnaire

A writing process questionnaire was developed in this study to allow the participants to self-report the processes they employed after completing a) the two selected real-life writing tasks, and b) the two reading-into-writing test tasks (Test Task A and Test Task B).

Theoretically, the framework of the questionnaire was developed based upon Field's (2004, 2008, 2011, 2013) model of different phases of receptive and productive skills, and Shaw & Weir's (2007) model of writing processes. The questionnaire was designed to measure the processes involved in five hypothesised phases of academic writing, namely a) *conceptualisation*, b) *meaning and discourse construction*, c) *organisation*, d) *low-level monitoring and revising* and e) *high-level monitoring and revising*, which are considered to be relevant to the context of this study (See Section 2.5.2.1 for a detailed discussion). In addition, a number of relevant cognitive models including Hayes and Flower's (1980, 1983) writing model, Spivey's (1984, 1990, 1997, 2001) discourse synthesis model, and Khalifa & Weir's (2009) reading model were studied to determine the cognitive processes involved when writers write from sources in the academic context. The following 7 cognitive processes were identified from the literature:

- Task representation
- Macro-planning
- High-level reading
- Connecting and generating
- Organising ideas
- Low-level editing

- High-level editing

Working definitions for each cognitive process were also developed as a result of the literature review (for details, see Table 2.4).

Individual questionnaire items were largely developed by adapting items from the work of Weir et al's (2007) writing processing questionnaire, Weir et al's (2007) survey on reading behaviours, Segev-Miller's (2007) taxonomy of discourse synthesis strategies, and Esmaeili's (2002) writing strategies for integrated reading and writing tasks. These items were then reviewed for content, form and classification by the researcher. The results were discussed by the researcher and two members of the CRELLA research team. The questionnaire was then translated to Chinese by the researcher. The translation was checked by an independent qualified English-Chinese translator. The questionnaire was then ready for trialling prior to a bigger pilot study.

The preliminary Writing Process Questionnaire consisted of 60 items, which were organised in five sections: *reading task prompt*, *reading input texts*, *before writing*, *writing the first draft* and *after writing the first draft*.

The participants can score the extent to which they agree or disagree with each item's description (4= definitely agree; 3=mostly agree; 2=mostly disagree; 1=definitely disagree).

Trial

The questionnaire was trialled with two Chinese MA students. Both participants finished the questionnaire in about fifteen minutes. They were asked to identify items which were unclear to them. The original English version of the questionnaire was also reviewed by three other members of the CRELLA research team. Based on their feedback, six items were removed because they were too similar to other items or were difficult to understand. As a result, the questionnaire was reduced to 54 items (See Appendix 3.4 for the pilot questionnaire).

Pilot

The questionnaire was then piloted with 97 undergraduates. The participants were encouraged to write down any processes they employed which were not mentioned in the questionnaire. The data was submitted to a series of item and reliability analyses. The subsequent changes made to the questionnaire after the pilot are presented below.

(1) Items dropped due to inadequate sampling

The sampling adequacy of each individual item, i.e. whether an item was completed by an adequate proportion of the population, was assessed by examining the anti-image correlation matrix[5]. A value less than 0.05 indicates an inadequate sampling. Therefore, one item (Item 11) with a value of below the threshold was eliminated from the questionnaire.

(2) The revision of items with unsatisfactory correlation with other items.

An initial analysis of the correlations of individual items was then performed to investigate if any items either have no correlation to any other items or correlate too closely to other individual items (i.e. >0.70). The results showed that all items had a correlation coefficient higher than 0.30 with at least some items in the questionnaire. However, some items measuring the low-level editing processes seemed to correlate too highly to each other (i.e. >0.70). In other words, some items were perhaps redundant. As these processes contained more items than the others in the questionnaire, it was felt appropriate to revisit the items with unsatisfactorily high correlations with each other. As a result, four items were combined into two items:

New combined items:
I checked the accuracy and range of the sentence structures.
I checked the appropriateness and range of vocabulary.

(3) Qualitative feedback

[5] Anti-image correlations matrix indicates the part of a variable which is not predictable by regressing each variable on all the other variables. The matrix is a matrix of the negatives of the partial correlations among variables. Partial correlations represent the degree to which the factors explain each other in the results. The diagonal of the anti-image correlation matrix is the KMO measure of sampling adequacy for the individual variables.

Some open space was provided at the end of each section in the questionnaire. The participants were asked to provide some qualitative feedback, e.g. to identify items that were unclear, to name any additional processes that they employed. Four items were dropped from the questionnaire because the participants thought that they were too general and were already covered by other items in the questionnaire. On the other hand, three items were added to include the processes commonly reported by the participants in the open space provided.

I prioritised these ideas in my mind.
My initial writing plan was changed while or after reading the source texts.
My writing plan (e.g. structure, content) was changed while I was writing.

(4) The examination of internal consistency of each cognitive process

The questionnaire was designed to measure seven cognitive processes, i.e. *task representation, macro-planning, high-level reading, connecting and generating, organising ideas, low-level editing* and *high-level editing* involved in five hypothesised academic writing phases, *i.e. conceptualisation, meaning and discourse construction, organising, low-level monitoring and revising.* A series of reliability analyses were performed to assess the internal consistency of the questionnaire items designed to measure the same cognitive process. Overall estimated reliability of each of the seven cognitive processes and each of the academic writing phases (using *Cronbach's alpha*) were obtained. *Item-total correlations* for each item within each cognitive process were obtained. The *adjusted alpha if the item were to be deleted* was also used to inform possible changes to the questionnaire. Cognitive processes which had an alpha lower than 0.50 and items whose item-total correlations were 0.30 were revisited. The overall estimated reliability of the individual items within each cognitive process was also obtained.

The results are presented in Table 3.10. The reliability analysis showed that all five academic writing phases achieved an alpha of 0.50 or above, ranging from 0.52 to 0.84. Out of the seven cognitive processes, items assigning to *task representation* (r=0.39) and *macro-planning* (r=0.35) did not report satisfactory internal reliability of 0.50 or above. Out of the 54 individual

items, only two individual items (Item 1 and Item 8) did not report satisfactory item-total correlation of 0.30 or above.

Item 1 reads as: *I read the task prompt (i.e. instructions) carefully to understand each word in it.* This item was meant to measure a process of *task representation*, i.e. building an initial understanding of the writing task. However, it did not yield an item-total correlation higher than 0.30, which means that the participants did not employ this process similarly as how they employed other *task representation* processes. As the wording of the item contained 'read' and 'carefully', the item might measure a careful reading process. The item was regrouped to *reading*. It reported an item-total correlation to the *reading* process at 0.48. After removing Item 1 from *task representation*, the process's internal consistency improved to 0.54.

Item 8 reads as: *I used my knowledge of how texts like these are organised to find parts to focus on.* This item was meant to measure a process of *generating and connecting* - relating a writer's priori genre knowledge to the source texts. However, it did not yield an item-total correlation higher than 0.30 as other *generating and connecting* items did. As the wording of the item contained 'organised', the item was regrouped to *organising ideas.* It reported an item-total correlation to the *organising ideas* process at 0.56.

The results from the above-mentioned four analyses (i.e. sampling adequacy, correlations, internal consistency and qualitative feedback) conducted with the pilot data and the subsequent changes made to the questionnaire are summarised in Table 3.10. As a result, the revised questionnaire for the main study consisted of 48 items (See the revised questionnaire in Appendix 3.5). The structure of the main study questionnaire is presented in Table 3.11 below.

Table 3.10 Reliability analysis on pilot questionnaire (54 items)

Item No.	Questionnaire items	Mean	Std. dev.	Item-total correlation	Reliability	Changes made	New item no.
Conceptualisation phase							
Task representation							
1	I read the task prompt (i.e. instructions) carefully to understand each word in it.	2.851	0.857	<0.300	0.386	Item1 dropped from Task Representation. The reliability increased to **0.538.** The item was added to *high-level reading*.	1.1
4	I understood the instructions for this writing task very well.	3.361	0.575	0.379			1.4
12	I read the task prompt again while reading the source texts.	2.753	0.915	0.314			2.6
25	I re-read the task prompt while writing.	2.557	0.903	0.398			4.4
Macro-planning							
2	I thought of what I might need to write to make my essay relevant and adequate to the task.	3.219	0.658	0.427	0.349	2 new items added based on qualitative data	1.2
3	I thought of how my essay would suit the expectations of the intended reader.	2.552	0.807	0.324			1.3
5	I thought about the purpose of the task.	2.887	0.794	0.363			1.5
				Final overall reliability:	**0.522**		

Meaning and discourse construction phase							
High-level reading							
6	I read through the whole of each source text carefully.	2.744	0.815	0.407	0.592	Item 1was added.	2.1
7	I read the whole of each source text more than once.	2.536	0.927	0.301			2.2
9	I searched quickly for part(s) of the texts which might answer the question.	3.371	0.660	0.538			2.4
10	I read some relevant part(s) of the texts carefully.	3.323	0.692	0.494			2.5
13	I took notes on or underlined the important ideas in the source texts.	3.392	0.815	0.363			2.7
26	I selectively reread the source texts.	2.660	0.868	0.396			4.5
Connecting and generating							
8	I used my knowledge of how texts like these are organized to find parts to focus on.	3.340	0.768	<0.300	0.629	Item 8 was dropped from *connecting and generating.* It was added to *organising.*	2.3
11	I used my knowledge of the topic to help me to understand the texts.	Dropped due to inadequate sampling					-
14	I linked the important ideas in the source texts to what I know already.	2.866	0.841	0.531			2.9
16	I developed new ideas or a better understanding of existing knowledge.	2.833	0.899	0.303			2.12
23	I developed new ideas while I was writing.	2.705	0.819	0.431			4.2

24	I made further connections across the source texts.	2.729	0.749	0.529			4.3
				Final overall reliability:	0.683		
Organising phase							
Organising							
15	I worked out how the main ideas across the source texts relate to each other.	3.216	0.732	0.408	0.575	Item 8 was added. Another new item was added based on qualitative data	2.11
19	I prioritised the important ideas in the source texts in my mind.	2.771	0.762	0.412			2.8
17	I organized the ideas I plan to include in my essay.	3.280	0.774	0.364			3.1
18	I recombined or reordered the ideas to fit the structure of my essay.	2.773	0.863	0.406			3.2
20	I removed some ideas I planned to write.	2.557	0.914	0.405			3.3
21	I tried to use the same organizational structure as in one of the source texts.	3.072	0.799	0.392			3.4
22	I sometimes paused to organize my ideas.	3.255	0.728	0.457			4.1
				Final overall reliability:	**0.630**		
Low-level monitoring and revising phase							
Low-level editing							
33	I checked that the quotations were properly made.	3.155	0.799	0.434	0.839	Nil	4.12
34	I checked that I had put the ideas of the source texts into my own words.	3.227	0.749	0.422			4.13

36	I monitored and edited the linguistic aspect of my text.	Dropped due to qualitative feedback					-
37	I checked the accuracy of the sentence structures.	Item 37 and 38 were combined due to high correlation between two items					4.15
38	I checked if the range of sentence structures was adequate.						
39	I checked the appropriateness of vocabulary.	Item 39 and 40 were combined due to high correlation between two items					4.16
40	I checked the range of vocabulary.						
47	After revising the first draft, I checked that the quotations were properly made.	2.990	0.909	0.359			5.12
48	After revising the first draft, I checked that I had put the ideas of the source texts into my own words.	2.969	0.897	0.477			5.13
50	After revising the first draft, I monitored and edited the linguistic aspect of my text.	Dropped due to qualitative feedback					--
51	After revising the first draft, I checked the accuracy of the sentence structures.	Item51 and 52 were combined due to high correlation between two items.					5.13
52	After revising the first draft, I checked if the range of sentence structures was adequate.						
53	After revising the first draft, I checked the	Item 53and 54 were combined due to high correlation between two items.					5.14

	appropriateness of vocabulary.						
54	After revising the first draft, I checked the range of vocabulary.						
					Final overall reliability:	**0.841**	
High-level monitoring and revising phase							
High-level editing							
27	I monitored and edited the content development of my text.	Dropped due to qualitative feedback			0.738	Nil	-
28	I checked that the content was relevant.	3.082	0.8532	0.631			4.7
29	I checked that I included all appropriate main ideas from all the source texts.	3.242	0.7244	0.608			4.10
30	I checked that I included my own viewpoint on the topic.	3.216	0.8110	0.531			4.11
31	I checked that the essay was well-organized	2.979	0.8804	0.632			4.8
32	I checked that the essay was coherent.	3.103	0.8267	0.659			4.9
35	I checked the possible effect of my writing on the intended reader.	2.495	0.9596	0.403			4.14
47	After revising the first draft, I monitored and edited the content development of my text.	Dropped due to qualitative feedback					-
42	After revising the first draft, I checked that the content was relevant.	3.011	0.974	0.663			5.7

43	After revising the first draft, I checked that I included all appropriate main ideas from all the source texts.	3.168	0.816	0.706			5.10
44	After revising the first draft, I checked that I included my own viewpoint on the topic.	3.198	0.841	0.709			5.11
45	After revising the first draft, I checked that the essay was well-organized	2.947	0.884	0.710			5.8
46	After revising the first draft, I checked that the essay was coherent.	3.115	0.856	0.561			5.9
49	After revising the first draft, I checked the possible effect of my writing on the intended reader.	2.432	0.955	0.447			5.14
				Final overall reliability:	**0.738**		

Table 3.11 Structure of the main study questionnaire (48 items)

Phases of academic writing	Cognitive processes	Sections of the questionnaire					No. of items
		Reading task prompt	Reading source texts	Before writing	While writing the 1st draft	After writing the 1st draft	
Conceptualisation	Task representation	1.4	2.6		4.4		8
	Macro-planning	1.2 1.3 1.5	2.13		4.6		
Meaning and discourse construction	High-level reading	1.1	2.1 2.2 2.4 2.5 2.7		4.5		11
	Connecting and generating		2.9 2.12		4.2 4.3		
Organising	Organising		2.3 2.8 2.10 2.11	3.1 3.2 3.3 3.4	4.1		9
Low-level monitoring and revising	Low-level editing				4.12 4.13 4.14 4.16	5.12 5.13 5.15 5.16	8
High-level monitoring and revising	High-level editing				4.7 4.10 4.11 4.8 4.9 4.14	5.7 5.10 5.11 5.8 5.9 5.14	12

3.4.3 Data collection procedures

A total of 443 writing process questionnaires were collected from the 219 participants in two phases. The data on test tasks was collected at the beginning of the term while the data on the real-life writing tasks was collected during the term.

Phase 1: Collecting data on test tasks at the beginning of the term

The data on test tasks was collected at the beginning of the term. The two test tasks were administered to the participants under strict test conditions during their language classes, following the instructions provided by the test providers (i.e. LTTC for Test Task A and CRELLA for Test Task B). Immediately after the participants had completed a test task (see Appendix 3.1.3 and Appendix 3.1.4), the questionnaire was used to prompt the participants to self-report the extent to which they employed different cognitive processes when completing the task. The ideal setting would be to assign all participants to do both test tasks (Test Task A and Test Task B). However, this was not achievable due to practical constraints. As a result, about 40% of the participants (n=81, from 4 classes) did both tasks to serve as anchor students. The order of the test administered to them was counter-balanced. Two classes did Test Task A first and the other two classes did Test Task B first.

The remaining 138 students (from 6 classes) did either one of the test tasks. Three classes (n=79) were assigned to do Test Task Aand the other three classes Test Task B (n=59). Independent samples t-tests were performed on the IELTS reading and writing bands of these 129 students who did either Test Task A or Test Task B. The results (see Table 3.12) showed that there was no significant difference between the two groups' proficiency level in terms of their IELTS reading and writing bands.

Table 3.12 Comparisons of the proficiency of the participants who did Test Task A and Test Task B

	Participants who did only Test Task A (n=79)		Participants who did only Test Task B (n=59)		Independent samples t-test
	Mean	Std Dev	Mean	Std Dev	
IELTS Reading	5.91	0.481	5.73	0.601	t(275)=1.901, p=0.060 (n.s.)
IELTS Writing	5.59	0.534	5.58	0.513	t(270)=1.177, p=0.860 (n.s.)

Phase 2: Collecting real-life data during the term

Real-life data, on the other hand, was collected during the term. As described earlier, two writing tasks (see Appendix 3.1.1 and Appendix 3.1.2), were selected from two different modules (Module A and Module B) for investigation in this study. The 219 participants in the study, like other students in the Business School, were allowed to choose two to four selective modules, depending on the structure of their programme. As a result, 70 of the Module A students and 73 of the Module B students participated in the cognitive investigation of this study.

Although the researcher did not have any control over which modules the participants chose to enrol on, it was felt appropriate to investigate the comparability of the level of these two groups of students. Independent samples t-tests were performed on the two groups' IELTS reading and writing bands. The results showed that there was no significant difference between the two groups' proficiency level in terms of their IELTS reading and writing bands (See Table 3.13).

Table 3.13 Comparisons of the proficiency between the participants who did Real-life task A and Real-life task B

	Students who did real-life Task A (n=70)		Students who did real-life Task B (n=73)		Independent samples t-test
	Mean	Std Dev	Mean	Std Dev	
IELTS Reading	5.92	0.690	5.95	0.649	t(254)=0.212, p=0.832 (n.s.)
IELTS Writing	5.51	0.462	5.53	0.471	t(210)=0.260, p=0.796 (n.s)

During the term, 70 students completed Real-life Task A (Essay) and 73 completed real-life Task B (Report) as part of their course work. The questionnaire was administered online through an online survey tool called Survey Monkey one week ahead of the submission deadline of each of the real-life tasks (in middle of the term for Real-life Task A and at the end of the term for Real-life Task B). Students were encouraged to complete the questionnaire online as soon as they had finished their writing assignment task. For those students, about 15 %, who did not complete the questionnaire online, their responses were collected by the researcher during the subject class in the week following the submission.

In summary, a total of 443 writing process questionnaires were analysed in the study, 143 real-life questionnaires and 300 testing questionnaires (See Table 3.14 for the number of questionnaires collected on each task).

Table 3.14 Questionnaire data collected for RQ2

Conditions	Tasks	N	Total
Real-life	A (Essay)	70	143
	B (Report)	73	
Test	A (Multiple verbal inputs)	160 (81 did both + 79 did only A)	300
	B (Multiple verbal and non-verbal inputs)	140 (81 did both + 59 did only B)	
		Total	443

3.4.4 Methods of data analysis

The purpose of RQ2 was to find out what cognitive constructs should be targeted when we assess academic writing ability. Through the self-report questionnaire, the cognitive processes employed by the participants in the real-life academic context were identified in order to define the target cognitive constructs. Using the same instrument, the cognitive processes employed by the participants under the test conditions were also measured to investigate how well the two types of reading-into-writing test tasks elicited from the participants processes that resembled the real-life cognitive constructs. This sub-section explains how the questionnaire data was analysed.

The 443 questionnaires collected from the participants on the two real-life tasks and the two reading-into-writing test tasks, were computed for statistical

analysis. The real-life data was analysed first, and the data on the two test tasks were analysed second.

Investigating the real-life cognitive constructs.

Descriptive statistics of individual questionnaire items from each of the real-life tasks were obtained. The Mann-Whitney U test, which is a non-parametric independent-sample test, was performed on individual questionnaire items to compare the results of the two real-life tasks. Based on the results that the means of the majority of the questionnaire items showed no significant difference between the two tasks, the data sets collected from the two real-life tasks were analysed together in the subsequent factor analyses (for details, see Section 5.2.2 in Chapter Five).

The real-life data was then submitted to exploratory factor analysis (EFA) to investigate the underlying structure of the cognitive processes involved in each of the five academic writing phases elicited by the real-life tasks. As explained in Chapter Two, this study builds on Field's models of receptive and productive language skills, and considers five of the cognitive phases to be most relevant to the discussion of the cognitive validity of academic writing. The exploratory factor analysis conducted was not to build an overall model of the cognitive phases involved in academic writing, but to examine the number of distinctive cognitive processes and the underlying structure of these cognitive processes involved in the five phases of academic writing.

After defining a set of EFA-generated target cognitive processes involved in each academic writing phase, further comparisons between the two real-life tasks was made to compare the extent to which each cognitive process was elicited by the two real-life tasks.

Finally, another set of Mann-Whitney U tests was conducted to examine if these EFA-generated cognitive parameters could reflect a difference in how high-achieving and low-achieving participants approached each of the real-life tasks.

Investigating the cognitive processes elicited by test tasks

Descriptive and inferential statistics on each cognitive process elicited from the participants as a whole group by Test Task A and Test Task B were obtained. The means and standard deviations of each individual item were obtained. Wilcoxon signed ranks test, which is a non-parametric related-sample test, was performed on the processing data from a) Test Task A (n=160) and the two real-life tasks (n=143), and b) Test Task B (n=140) and the two real-life tasks (n=143) to examine the extent to which the processes elicited by two types of reading-into-writing test tasks are comparable to those elicited by the real-life tasks.

A set of Mann-Whitney U tests was then performed to investigate if high-achieving and low-achieving participants performed the cognitive processes differently on Test Task A and Test Task B.

After that, another set of descriptive and inferential (Wilcoxon signed ranks test) statistics was performed to compare the processes employed by the high-, medium-, and low-achieving participants on a) Test Task A and the two real-life tasks, and b) Test Task B and the two real-life tasks.

Finally, exploratory factor analysis (EFA) was then performed on the Test Task A and Test Task B data separately to investigate the underlying structure of the cognitive processes involved in each of the five cognitive phases elicited by each of the reading-into-writing test types. These analyses were to reveal the extent to which the processes elicited by the two reading-into-writing test tasks resemble the target real-life cognitive constructs.

3.5 Research methods for establishing criterion-related validity

Following the description of the research methods for establishing the context and cognitive validity of reading-into-writing test tasks to assess academic writing ability, this section will describe the research methods used to establish the criterion-related validity of reading-into-writing test tasks.

The purpose of analysing students' performances was to investigate the extent to which the reading-into-writing test tasks would demonstrate a link between test performance and test takers' real-life academic writing performance. Without such evidence, we cannot be confident that reading-into-writing tasks are a good format for assessing academic writing ability.

3.5.1 Participants

The 219 participants who participated in the investigation of the criterion-related validity of the reading-into-writing test tasks were the same participants who participated in the investigation of the cognitive validity of the reading-into-writing test tasks (RQ2) of this study (For the profile of the participants, See Section 3.4.1).

3.5.2 Data collection methods and instrument

3.5.2.1 Real-life scores

As the primary purpose of this analysis was to investigate the relationship between the students' performance on the two test tasks and their real-life performance, it was felt that more points of reference as external criteria of the test performance were needed. Therefore, apart from the real-life essay task and the real-life report task, one in-class question and answer test and one end-of-term case study examination from two other modules were collected. Table 3.15 summarises the basic features of the additional real-life measurements.

For the real-life tasks, all marking followed university departmental marking procedures (see Appendix 3.1.1 and 3.1.2 for the marking schemes of the two real-life tasks, presented in the sample tasks). Lecturers who marked the real-life performances were not informed of the present study and they did not know the students' IELTS scores, or their scores on the two reading-into-

writing test tasks. All score data used in this study were the final standardised marks submitted to the University.

Table 3.15 Additional real-life measurements selected for RQ3

Module	Tasks	Task instructions	Input	Time	Output
C	In-class test	Demonstrate understanding of core concepts and theories	A few questions (about 30 words each)	1 hour	No specific word limits. Students were expected to answer each question with about 100 words.
D	End-of-term exam	Write an essay based on a case study (provided in advance) - Critically analyse the issues presented in the case study - Make recommendations and justify with reasons	Verbal and non-verbal inputs (about 2500 words)	2 hours	No specific word limits

3.5.2.2 Reading-into-writing test scores

As presented earlier, a total of 160 participants completed Test Task A, 140 completed Test Task B, 70 completed the real-life task A, and 73 completed the real-life Task B. The two test tasks were marked by the test providers following their standard procedures (see Appendix 6.1 and 6.2 for the marking schemes of the two test tasks). Test Task A was rated by LTTC while Test Task B by CRELLA. Reliability of the test results was checked by the test providers respectively.

3.5.3 Data collection procedures

Real-life scores were collected from the Business School at the end of the term. Test Task A scores were collected from LTTC whereas Test Task B scores were collected from CRELLA about three-months after the test events (See Table 3.16).

Table 3.16 Scores collected for RQ3

Condition	Tasks	No of scores collected
Real-life academic context	Essay	161
	Report	136
	In-class question and answer test	145
	End-of-term case study examination	143
Reading-into-writing language tests	Test Task A	160
	Test Task B	140

3.5.4 Method of data analysis

The first step was to analyse the participants' performance on individual tasks. Descriptive statistical analyses of the participants' scores were performed on individual tasks regarding the overall pattern and the score distribution across levels (For results, see Section 6.2).

The second step was to analyse the predictive power of the test tasks. Correlational analyses were performed on scores between a) Test Task A and real-life measurements (i.e. essay, report, in-class question and answer test and end of term case study examination) and b) Test Task B and real-life measurements. *Correlations* between the performances in the real-life and test conditions as well as *the percentage of variance explained* were discussed (For results see Sections 6.3.1 and 6.3.2). After that, using scatterplots, the patterns of the correlations between the test scores and overall real-life scores were analysed to discuss the strengths and weaknesses of the correlations, i.e. whether the test scores predicted real-life academic outcome at some levels better than the other. The findings would provide information about how well two types of reading-into-writing test tasks could predict the test takers' real-life academic performance in their degree courses.

3.6 Summary

The chapter has presented the research design of this study with respect to the investigation of the context validity, cognitive validity and criterion-related validity of the two EAP reading-into-writing test tasks.

Expert judgement and automated textual analysis were used to investigate the contextual features of the two real-life tasks and the two reading-into-writing test tasks. A self-report questionnaire was used to investigate the cognitive processes involved in five academic writing cognitive phases elicited by the two real-life tasks and the two reading-into-writing test tasks. Three sets of exploratory factor analyses were used to investigate the number of distinct cognitive processes and the underlying structure of these cognitive processes within each cognitive phase elicited by the two real-life tasks, Test Task A, and Test Task B. Participants' performances on the two reading-into-writing test tasks and four real-life measurements (i.e. essay, report, question-and-answer test, case-study exam) were collected. Correlations between the test scores and real-life scores were analysed.

The next chapter presents the results of the contextual features of the two real-life tasks and the two reading-into-writing test tasks in terms of overall task setting and input text features, and discusses the extent to which the salient contextual features of target academic writing tasks are represented by the test tasks.

4 ESTABLISHING THE CONTEXT VALIDITY OF READING-INTO-WRITING TESTS TO ASSESS ACADEMIC WRITING ABILITY

4.1 Introduction

This chapter concerns the context validity of EAP reading-into-writing tests. The chapter reports the results of the salient contextual features of the two selected real-life academic writing tasks (the *essay* task and the *report* task) and the two types of reading-into-writing test tasks (Test Task A *essay task with multiple verbal inputs* and Test Task B *essay task with multiple verbal and non-verbal inputs*), and discusses the extent to which two reading-into-writing test tasks resemble the contextual features of the writing tasks in the real-life academic context (see Appendix 3.1.1 for the real-life essay task, Appendix 3.1.2 for the real-life report task, Appendix 3.1.3 for Test Task A and Appendix 3.1.4 for Test Task B). Chapter Two identified a range of contextual parameters that are likely to influence the difficulty of a reading-into-writing task in terms of **overall task setting** and **input text features** (See Section 2.5.1). The overall task setting was analysed by expert judgement while the input text features were analysed by both expert judgement and automated textual analysis (For details of the research methods, see Section 3.3).

First of all, Section 4.2 reports and discusses the results of the overall task setting between the real-life writing tasks and the reading-into-writing test tasks. The overall task setting of tasks were analysed by expert judgement in this study. Ten judges analysed, individually and then in pairs, the four tasks in terms of *genre, purpose, topic domain, cognitive demands, language functions, clarity of intended reader* and *knowledge of criteria,* i.e., item 1-7 of the Contextual Parameter Proforma (See Table 3.4).

Section 4.3 then reports and discusses the results of the input text features of the real-life tasks and the reading-into-writing test tasks. The analysis of the input texts in this study involved 1) ten sample texts for each of the two real-life tasks, and 2) two passages from ten testlets of Test Task A and two passages from one testlet of Test Task B, totalling 20 real-life input texts and 22 testing input texts in all (The procedures of sampling the input texts were reported in Section 3.2). Two methods: *expert judgement* and *automated textual analysis* were used to analyse the features of the input texts of the tasks. Section 4.3.1 reports and discusses the results from expert judgement. Expert judgement was used to analyse the textual features which cannot be analysed effectively by automated tools. Two judges analysed all the sampled input texts of the four tasks in terms of *input format*, *verbal input genre*, *non-verbal input*, *discourse mode*, *concretenesss of ideas*, *explicitness of textual organisation* and *cultural specificity* (i.e. item 8-14 of the Contextual Parameter Proforma, see Table 3.4). Section 4.3.2 reports and discusses the results from automated textual analysis. CohMetrix version 2.1 (Graesser et al., 2004) and VocabProfile version 3 (Cobb, 2003) were used to analyse the input texts in terms of *lexical* complexity (7 indices), *syntactic* complexity (5 indices) and degree of *cohesion* (5 indices) (The procedures of selecting the indices were reported in Section 3.3.2.2). The results of the real-life input texts are reported first in terms of lexical complexity, syntactic complexity and degree of cohesion (Section 4.3.2.1 - 4.3.2.3). Section 4.3.2.4 compares the textual indices of the real-life input texts in this study with those reported in Green et al's (2010) study of real-life undergraduate reading texts. Section 4.3.2.5 and 4.3.2.6 then compare the textual indices of the real-life input texts with the textual indices of the two reading-into-writing test tasks. The chapter concludes with a synopsis of the main results concerning the context validity of reading-into-writing tests to assess academic writing ability (Section 4.4).

4.2 Overall task setting between real-life writing tasks and reading-into-writing test tasks

Based on the ten judges' responses in pairs to the Contextual Parameter Proforma Item 1- 7, the overall task setting of the two real-life tasks and the two reading-into-writing test tasks are presented and discussed below (The tasks are provided in Appendix 3.1).

4.2.1 Genre

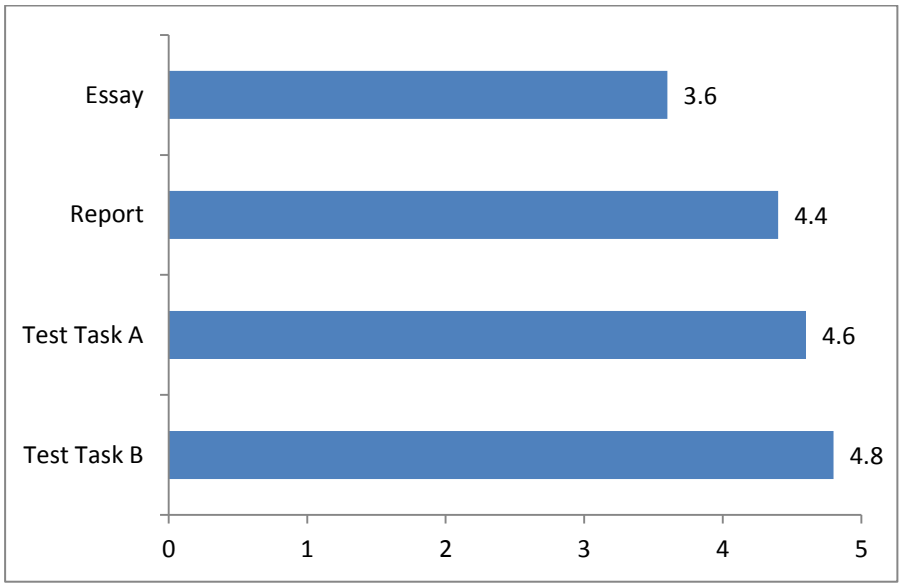
Regarding the genre of the output text of the two real-life tasks, the judges' responses showed total agreement on their responses. For real-life task A, students were expected to produce an essay whereas for real-life task B, students were expected to produce a report (Real-life task A hereafter will be called the essay task, real-life task B the report task).

Regarding the genre of the output text of the two reading-into-writing test tasks, there was some variation among the judges' responses. For Test Task A, four pairs of the judges considered the genre to be essay while one pair decided that it was 'essay and summary'. For Test Task B, three pairs of the judges considered the genre was essay while two pairs regarded it as 'essay and summary'. Pair 1 explained that, 'although the test tasks both require the test takers to write "an essay", both tasks require the test takers to write a summary in more specific terms. Essay can be anything'. While the majority of the judges considered the two reading-into-writing test tasks to be an essay task. One to two pairs of the judges argued that the test tasks incorporated the characteristics of different genres in a single task.

Hyland (2002), in his book on genre, argued that genre represents 'how writers typically use language to respond to recurring situations' (p.4). It is important for writers to be able to identify the genre when they approach a task because their 'choices of grammar, vocabulary, content, and organisation depend on the situations in which they are writing' (p.9). It might be problematic if test takers are required to produce a combined form of different genres that only exist in the testing conditions. Real-life tasks in this study apparently presented the genre of the output text more clearly than the two test tasks did. Another issue is that, as pointed out by one pair of the judges, the genre 'essay' is often used in a too general sense, especially in the test papers.

4.2.2 Purpose of the task

Regarding the clarity of the communicative purpose set in the task, i.e. 'a reason for completing the task that goes beyond a ritual display of knowledge for assessment' (Shaw & Weir, 2007: 71), the judges' responses fell toward the positive end of a five-point Likert scale for all four tasks (see Figure 4.1).



(1=unclear; 5=clear)

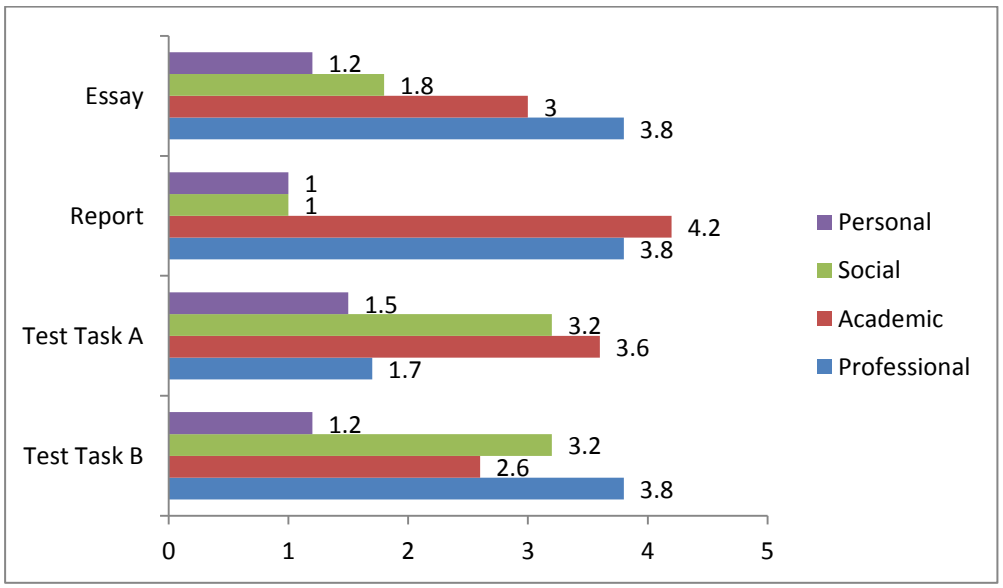
Figure 4.1 Clarity of the purpose of the tasks

However, it is interesting that Test Task A and Test Task B seemed to present a clearer purpose than the real-life tasks did. Pair 5 commented that the communicative purpose presented in the two test tasks were straightforward and unambiguous. The purpose of the real-life essay task was perceived to be the least clear among the four. Pair 2 commented that 'there was hardly a real communicative purpose to achieve on this task, apart from following the instructions.' Nevertheless, although the communicative purpose presented on the two real-life tasks seemed to be less transparent than the test tasks, students may well receive further explanations from the lecturer. As the test takers would not receive any verbal explanation of the test task under test conditions, it is essential for the test tasks to present a clear communicative purpose. The two reading-into-writing test tasks in this study did very well in this regard.

4.2.3 Topic domain

While topic is regarded as one of the major contextual variables which have a significant impact on writing performance (Clapham, 1996; Douglas, 2000; Feak & Dobson, 1996; Read, 1990), it is not always straightforward to analyse the topic domain of a reading-into-writing task. The topic domain of a task can be determined intrinsically by, for instance, the context described in the prompt, the suggested title of the output text, the common theme of the input texts which typically include different perspectives of the 'topic', and the original sources of the input texts. Determining the topic domain of a task by looking at these intrinsic contextual features can be complicated enough. For example, a reading-into-writing task may have an 'academic' context (e.g. writing an academic essay), a 'professional' topic (e.g. Business Law) and input texts originally from a comparatively more 'social' domain (e.g. newspaper and magazine articles). In addition, the topic domain can be determined extrinsically by how the writer would interpret the comparative importance among these contextual features.

In this study, as explained in Section 3.3.2.1, the judges were asked to rate the extent to which each task falls into each of the four topic domains, i.e. professional, academic, social and personal. The results showed that all tasks fell into more than one topic domain (see Figure 4.2).



(1=Not at all; 5=Definitely)

Figure 4.2 Topic domains of the tasks

Based on the judges' response, the topic domain of the essay task was largely *professional*, i.e. Business in the context of the study, but also, to a slightly lesser degree, *academic*. The topic domain of the report task was regarded as primarily *academic*, followed by *professional*. Both the real-life tasks were predominantly in the *professional* and *academic* domains. Agreeing with the literature (Khalifa & Weir, 2009), the *personal* and *social* domains did not play an important role in the academic writing context. Tasks are seen to be in the academic domain when they are concerned with the teaching/learning sectors. The topic in the academic domain can be related to a particular discipline or field of study which may have no practical purpose or use. The professional domain refers to the occupational contexts. The topic is usually related to the specialised knowledge of a profession. According to judges' responses, both real-life tasks fell into the academic and professional domains but the essay task was more '*professional*' while the report task was more '*academic*'.

The two test tasks also possessed multiple topic domains. However, based on the judges' responses, Test Task A was both *academic* and *social* while Test Task B fell into the *professional* and *social* domains. The social domain refers to the contexts connected with general social interaction in a public domain, one usually adopted in language tests of general proficiency. The topic of Test Task A was about whether it is worth saving endangered languages, whereas topic of Test Task B was about the causes of work-related stress and its solutions. The academic domain and professional domain involve specific content. However, the judges felt that although Test Task A fell into the *academic* domain and Test Task B fell into the *professional* domain, both test tasks' input texts contained rather general content, which was usually connected to the social domain.

Test Task A serves as a means of measuring the English language ability of Taiwanese applicants who wish to pursue further studies overseas (LTTC, 2012). Test Task B is a university diagnostic test which aims to differentiate

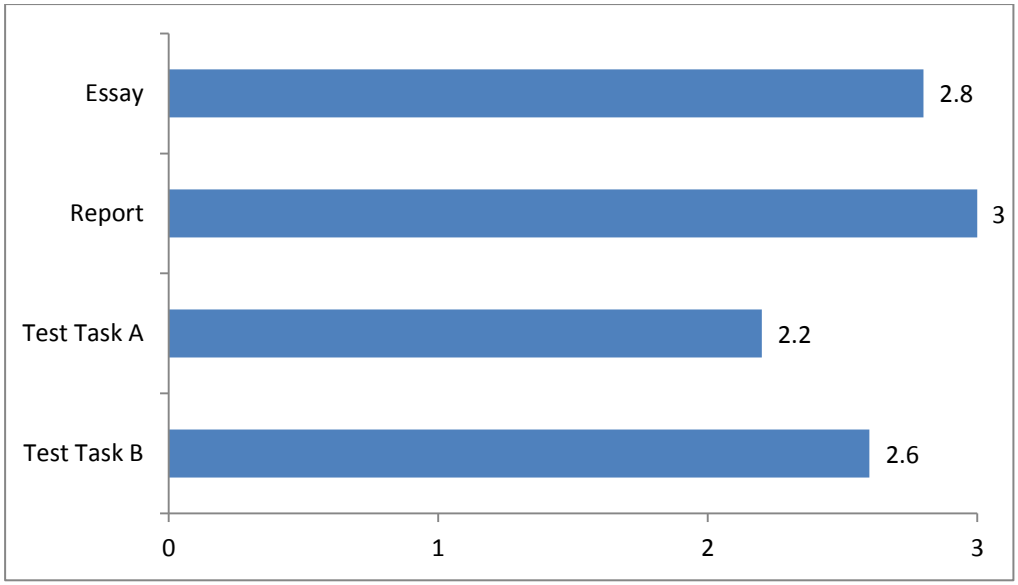
the new international students who would need to receive support in writing and diagnose the weaknesses in their academic writing ability. As both test tasks serve in academic contexts, the social domain does not seem to be entirely appropriate from this perspective. It is, however, understandable why both reading-into-writing test tasks did not contain very specific content. As argued in Chapter Two, one advantage of reading-into-writing tasks is that they can provide an equal access of background knowledge of the topic to prevent bias against test takers. If the content provided in the input texts is too specific, it may impose the background effect on test takers (J. M. Ackerman, 1990; Clapham, 1996; Kellogg, 1987). Unlike ESP tests, writing tasks in EAP language tests should not require a high level of specific knowledge (Douglas, 2000).

4.2.4 Cognitive demands

The cognitive demand of a reading-into-writing task depends largely on the expected 'scope' of the interaction between input and output (Douglas, 2000: 65). Building upon the literature review (Bereiter & Scardamalia, 1987; Galbraith & Torrance, 1999; Purves et al., 1984), the level of cognitive demand of a writing task can be broadly divided into three levels:

1. Telling/retelling content
2. Organising/reorganising content
3. Transforming content

The judges were asked to determine the level of the cognitive demands the tasks impose on the writers by considering the nature of the cognitive processes required and in what way writers should draw upon the input texts. The results are shown in Figure 4.3.



(1=telling/retelling content;
2=organising/reorganising content; 3=transforming content)

Figure 4.3 The cognitive demand of the tasks

Writing tasks at the lowest level of cognitive demand require writers to retell their own priori knowledge on the topic and/or reproduce information provided in the input texts. This primarily involves a linear process of retrieving the writer's internal resources from long-term memory and/or reproducing (i.e. without using the writer's own words or ideas) relevant information from the input texts in response to the communicative purpose of the writing task. Galbraith & Torrance (1999: 3) described such a writing process as 'think-say' or 'what next?' writing. Writing tasks at this level do not explicitly require writers to organize the information they retrieved from long-term memory and/or copied from the input texts. Hence, the structure of most texts produced on such tasks would largely reflect the sequences of how the writer has retrieved the content from the internal resources and/or select the information from the external input texts. Scardamalia & Bereiter (1987) regarded such writing process as *knowledge telling* writing, which is an approach typically employed by immature writers. The standard test format of the impromptu writing-only task type has been criticised as being inauthentic, knowledge-telling tasks, which merely require writers to draw upon internal resources (Cumming, 1997; Feak & Dobson, 1996; Lumley, 2005; Weigle, 2002, 2004; Weir et al., 2013). As shown in Figure 4.3, none of the real-life tasks or the reading-into-writing test tasks were knowledge telling tasks.

Writing tasks at the level of *organising and/or reorganising content* require writers to develop an explicit representation of the rhetorical problem of the writing task and purposefully organise the content they retrieved from long-term memory and/or selected from the input texts in order to solve the rhetorical problem of the writing task. Examples include letters to inform, statements of personal views, technical descriptions, summaries, letters of advice (Purves et al., 1984).

Writing tasks at the highest level of cognitive demand: *transforming content* require writers to establish a high awareness of the rhetorical situation of the writing task. Writers are required to strategically organise as well as transform the content they retrieved from long-term memory and/or selected from the input texts to fulfil writing goals. Such tasks require from the writers a contribution of transformed or new knowledge through the activation of high-level processes, such as *defining* the rhetorical situation of the writing tasks, *integration* of (contradictory) content from multiple internal and external sources as well as *interpreting, elaborating, evaluating*, and *modifying* ideas to satisfy rhetorical goals. Flower (1990), therefore, argued that writing tasks at the highest level would promote the development of 'critical literacy' (See Section 2.2.2 for a discussion of the nature of academic writing as knowledge transforming). Examples include book reviews, commentaries, critical essays, reports (Purves et al, 1984).

As shown in Figure 4.3, the two real-life tasks and the reading-into-writing test tasks were mapped towards the knowledge-transforming end of the cognitive demand's continuum (Scardamalia & Bereiter, 1987). The real-life report task scored an average of 3 out of a scale of 1 (telling/retelling content), 2 (organising/reorganising content) and 3 (transforming content), the real-life essay task an average of 2.8. On the other hand, Test Task A scored an average of 2.2, and Test Task B an average of 2.6. The results showed that the real-life tasks were deemed to be more towards the highest level of *transforming content*, whereas the two reading-into-writing tasks were deemed to be more towards the level of *organising/reorganising content*.

In other words, the real-life academic writing tasks were knowledge transforming tasks. In order to complete a knowledge transforming task, writers are expected to employ high-level processes mentioned above, such as planning rhetorical goals, integrating ideas from different sources and transforming ideas (The actual processes elicited by the tasks are discussed in Chapter Five). The two reading-into-writing test tasks were apparently easier than the real-life tasks in terms of the cognitive demands. Both required the test takers to transform the ideas by selecting, organising and summarising relevant ideas from the input sources as well as evaluating different points of view. However, the test tasks might not require test takers to interpret, evaluate, and apply ideas in context to the extent that the real-life tasks did. Perhaps this is not surprising given the constraints on an exam essay as compared to the wider possibilities of transformational activity in university writing tasks.

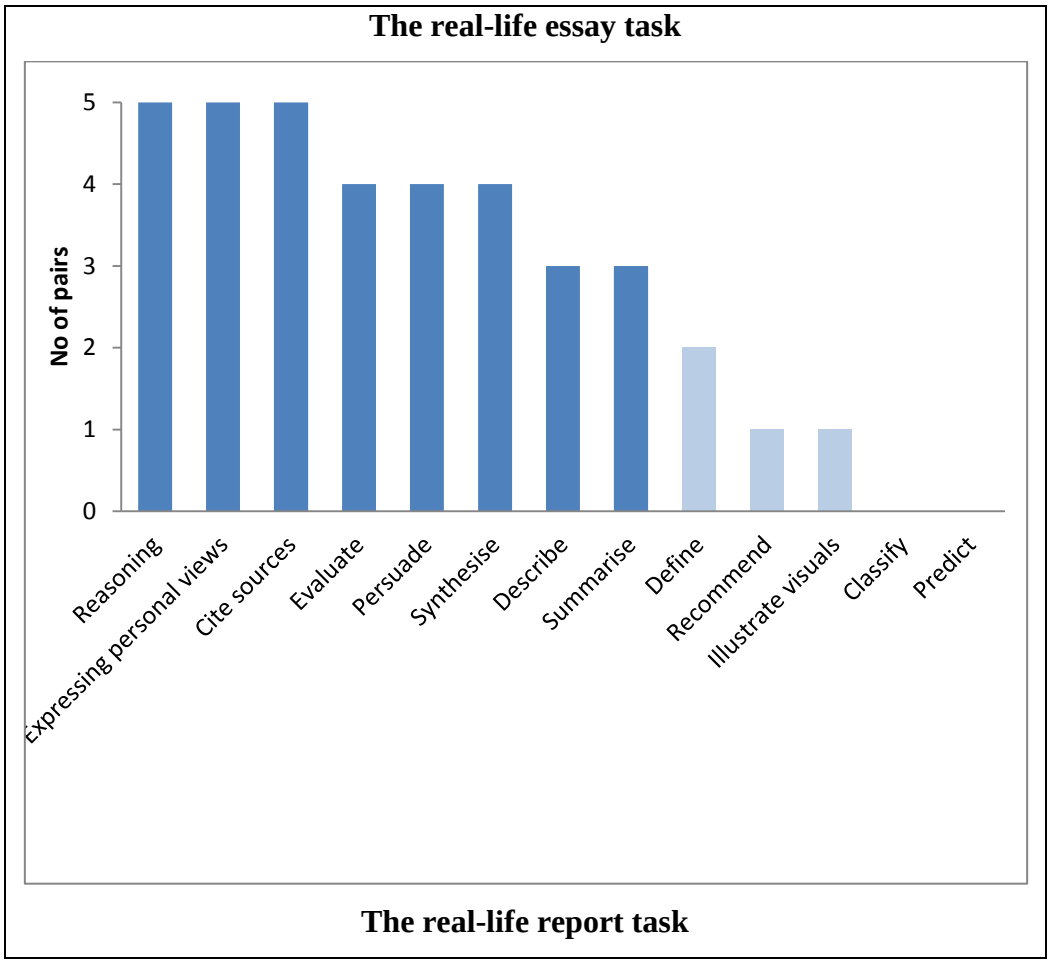
4.2.5 Language functions to be performed

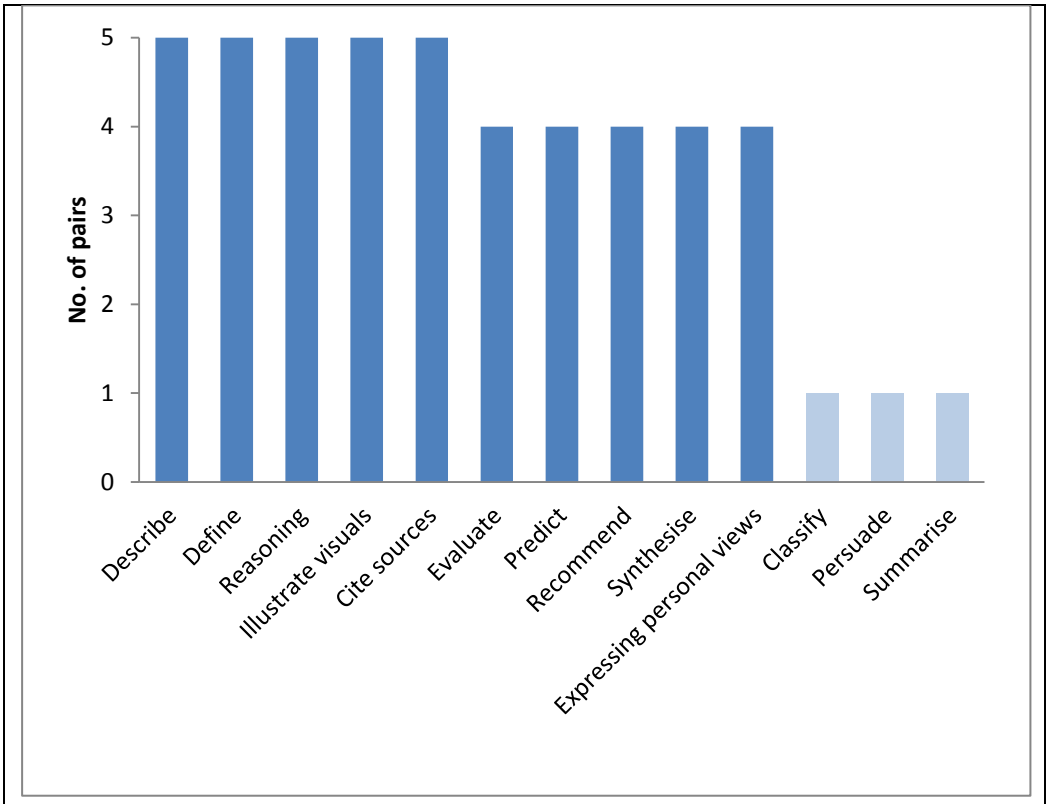
The judges were asked to analyse the language functions that the writers are expected to perform on the four tasks. The judges' evaluations of the language functions varied the most among all categories in the exercise (See Figure 4.4). Pair 4 explained that it was comparatively subjective to determine what language functions are expected from a task because different people might approach the task differently.

The findings showed that, according to the judges, the real-life report task required students to perform more language functions than the essay task and the two test tasks. The judges deemed that the most important functions (i.e. those determined by three or more pairs of the judges) included *describing, defining, reasoning, illustrating visuals* and *citing sources*, followed by *evaluating, predicting, recommending, synthesising* and *expressing personal views.* The most important language functions elicited by the essay task were deemed to be *reasoning, expressing personal view* and *citing sources*, followed by *evaluating, persuading* and *synthesising.* Three pairs of the judges regarded *describing* and *summarising* to be also necessary (See Figure 4.4).

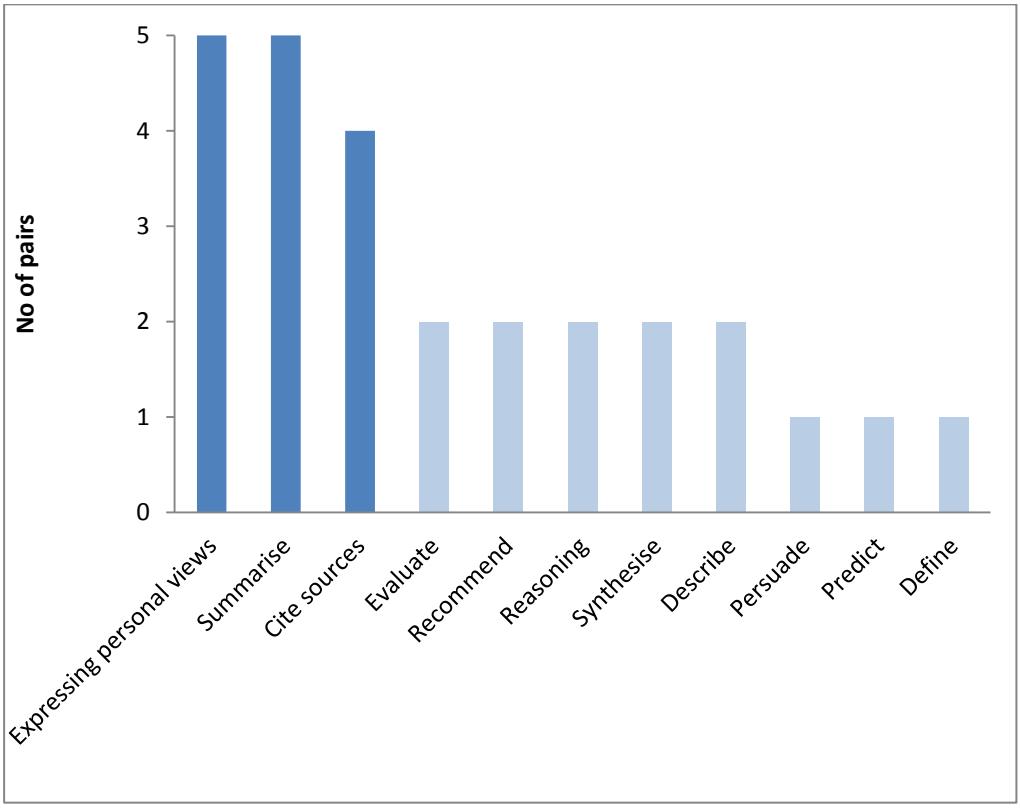
According to the judges, Test task A apparently required fewer language functions. Only two language functions, i.e. *expressing personal view* and *summarising,* were determined by five pairs of the judges as necessary and one function, i.e. *citing sources,* by four pairs. Two or more pairs of judges identified *evaluating*, *recommending*, *reasoning*, *synthesising* and *describing*. One pair of judges identified *persuading*, *predicting* and *defining* (See Figure 4.4).

The judges deemed that Test Task B required test takers to perform, mostly necessarily, *reasoning, summarising* and *express personally viewpoints*, followed by *evaluating, recommending, synthesising and illustrating visuals* (See Figure 4.4).




No. of pairs
5
4
3
2
1
0
Describe
Define
Reasoning
Illustrate visuals
Cite sources
Evaluate
Predict
Recommend
Synthesise
Expressing personal views
Classify
Persuade
Summarise

Test Task A


No of pairs
5
4
3
2
1
0
Expressing personal views
Summarise
Cite sources
Evaluate
Recommend
Reasoning
Synthesise
Describe
Persuade
Predict
Define

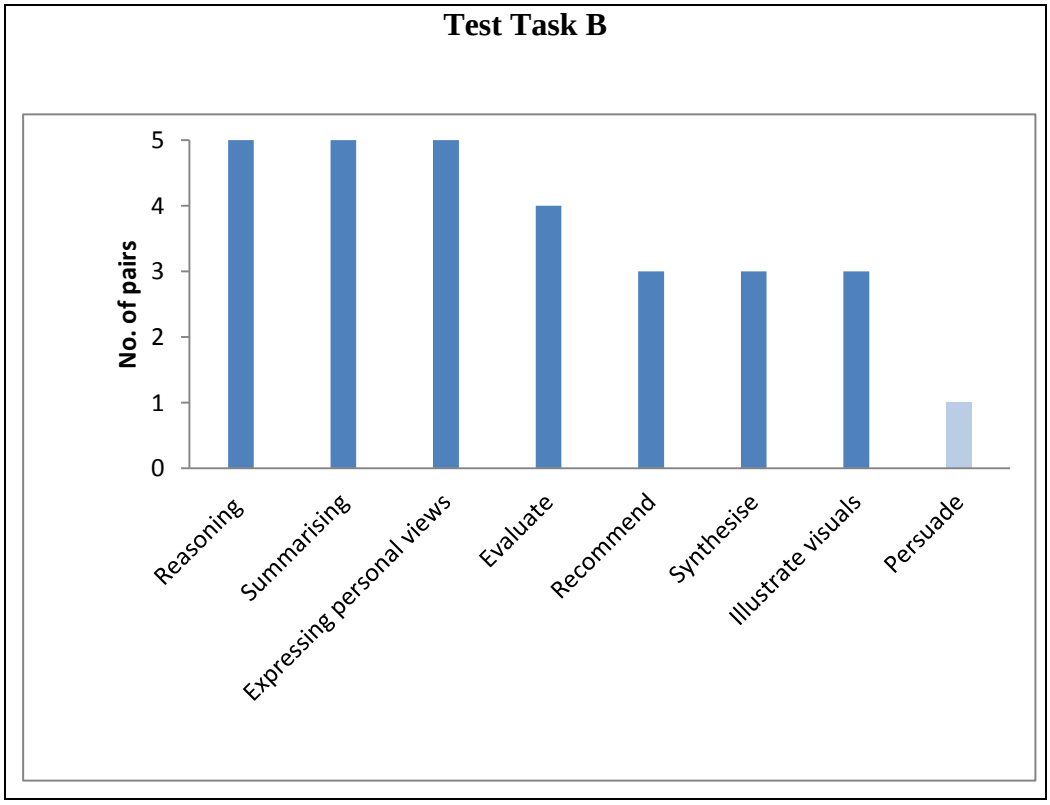


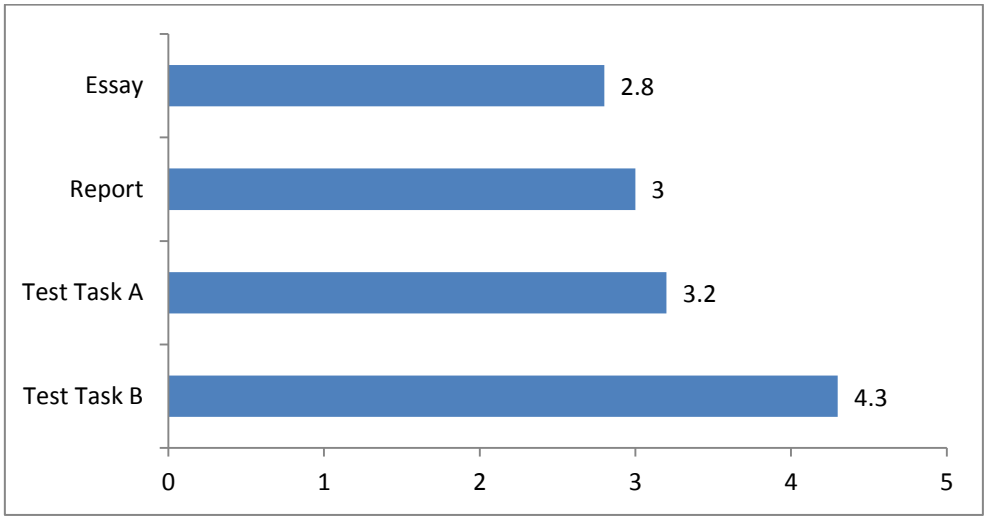
Figure 4.4 Language functions required by the tasks

The findings reveal that some language functions, e.g. *reasoning, expressing personal views, evaluating, synthesising* and *citing sources*, were deemed to be important for both real-life tasks. This indicates the need for these 'core' language functions to be tested in EAP tests. It is worth noting that these core functions were also considered to be those expected in Test Task A by at least two of the pairs of judges. These core language functions, apart from *citing sources*, were considered essential in Test Task B by at least three pairs of the judges. While expert judgements offered useful information about the language functions likely to be elicited by these test tasks, it is also essential to check at the piloting stage of test development whether the expected functions are actually carried out by test takers (Weir & Wu, 2006).

4.2.6 Clarity of intended reader

With respect to the clarity of intended reader presented, the judges considered that both real-life tasks did not do very well (See Figure 4.5). Pair 5 commented that while it might be obvious to the students that the 'real' intended reader of the real-life tasks were the lecturers, both tasks did not provide any information about the intended reader. Mature writers would

consider the needs of the reader while they plan, write and edit their text (Scardamalia & Bereiter, 1987). A valid writing task should always clearly present the intended reader, e.g. self, well known other, distant other (Shaw & Weir, 2007).



(1=Unclear; 5=Clear)

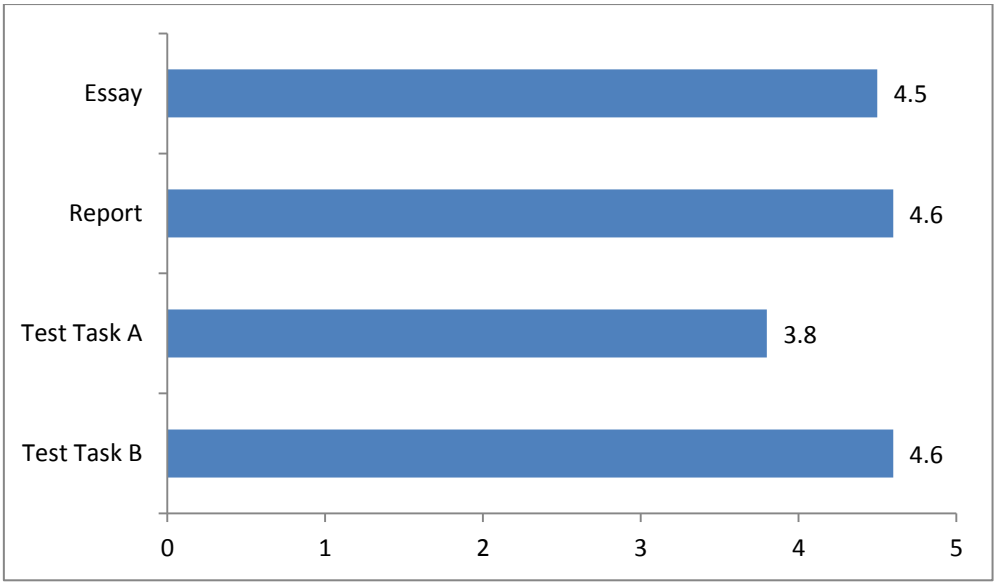
Figure 4.5 Clarity of intended reader of the tasks

The judges gave an average of 3.2 out of 5 for the Test Task A (see Figure 4.6). However, there was some obvious disagreement in the focus group meeting (the rating ranged from 2 to 5). The context of the task was a national essay contest. While some judges considered the implied reader was clear, i.e. the judges of the writing contest, while other judges thought that no actual information was provided regarding the intended reader. It was unclear to them whether the reader would be a single judge, a group of judges, or even a bigger community which could get access to the writing contest.

Test Task B received the highest rating (4.5 out of a score of 5) regarding its clarity of the intended reader among the tasks (see Figure 4.5). Test Task B required the test takers to write to a single lecturer. The judges believed that the relationship between test takers and the intended reader was made clear.

4.2.7 Knowledge of criteria

The overall ratings concerning the provision of the knowledge about marking criteria on the tasks are presented in Figure 4.6.



(1=Unclear; 5=Clear)

Figure 4.6 Provision of the knowledge of criteria

The judges felt that both real-life tasks provided students with very clear and detailed marking criteria. On the other hand, the judges gave an average rating of 3.8 out of 5 for Test Task A. The task stated that test taker's performance would be scored according to four criteria, a) *relevance* and *adequacy*, b) *coherence and organisation*, c) *lexical use* and d) *grammatical use*. However, most judges reported that some more specific descriptions of the criteria would be helpful. Regarding Test Task B, the judges thought that although the descriptions of the marking criteria were much less detailed than those provided on the real-life tasks, the criteria were clear and precise enough in a test situation (4.6 out of 5). Shaw & Weir (2007) argued that test takers' knowledge of criteria would have an impact on whether and how they monitor and revise their texts.

The chapter has so far discussed the overall task setting of the two real-life tasks and the two reading-into-writing test tasks. The next section looks at the results regarding input text features analysed by expert judgement and automated textual analysis tools.

4.3 Features of input texts between real-life writing tasks and reading-into-writing test tasks

Ten sample texts were analysed for each of the real-life tasks, twenty from ten testlets of Test Task A and two from one testlet of Test Task B (see Section 3.2.1 for the procedures of sampling the input texts). Results from the expert judgement will firstly be presented and discussed, followed by the automated analysis results.

4.3.1 Results from expert judgement

Two judges individually analysed all the input texts in terms of *input format*, *verbal input genre*, *non-verbal input*, *discourse mode*, *concreteness of ideas*, *explicitness of textual organisation* and *cultural specificity* (i.e. item 8-14 of the Contextual Parameter Proforma, see Table 3.5).

For the classification criteria, the judges showed total agreement on *the input format*, *verbal input genre*, and *non-verbal verbal input type*. Their agreement rate on the *discourse mode* was 93%. All divergent responses lay between the options of *expository* and *argumentative.* The judges explained that some input texts seemed to serve both discourse modes. They were asked to identify together the primary discourse mode in those texts.

For the three rating scale criteria, i.e. *concreteness of ideas*, *explicitness of textual organisation* and *cultural specificity*, the majority of the two judges' responses (97%) closely coincided with 62% in exact agreement and 35% within one scale point. The remaining 3% of responses displayed a disparity of two scale points. Their responses, if different, were averaged.

4.3.1.1 Input format, verbal input genre and non-verbal input types

The input format of the two reading-into-writing test tasks was standardised. Test Task A contained two reading passages while Test Task B contained two passages with a non-verbal input in each. For the two real-life tasks, verbal input was more dominant on the essay task while the combination of verbal and non-verbal input was more dominant on the report task (See Figure 4.7). 80% of the essay input texts were verbal and 20% contained both verbal and non-verbal information. In contrast, 30% of the report input texts were verbal

and 70% was verbal and non-verbal. None of the input texts contained solely non-verbal input.

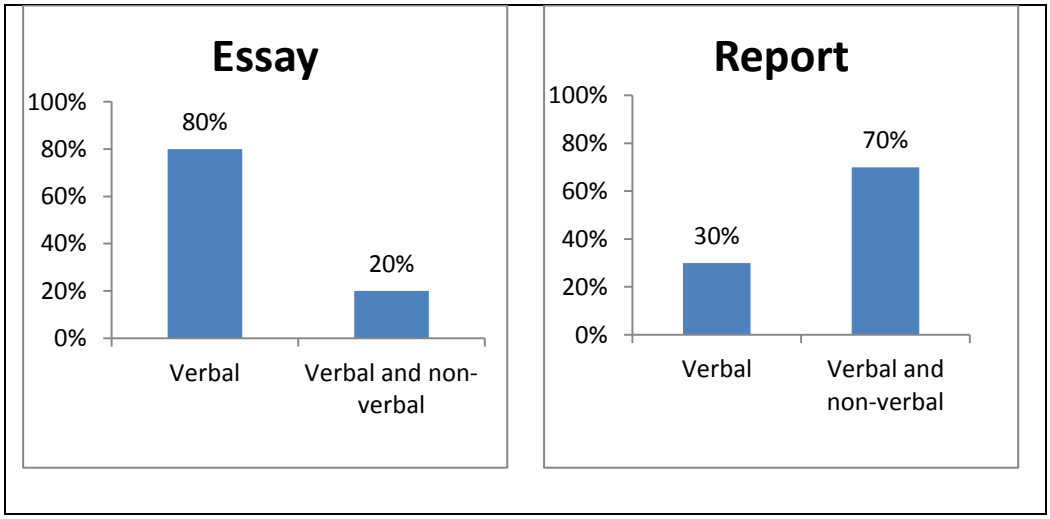


Figure 4.7 Distribution of input format

Regarding the distribution of the verbal input genres, the real-life tasks consisted of a wider range of different genres than the test tasks (See Figure 4.8). News / magazine articles (50%) and journal articles (30%) were the most frequently occurring genres read by the participants on the real-life essay task while book chapters (60%) were the dominant genre for the real-life report task.

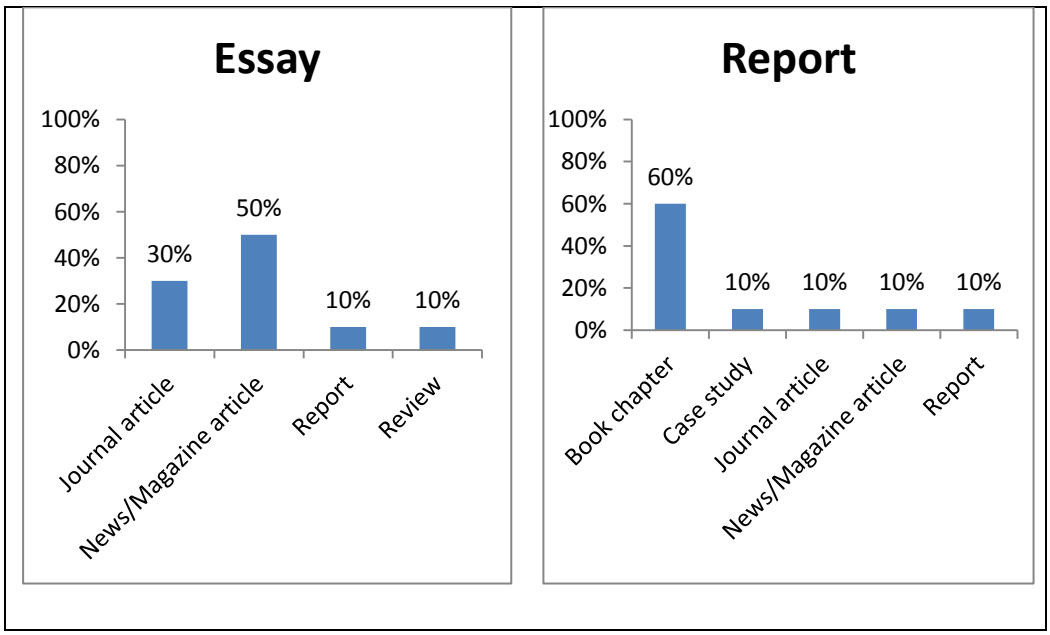


Figure 4.8 Distribution of verbal input genre

As mentioned earlier, each Test Task A contained two input texts. All input texts from the 10 testlets collected in this study were more difficult to associate with these genres and seemed to belong to a rather non-specific text created specifically for the exam, perhaps a simplified version of the essay genre. For Test Task B, only one testlet was available at the time of the study. Test Task B contained two input texts, of which one was identified by the judges as a news/magazine article and the other as a report. Nevertheless, Pair 3 commented that although they were able to identify the genres of the input texts in Test Task B, they did not appear totally authentic. This raises an issue of how to develop or modify texts for test purposes. Recommendation of test design of reading-into-writing tests for item writers will be provided in Chapter Seven.

Regarding the non-verbal input identified in the real-life input texts, diagrams (50%) or pictures (50%) were the most frequently occurring non-verbal information read by the participants for the real-life essay task. In contrast, graphs (71%) and tables (29%) were used more frequently for the report task (See Figure 4.9). Test Task A contained no non-verbal input. Test Task B contained two non-verbal inputs - both were diagrams.

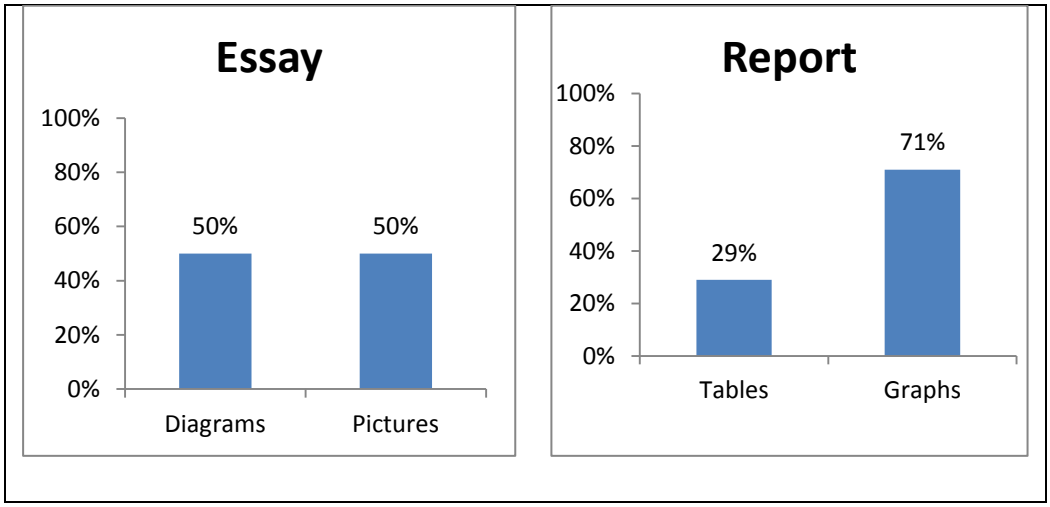


Figure 4.9 Distribution of non-verbal input type

4.3.1.2 Discourse mode

As discussed earlier, both real-life tasks were knowledge-transforming tasks. Students were expected to actively interact with the input texts. The discourse

mode of the input texts would have a direct impact on the task difficulty. Brewer (1980: 225) argued that processing descriptive texts would require the reader to build a visual and spatial cognitive structure; narrative texts would require creating a mental representation of a series of occurring events; and expository texts would require the cognitive processes of constructing induction, classification and comparison.

With respect to the primary discourse mode of the input texts, both real-life texts contained expository and argumentative texts (See Figure 4.10). The report task contained more expository texts while the essay task contained more argumentative texts. No input texts on both real-life tasks were considered as narrative or descriptive.

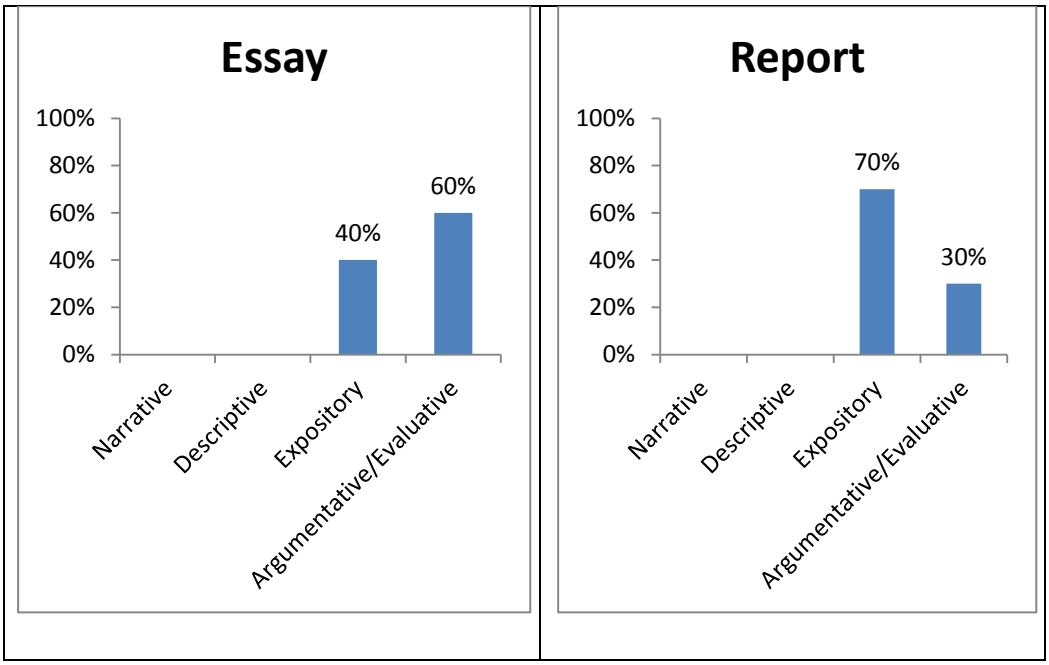


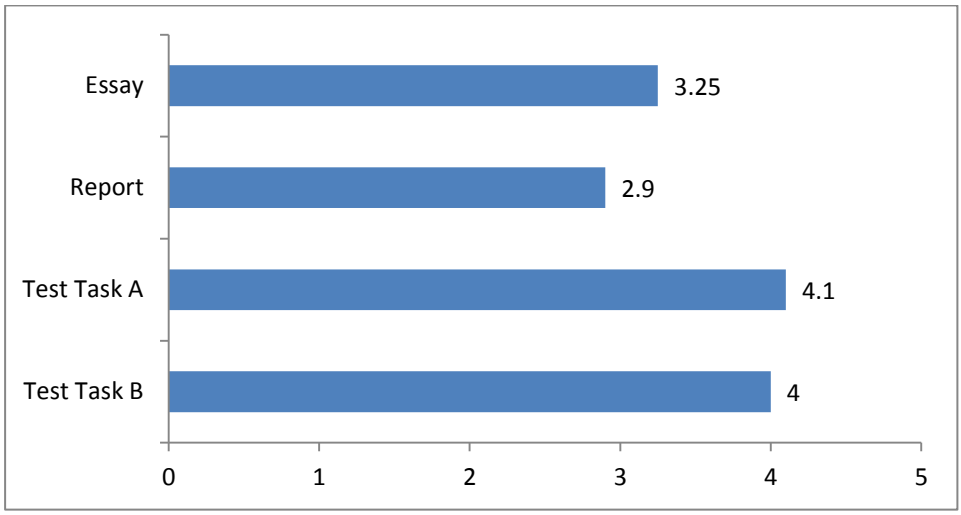
Figure 4.10 Distribution of the discourse mode

In contrast, the input texts on the test tasks were dominated by single discourse mode. All texts in Test Task A were identified as argumentative texts (100%) while all texts in Test Task B contained only expository texts.

4.3.1.3 Concreteness of the ideas

With respect to the concreteness of the ideas in the input texts, the ideas in the test task input texts were considered more concrete than those in the real-life input texts (See Figure 4.11). This is perhaps not surprising. As discussed

earlier, the real-life tasks were in the topic domains of *academic* and *professional*, whereas the test tasks incorporated the *social* domain as well. Therefore, the input texts of the test tasks contained more concrete (i.e. less knowledge specific) ideas.



(1=abstract; 5=concrete)

Figure 4.11 Concreteness of ideas

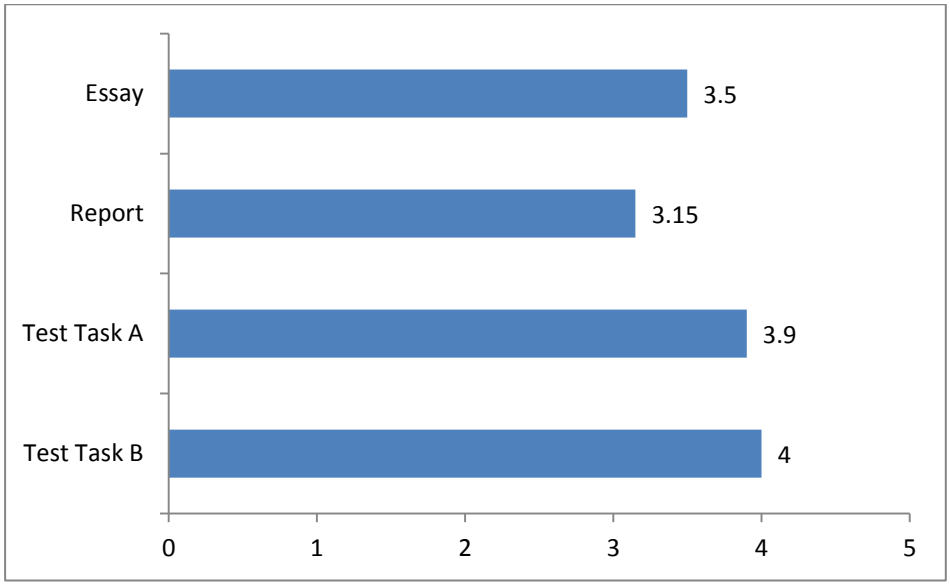
For example, the input texts of one tasklet of the real-life essay task were related to the phenomenon of the feminization in the public relations (PR) industry, another tasklet related to the advertisement strategies of John Lewis to focus on core family values. The input texts of one tasklet of the real-life report task were related to business-specific knowledge such as different techniques to predict the uncertain nature of business trends and graphics useful for modeling and forecasting time series. On the other hand, the input texts of the two test tasks are much less knowledge-specific. For example, one testlet of Test Task A was about the reasons why saving the disappearing languages is important (e.g. every language has unique characterises) as well as the reasons why people should not be concerned about saving disappearing languages (e.g. resources need to be allocated to more important concerns such as education, health and jobs). The input texts of Test Task B were about different methods of handling work-related stress.

The content that test takers are required to process under the test conditions should be more concrete than the content they have to process in a real-life academic context. In other words, the content of the input texts should not be

too specialised or abstract to hinder test takers' ability to apply their writing skills.

4.3.1.4 Explicitness of the textual organisation

Regarding the explicitness of the textual organisation of the input texts, the input texts of the test tasks were more explicitly organised than the real-life input texts (See Figure 4.12). The judges felt that most of the test task input texts were organised mechanically into 3 to 5 paragraphs, each containing a main idea. And these paragraphs were sometimes too explicitly linked with the use of formulaic markers such as 'firstly', 'in addition' and 'lastly'. On the other hand, formulaic markers were found to be less frequent in real-life input texts. There was a higher demand for the students to figure out how each paragraph relates to each other, and to the whole text (i.e. the process of discourse construction).



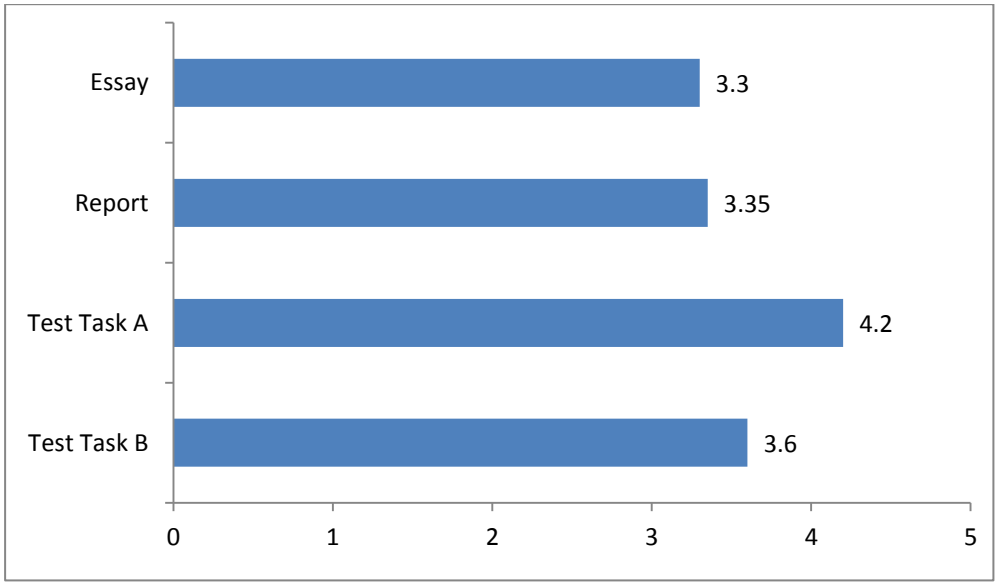
(1=Inexplicit; 5=Explicit)

Figure 4.12 Explicitness of textual organisation

4.3.1.5 Cultural specificity

Regarding the cultural specificity of the input texts, all input texts were rated towards the specific end of the cultural specificity. Test Task A input texts were considered more culturally specific than Test Task B and the two real-life tasks (See Figure 4.13). Many of the Test Task A input texts referred to

subjects which are believed to be familiar to test takers from the Taiwanese testing context.



(1=Neutral; 5=Specific)

Figure 4.13 Degree of cultural specificity

The results of the expert judgement have been reported so far. Next sub-section reports the feedback of the judges regarding their experience of the exercise.

4.3.1.6 Feedback from judges

To evaluate how confident the judges felt when evaluating the contextual features of the two real-life tasks and the two reading-into-writing test tasks using the Contextual Parameter Proforma (see Table 3.4 in Chapter Three), an evaluation questionnaire (see Appendix 3.3) was developed to collect the judges' feedback of their experience. Overall, the judges reported that they were confident when responding to the items of the Proforma in a scale of 4 (4=very confident; 3=confident; 2=not confident; 1=not confident at all) (see Table 4.1). However, they were comparatively least confident when responding to the item of topic domain.

Table 4.1 Feedback from judges

	Mean	Standard Deviation
Part 1 (No. of judges: 10)		
Purpose	3.80	0.42
Topic domain	2.60	0.52
Genre	4.00	0.00
Cognitive demands	3.30	0.48
Language functions	3.50	0.71
Clarity of intended reader	3.60	0.52
Clarity of knowledge of criteria	3.60	0.70
Part 2 (No. of judges: 2)		
Input format	4.00	0.00
Verbal input genre	4.00	0.00
Non-verbal input	4.00	0.00
Discourse mode	3.50	0.71
Concreteness of ideas	3.50	0.71
Explicitness of textual organisation	3.00	0.00
Cultural specificity	3.50	0.71

As argued in Section 4.2.3, it is complicated to determine the topic domain of a reading-into-writing task. The decision might be influenced by a range of task features, such as the context described in the prompt, the suggested title of the output text, the common theme of the input texts, and the original sources of the input texts. The participants commented in the first pilot that it was difficult to choose one topic domain (For details, see Section 3.3.2.1 in Chapter Three). In the second pilot and the main study, the judges were asked to rate the extent to which each task falls into each of the four topic domains, i.e. professional, academic, social and personal. While the judges commented that the change was helpful, the results of the evaluation questionnaire showed that they were still least confident in responding to this item. Further research might need to provide sample tasks in each topic domain to help with the evaluation.

The next sub-section discusses the results regarding the level of difficulty of the input texts analysed by automated textual analysis tools.

4.3.2 Results from automated textual analysis

Regarding the automated textual analyses, 60 extracts from the 20 real-life input texts, 20 passages from 10 testlets of Test Task A, and two passages from one testlet of Test Task B were analysed. Three aspects of textual features of the input texts in terms of the lexical complexity, syntactic complexity and degree of cohesion were analysed by the 17 selected indices generated by two automated tools: CohMetrix and VocabProfile (The procedures of selecting the indices were reported in Section 3.3.3). The textual features of the real-life input texts will be discussed in terms of lexical complexity, syntactic complexity and degree of cohesion (See Section 4.3.2.1 – Section 4.3.2.3). The results are then compared descriptively to the textual indices reported in Green et al's (2010) study of undergraduate reading texts (Section 4.3.2.4). Section 4.3.2.5 and 4.3.2.6 compare the textual indices of the real-life input texts with the textual indices of the two reading-into-writing test tasks.

The two real-life tasks have been analysed and discussed separately so far, but the 60 sample extracts collected from the two tasks will be treated together in this section. The purpose of the automated textual analysis was to determine the difficulty level of the input texts in terms of lexical complexity, syntactic complexity and degree of cohesion. It was felt more beneficial to analyse all the real-life input texts as a whole group, so that the results will provide a more generalisable picture of the appropriate difficulty level of the input texts.

4.3.2.1 Lexical complexity of the real-life input texts

Lexical complexity has long been used to determine the difficulty level of reading texts in the second language learning context (Green, 2010). The lexical complexity of the input texts was analysed by 7 indices, namely, *high frequency words (K1), High frequency words (K1+K2), academic words, low frequency words (offlist), log frequent content words, average syllables per word* and *type-token ratio (content words).* Table 4.2 presents the mean, standard deviation, minimum and maximum of these indices obtained from all real-life input texts.

Table 4.2 Lexical complexity of the real-life texts

	Real-life input texts			
	Mean	Std. Dev.	Minimum	Maximum
High frequency words (K1)	77.20	4.60	68.96	85.94
High frequency words (K1+K2)	87.76	2.98	82.06	93.63
Academic words	10.37	3.79	2.62	19.83
Low frequency words (Offlist)	2.41	1.85	0.00	6.87
Log frequent content words	2.10	0.15	1.83	2.38
Average syllables per word	1.70	0.11	1.46	1.93
Type-token ratio (content words)	0.69	0.08	0.47	0.85

Proportion of high frequency, academic word and low frequency words

The first four lexical indices measure the frequency of all words in the real-life texts. The first two, i.e. the first 1000 and 2000 most frequent words in the BNC corpus, showed the proportion of frequent words in the real-life input texts. 77.2% of the real-life input texts were taken from the first 1000 frequent words and 87.76% the first 2000.

Real-life input texts on average consisted of 10.37% academic words. However, the percentage of academic words in this study ranged from 2.62% to 19.83%. The rather large variation seems to be associated with the results shown in the previous sub-section that the real-life input texts belonged to different genres. Apart from course books, participants also read non-academic texts, e.g. business articles, which probably contain much fewer academic words.

Real-life input texts contained a very low percentage (2.41%) of low frequency words (i.e. those are not included in frequency list of 15000 on the BNC). Some input texts contained no low frequency words. This study focused on a single discipline, i.e. Business. However, the input texts students used to complete the writing tasks did not seem to contain a high proportion of specialised or low frequent vocabulary.

Frequency level, average syllables and type-token ratio of content words

The next three lexical indices concern the content words in the real-life source text. Content words, i.e. nouns, main verbs, adjectives and adverbs, are those contain conceptual meaning.

The index (frequency level) showed the frequency level of the content words in the real-life texts. It computes the log frequency of all content words in the text, ranging from zero to six. The lower the score is, the less frequent the content word is. The mean score of the frequency level of the content words in the real-life input texts was 2.10. Undergraduate course book texts in Green et al's (2010) reported a similar score of 2.14. This means the frequency levels of the content words of the real-life input texts and the course book texts were very close. The comparison of the level of difficulty of the real-life input texts and course book texts is further discussed in Section 4.3.2.4 below.

The index (average syllables) measures the average syllables per content word in the real-life input texts. The content words in real-life input texts on average contained 1.70 syllables, ranging from 1.46 to 1.93. Content words with more syllables are more difficult to process because decoding multisyllabic word takes more time and cognitive effort than decoding a monosyllabic one (Rayner & Pollatsek, 1989). This index therefore partly reflects the decoding demand of the real-life input texts.

The last lexical index (type-token ratio of content words) measures the type-token ratio of all content words in the real-life input texts. The ratio reflects the proportion of unique content words which need to be decoded. The higher the ratio is, the more unique content words there are in the real-life texts. A type-token ratio of 1 means that all words of the text occur only once. Real-life input texts on average had a type-token ratio of 0.69.

4.3.2.2 Syntactic complexity of the real-life input texts

Syntactic complexity is believed to be an important indicator of the difficulty of a text. Researchers, for instance Crossley, Greenfield & McNamara (2008), have demonstrated that the more complex sentence structures a text contains, the more difficult it is for readers to process the text. Syntactic complexity is

particularly important in determining the difficulty of a reading-into-writing task where higher-level reading processing (such as creating textual and intertextual representations) rather than low-level lexical decoding is targeted. Based on the literature review (see Section 3.3.3), the syntactic complexity of the real-life input texts was analysed by five indices in this study. Table 4.3 below summarises the results.

Table 4.3 Syntactic complexity of the real-life texts

	Real-life input texts			
	Mean	Std. Dev.	Minimum	Maximum
Average words per sentence	21.38	3.27	15.32	28.56
Logical operator incidence score	45.12	12.97	16.17	77.28
Mean number of modifiers per noun-phrase	1.03	0.17	0.67	1.38
Mean number of words before the main verb of main clause in sentences	5.50	1.46	2.62	10.65
Sentence syntax similarity	0.08	0.02	0.05	0.11

Average words per sentence

The first syntactic index (average words per sentence) measures the average number of words per sentence in the real-life input texts. Generally speaking, the longer a sentence is, the more complex it is because it might contain more phrases and clauses. A text with many complex sentences is demanding to process because the reader needs to build many elaborate syntactic structures.

Processing long sentences is demanding also because it requires more working memory while the reader is building the syntactic structure (Graesser, Cai, Louwerse, & Daniel, 2006). Such process of building and analysing the syntactic pattern in a string of words is known as parsing (Rayner & Pollatsek, 1989). Real-life input texts on average contained 21.38 words per sentence, ranging from 15.32 to 28.56. Undergraduate course book texts in Green et al's (2010) study also contained a very similar number of words (i.e., 21.47) per sentence. The comparison of the difficult level of the real-life input texts and course book texts is further discussed in Section 4.3.2.4 below.

Syntax similarity index

This index (syntax similarity index) measures how syntactically similar the sentences in the real-life input texts are, by calculating the proportion of nodes in the two syntactic tree structures that are intersecting nodes between all sentences and across paragraphs. It is easier to process a text with more syntactically similar sentences than with more syntactically different sentences due to a syntactic parsing effect. The parsing effect is used to describe the high possibility that a speaker would produce an utterance with a structure similar to the previous utterance he or she produced (Pickering & Branigan, 1998). Ledoux, Traxler & Saab (2007) found that the syntactic parsing effect also presented in comprehension processes. The syntax similarity index between all sentences of the real-life input texts was 0.08, which means 8% of the nodes in the two syntactic tree structures are intersecting nodes between all sentences and across paragraphs. Undergraduate course book texts in Green et al's (2010) reported a mean syntax similarity index of 0.09. This means the range of sentence structures used in the real-life input texts was very close to the course book texts. The comparison of the difficult level of the real-life input texts and course book texts is further discussed in Section 4.3.2.4 below.

Mean number of modifiers per noun phrase and mean number of words before the main verbs of the main clauses

The next two syntactic indices concern the noun phrases and main clauses of the real-life input texts. They measure the mean number of modifiers per noun phrase and the mean number of words before the main verbs of the main clauses respectively. Noun phrases and main verbs in a text are believed to carry the key meaning of a text. Modifiers, e.g. adjectives, adverbs, or determiners, are used to describe the property of the head of a noun phrase or the main verb.

These two indices reflect the difficulty of building the syntactic structures in a text. The more modifiers or words the reader has to read before getting to the head nouns or the main verbs, the more demanding it is to build the syntactic structures.

In addition, the indices also reflect the complexity of the ideas in the text. The more *modifiers before the head nouns* and more *words before the main verbs* mean the more qualities these key ideas possess. One important aspect of academic writing ability is to synthesise, i.e. select, connect and organise, the ideas from the input texts based on the writing purpose. The more qualities the ideas in the source texts possess, the more demanding it is to synthesise these ideas.

The mean number of modifiers per noun-phrase in real-life input texts was 1.03, ranging from 0.67 to 1.38. The mean number of words before the main verbs of the main clauses in the real-life input texts was 5.5, ranging from 2.62 to 10.65.

Logical operator incidence score

The next syntactic index deals with the density of logical operators (connectives) in the real-life texts. Logical operators can be used to explicitly express the relations among the ideas in a text. According to CohMetrix, texts with a high density of these logical operators are difficult. This is perhaps true for low proficiency readers who have problems processing complex sentences. Otherwise, many researchers, e.g. Brown & Yule (1983) and Green et al (2012), argue that the lack of connectives actually increase the difficulty of a text because the reader has to build the relationships between the ideas. Building intertextual and intratextual representations from the input texts is another important aspect of academic writing abilities. Similarly to the previous two indices, this index not only reflects the difficulty of building the syntactic structures, but another academic writing process. The lower the mean logical operator incidence score is, the more difficult it is to build textual representations. Real-life input texts had a mean logical operator incidence score of 45.12, with a wide range from 16.17 to 77.28. However, as CohMetrix did not provide sufficient information concerning how the incidence was computed, the meaning of the index can only be interpreted indirectly as a direction that the lower the score is, the less density of logical operators (connectives) a text contains.

4.3.2.3 Degree of cohesion of the real-life input texts

Measurement of cohesion is used less frequently to determine the difficult level of a reading text than the measurements of lexical and syntactic complexity. However, the degree of cohesion of the input texts is particularly relevant to the discussion of the reading-into-writing tasks in the study. The more cohesive a text is, the easier it would be for the reader to build the textual representation because a cohesive text contains 'explicit features, words, phrases or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas and in connecting ideas to higher level global units, e.g. topics and themes' (Graesser et al., 2004: 193).

It is certainly easier for the reader to create textual representation of a more cohesive text. However, one reason why the measurement of cohesion is less popular in determining the text difficulty is because the cohesion of a text may not be reflected directly by the occurrence of cohesive devices. Kennedy & Thorp (2007) argued that, especially concerning a more advanced level, an overt occurrence of cohesive devices does not necessarily improve the cohesion of a text. Therefore, the results have to be interpreted with caution. The cohesion of the real-life input texts were analysed by five indices in this study. Table 4.4 summarises the results.

Table 4.4 Degree of cohesion of the real-life input texts

	Real-life input texts			
	Mean	Std. Dev.	Minimum	Maximum
Adjacent overlap argument	0.55	0.18	0.17	0.89
Adjacent overlap stem	0.58	0.19	0.09	0.83
Adjacent overlap content word	0.10	0.04	0.04	0.22
Proportion of adjacent anaphor references	0.25	0.18	0.03	0.74
Adjacent semantic similarity (LSA)	0.23	0.09	0.09	0.49

Overlap arguments, stems and content words between adjacent sentences

The first three indices (adjacent overlap argument, adjacent overlap stem and adjacent overlap content word) measure the proportion of adjacent sentences sharing one or more arguments (i.e. nouns, pronouns, noun-phrases), stems

and content words respectively. The occurrence of repeated arguments, stems or content words would make the text more cohesive and hence easier to be processed. These previously-occurring ideas would ease the demand on the reader to process new ideas.

Real-life input texts had a mean adjacent argument overlap score of 0.55, a mean adjacent stem overlap score of 0.58, and a mean adjacent content word overlap score of 0.10. In other words, about the 50% of the adjacent sentences shared one or more argument and word stem, and 10% of the content words in adjacent sentences shared one or more common content words.

Anaphor reference adjacent

The next index (anaphor reference adjacent) measures the proportion of anaphor references between adjacent sentences. It is easier for the reader to resolve the anaphor reference when the referent is in an adjacent sentence, rather than at a distance of a few sentences. Real-life input texts showed a mean adjacent anaphor reference score of 0.25. In other words, 25% of the anaphor references in the real-life input texts referred to their referents in an adjacent sentence.

Adjacent semantic similarity (LSA)

The last selected index (adjacent semantic similarity) measures how conceptually similar each sentence is to the next sentence by comparing the Latent Semantic Analysis (LSA) dimensions of their lexical items. The higher the score was, the more conceptually similar the adjacent sentences are with each other. A high proportion of adjacent sentences with conceptually-related words can help the reader to build the textual representation, e.g. the themes of the text. Green et al (2012) summarised the reasons as below:

a) Enhancing the reader to draw upon stored schematic knowledge relating to the theme (Barlett, 1932);

b) Supporting spreading activation of word/meaning recognition (Hutchison, 2003); and

c) Assisting the reader in building up a coherent information structure for the text (Gernsbacher, 1990).

Real-life input texts have a mean adjacent semantic similarity score of 0.23. The score can vary from 0 (low cohesion) to 1 (high cohesion). Undergraduate course book texts in Green et al's (2010) reported a mean adjacent semantic similarity score of 0.26. This means that the real-life input texts were slightly less conceptually similar across the text than the course book texts, though the difference was very small. The comparison of the difficult level of the real-life input texts and course book texts is further discussed in Section 4.3.2.4 below.

4.3.2.4 Comparison between the real-life input texts and undergraduate texts

The textual features of the real-life input texts have been discussed so far. In order to further discuss the level of difficulty of the input texts students read for their writing assignments, the textual features of the undergraduate course texts reported in Green et al (2010) are provided as a reference. They computed the values from 42 passages extracted from 14 undergraduate textbooks at a British university. Their results were compared descriptively to the results obtained from the real-life input texts in this study to explore if there was a difference in the difficulty level between undergraduate course book texts and the texts undergraduates used to complete their writing assignments.

Regarding the lexical complexity, as shown in Table 4.5, the real-life input texts showed similar figures to the undergraduate course texts analysed in Green et al (2010) in terms of the frequency level of the content words, average syllables per content word and the type-token ratio of all content words. Nevertheless, the real-life input texts had a slightly lower percentage of academic words than the undergraduate texts, though the difference was small. In addition, there were more frequent words (the first 1000 and the first 2000) and less low frequency words in the real-life input texts than the undergraduate course texts. This implies that the real-life input texts were

easier than the undergraduate course book texts in terms of the proportion of word frequency bands.

With regards to the five indices related to the syntactic features of the texts, the difficulty levels of the real-life input texts collected in this study and the undergraduate course book texts collected by Green et al (2010) were largely comparable (See Table 4.5). They contained almost the same average length of sentences (average words per sentence). In addition, the range of sentence structures (sentence syntax similarity) and the average number of modifiers per noun phrase were very close between the real-life input texts and the undergraduate course book texts. However, the real-life input texts contained a slightly higher number of *words before the main verbs of the main clauses* than the undergraduate texts, and a slightly lower *logical operator incidence score* than the undergraduate texts. The results indicate that it might be slightly more demanding to work out the meaning and syntactic structure embedded in the main clauses in the real-life input texts, and slightly more demanding to build the textual representation of the real-life input texts.

Based on the five indices which indicate text cohesion, the real-life input texts and undergraduate course book texts apparently had a very close degree of text cohesion (See Table 4.5). The only difference was that the real-life input texts had a slightly lower adjacent semantic similarity score than the undergraduate course book texts. In other words, sentences in the real-life input texts seemed to be less conceptually similar to the next sentence than those in the undergraduate course book texts.

In short, the difficulty level of the input texts undergraduates used to complete their writing assignments was very close to the undergraduate course book texts.

Table 4.5 Descriptive comparison between real-life input texts and undergraduate course book texts

	Real-life input texts (60 extracts from 20 texts) Mean	Course book texts (Green et al, 2010) (40 extracts) Mean	Descriptive comparison of the difficulty level between real-life input texts and undergraduate course book texts
Lexical features			
High frequency words (K1)	77.20	74.00	The real-life input texts had more first 1000 frequency words in proportion than the course book texts.
High frequency words (K1+K2)	87.76	85.89	The real-life input texts had more first 2000 frequency words in proportion than the course book texts.
Academic words	10.37	10.51	The real-life input texts had slightly fewer academic words in proportion than the course book texts.
Low frequency words (Offlist)	2.41	4.33	The real-life input texts had fewer low frequency words in proportion in proportion than the course book texts.
Log frequent content words	2.10	2.14	The real-life input texts had similar frequency level of the content words as the course book texts.
Average syllables per word	1.70	1.72	The real-life input texts had similar number of syllables per words as the course book texts.
Type-token ratio (content words)	0.69	0.65	The real-life input texts had similar type-token ratio as the course book texts.
Syntactic features			
Average words per sentence	21.38	21.47	The real-life input texts had a similar average sentence length as the course book texts.

Sentence syntax similarity	0.08	0.07	The range of sentence structures used in the real-life input texts was very close to the course book texts.
Mean number of modifiers per noun-phrase	1.03	0.95	The real-life input texts had a similar number of modifiers per noun-phrase than the course book texts.
Mean number of words before the main verb	5.50	4.59	The real-life input texts had more words before the main verb per verb-phrase than the course book texts.
Logical operator incidence	45.12	46.14	The real-life input texts input texts had slightly lower proportion of connectives than the course book texts.
Cohesion			
Adjacent overlap argument	0.55	0.56	The real-life input texts had a similar percentage of the adjacent sentences that shared one or more arguments (i.e. nouns, pronouns, noun-phrases) as the course book texts.
Adjacent overlap stem	0.58	0.58	The real-life input texts had almost the same percentage of the adjacent sentences that shared one or more word stems as the course book texts.
Adjacent overlap content word	0.10	0.10	The real-life input texts had almost the same percentage of the adjacent sentences that shared one or more content words as the course book texts.
Proportion of adjacent anaphor references	0.25	0.24	The real-life input texts had a similar percentage of the adjacent sentences that shared one or more argument as the course book texts.
Adjacent semantic similarity (LSA)	0.23	0.26	The real-life input texts were conceptually less similar across the text than the course book texts.

4.3.2.5 Comparison between the real-life and Test Task A input texts

Having investigated the difficulty level of the real-life input texts in terms of the automated indices, this section examines the extent to which the input texts set in the two reading-into-writing test tasks were similar to the real-life texts. The 17 indices obtained from Test Task A and Test Task B input texts were compared with those obtained from the real-life input texts. The differences between the real-life and Test Task A input texts were analysed by the Mann-Whitney test, which is a non-parametric, between-subjects test (The results will be discussed below). The differences between the real-life and Test Task B input texts were compared descriptively only, due to a small sample size of the Test Task B input texts (The results will be discussed in Section 4.3.2.6.

Overall, the difficulty level of sampled Test Task A input texts was comparable to the level of the real-life input texts (See Table 4.6). The differences in the 14 out of the 17 indices obtained between the two conditions were non-significant. In the remaining three indices, with the exception of *low frequency words (Offlist)*, the differences obtained were slight.

Table 4.6 Comparison of the difficulty level between real-life and Test Text A input texts

	Real-life tasks (60 extracts from 20 texts) Mean	Test Task A (20 texts) Mean	Mann-Whitney U	Wilcoxon W	z	Asymp. Sig. (2-tailed)
Lexical features						
High frequency words (K1)	77.20	76.54	556.000	766.000	-.489	0.63
High frequency words (K1+K2)	87.76	87.69	600.000	810.000	.000	1.00
Academic words	10.37	8.84	437.000	647.000	-1.811	0.07
Low frequency words (Offlist)	2.41	10.41	11.000	1841.000	-6.545	**0.00**
Log frequent content words	2.10	2.05	508.000	718.000	-1.022	0.31
Average syllables per word	1.70	1.72	591.000	2421.000	-.100	0.92
Type-token ratio (content words)	0.69	0.72	448.000	2278.000	-1.689	0.09
Syntactic features						
Average words per sentence	21.38	20.49	514.000	724.000	-.956	0.34
Sentence syntax similarity	0.08	0.09	401.500	2231.500	-2.206	**0.03**
Mean number of modifiers per noun-phrase	1.03	0.91	336.000	546.000	-2.933	**0.00**
Mean number of words before the main verb	5.50	5.76	493.500	2323.500	-1.183	0.24
Logical operator incidence	45.12	43.76	560.500	770.500	-.439	0.66
Cohesion						
Adjacent overlap argument	0.55	0.60	520.500	2350.500	-.884	0.38
Adjacent overlap stem	0.58	0.65	500.500	2330.500	-1.106	0.27
Adjacent overlap content word	0.10	0.09	490.500	700.500	-1.217	0.22
Proportion of adjacent anaphor references	0.25	0.28	457.000	2287.000	-1.589	0.11
Adjacent semantic similarity (LSA)	0.23	0.25	488.000	2318.000	-1.245	0.21

With regards to lexical complexity, Test Task A input texts contained similar proportions of high frequency words (K1 and K1+K2) as the real-life input texts. However, they contained slightly fewer academic words (though the difference was not significant) (See Table 4.6). Interestingly Test Task A contained significantly ($z=-6.545$, $p<0.01$) more *low frequency* words than the real-life input texts and the mean difference was as large as 8% (real-life: 2.41%, Test Task A: 10.41%). The *low frequency* words on the Test Task A input texts were mainly proper names of places and organisations/companies. For the remaining lexical indices concerning the content words, there was not much difference in terms of the frequency of content words and the average syllables per word. Test Task A input texts had a slightly higher type-token ratio but the difference was not significant.

Regarding the syntactic complexity, there was no significant difference in three syntactic indices (average words per sentence, mean number of words before the main verb and logical operator incidence) between the Test Task A and real-life input texts (See Table 4.6). However, Test Task A input texts had a significantly higher sentence syntax similarity index than the real-life input texts, and contained significantly fewer modifiers per noun-phrase than the real-life input texts. This suggests that the Test Task A texts might be less complex to process than the real-life input texts in terms of syntactic complexity, although the actual mean differences were very small.

The degree of text cohesion in Test Task A and the real-life input texts was similar. There was no significant difference in all cohesion indices obtained between Test Task A and the real-life input texts (See Table 4.6).

4.3.2.6 Comparison between the real-life and Test Task B input texts

At the time of the study, only one set of operationalised Test Task B was available. Due to a limited sample size, only descriptive statistics of the textual indices of the two Test Task B input texts are presented here (See Table 4.7). The indices obtained from the real-life input texts are provided in the table for a descriptive comparison. For the 17 textual indices, larger descriptive discrepancies were found in 6 indices (2 lexical, 1 syntactic and 3 coherence indices) between the Test Task B input texts and the real-life input texts.

Table 4.7 Descriptive comparison of the difficulty level between real-life source texts and Test Task B input texts

	Real-life tasks (60 extracts from 20 texts) Mean	Test Task B (2 texts) Mean	Descriptive comparison of the difficulty level between real-life and Test Task B input texts
Lexical features			
High frequency words (K1)	77.20	81.9	The Test Task B input texts had slightly more first 1000 frequency words in proportion than the real-life input texts.
High frequency words (K1+K2)	87.76	91.99	The Test Task B input texts had slightly more first 2000 frequency words in proportion than the real-life input texts.
Academic words	10.37	14.46	The Test Task B input texts had more academic words in proportion than the real-life input texts.
Low frequency words (Offlist)	2.41	6.63	The Test Task B input texts had more low frequency words in proportion in proportion than the real-life input texts.
Log frequent content words	2.10	2.11	The Test Task B input texts had almost the same frequency level of the content words as the real-life input texts.
Average syllables per word	1.70	1.79	The Test Task B input texts had almost the same number of syllables per word as the real-life input texts.
Type-token ratio (content words)	0.69	0.77	The Test Task B input texts had a slightly higher type-token ratio than the real-life input texts.
Syntactic features			
Average words per sentence	21.38	20.32	The Test Task B input texts had a slightly shorter average sentence length than the real-life input texts.
Sentence syntax similarity	0.08	0.10	The sentence structures used in the Test Task B input texts were slightly more similar to each other than those in the real-life input texts.

Mean number of modifiers per noun-phrase	1.03	1.41	The Test Task B input texts had slightly more modifiers per noun-phrase than the real-life input texts.
Mean number of words before the main verb	5.50	4.75	The Test Task B input texts had fewer words before the main verb per verb-phrase than the real-life input texts.
Logical operator incidence	45.12	28.16	The Test Task B input texts input texts had a much lower proportion of connectives than the real-life input texts.
Cohesion			
Adjacent overlap argument	0.55	0.73	The Test Task B input texts had a higher percentage of the adjacent sentences that shared one or more arguments (i.e. nouns, pronouns, noun-phrases) than the real-life input texts.
Adjacent overlap stem	0.58	0.73	The Test Task B input texts had a higher percentage of the adjacent sentences that shared one or more word stems as the real-life input texts.
Adjacent overlap content word	0.10	0.80	The Test Task B input texts had a higher percentage of the adjacent sentences that shared one or more content words as the real-life input texts.
Proportion of adjacent anaphor references	0.25	0.18	The Test Task B input texts had a lower proportion of adjacent anaphor references than the real-life input texts.
Adjacent semantic similarity (LSA)	0.23	0.28	The Test Task B input texts were slightly more conceptually similar across the text than the real-life input texts.

Regarding the lexical complexity, Test Task B input texts seemed to be slightly easier than the real-life input texts due to a higher proportion of the first 1000 and 2000 frequency words. However, larger discrepancies were obtained in other indices (academic words and low frequency words), which apparently suggested that Text Task B input texts were actually more difficult than the real-life input. Text Task B input texts contained more academic words (14.46% vs 10.37%) and low frequency words (6.63% vs 2.41%) than the real-life input texts. In addition, Text Task B input texts had a slightly higher type-token ratio of the content words than the real-life input texts. There was not much difference between the Test Task B input texts and the real-life input texts in terms of the frequency level of the content words and the number of syllables per word. Therefore, while containing a slightly higher proportion of high frequency words, Test Task B input texts could be more difficult to process than the real-life input texts, due to a higher proportion of academic words and low frequency words and a higher proportion of unique content words (type-token ratio).

Regarding the syntactic features, Test Task B had a much lower proportion of connectives (logical operator incidence score) and a slightly more modifiers per noun-phrase than the real-life input texts. This could indicate a higher demand to process the noun-phrases and to sort out the logical connections between ideas in the Test Task B input texts than in the real-life input texts. On the other hand, Test Task B input texts contained a lower average number of words per sentence, a higher sentence syntax similarity score, and a lower number of words before the main verbs in verb phrases, but the actual differences were very small. Therefore, the results seemed to suggest that Test Task B input texts were more syntactically challenging than the real-life input texts due to a noticeable lower proportion of connective in the texts.

Regarding the degree of text cohesion, the Test Task B input texts had a lower proportion of adjacent anaphor references than the real-life input texts. This indicates a more demanding process of resolving the anaphor references in the Test Task B input texts and the real-life input texts. However, the other four text cohesion indices seemed to suggest that Test Task B input texts had a better cohesion than the real-life input texts, and hence were less challenging.

Test Task B input texts had higher proportions of adjacent sentences sharing one or more arguments (i.e. nouns, pronouns, noun-phrases), word stems and content words than the real-life input texts. This means it would be easier to process the main themes in Test Task B input texts than in the real-life input texts. Test Task B input texts also had a higher adjacent semantic similarity score than the real-life input texts, which indicates that the adjacent sentences in the Test Task B input texts were more conceptually similar than those in the real-life input texts.

In short, when compared descriptively to real-life input texts, Test Task B input texts were more demanding in terms of lexical complexity (more academic words and more low frequency words) and syntactic complexity (less proportion of connectives), but less demanding in terms of text cohesion (higher proportions of shared arguments, words stems and content words). Due to the small number of testlets available for Test Task B, it was not possible to do any inferential statistics on the textual indices between Test Task B and real-life input texts. The descriptive results reported here are only suggestive.

4.4 Summary

Chapter Four aims to address RQ1: What are the most appropriate contextual parameters of the EAP writing tasks? To what extent do the reading-into-writing tests resemble these contextual features in the testing conditions?

The chapter has reported the results of the salient contextual features of the two selected real-life writing tasks to shed light on the most appropriate contextual parameters for EAP writing tests. The chapter has also reported the contextual features of two types of reading-into-writing test tasks (*essay with multiple verbal inputs* and *essay with multiple verbal and non-verbal inputs*), and discussed the extent to which the contextual features of the reading-into-writing test tasks resembled the target contextual features of the real-life academic writing tasks. This section provides a summary of the findings regarding the contextual validity of the two reading-into-writing test tasks.

4.4.1 Overall task setting

The results regarding the overall task setting are summarised in Table 4.8. Based on the expert judgement analysis, the two reading-into-writing test tasks resembled the overall task setting of the real-life tasks in a number of important ways. Both reading-into-writing test tasks (Test Task A and Test Task B) required an output of an essay, which was one of the most common genres required in the real-life academic context in this study. In terms of topic domains, the *academic* and *professional* domains were dominant in the selected real-life tasks. Test Task A was considered to be in the *academic* and *social* domains while Test Task B fell into the *professional* and *social* domains. Regarding the cognitive demands imposed on the writer, both real-life tasks were knowledge-transforming tasks which required high-level processes. The two reading-into-writing test tasks were apparently easier than the real-life tasks in terms of the cognitive demands. Both required the test takers to transform the ideas by selecting, organising and summarising relevant ideas from the input sources as well as evaluating different points of view. Nevertheless, the test tasks might not require test takers to interpret, evaluate, and apply ideas in context to the extent that the real-life tasks did. In terms of language functions, real-life tasks seemed to have elicited a wider range of language functions than the test tasks. However, core language functions such as *reasoning, expressing personal views, evaluating, synthesising* and *citing sources* were also considered to be expected in the test tasks. With respect to the clarity of intended reader presented, the judges considered that both real-life tasks did not do very well. Test Task B received higher rating than the real-life tasks while Test Task A received a similar rating to the real-life tasks. Regarding the knowledge of criteria, the judges felt that both real-life tasks provided students with very clear and detailed marking criteria. For Test Task A, most judges considered that the criteria were presented clearly but some additional descriptions of the criteria might be helpful. For Test Task B, the judges thought that although the descriptions of the marking criteria were much less detailed than those provided on the real-life tasks, the criteria were clear and precise enough in the test conditions.

Table 4.8 Summary of results of the overall task setting (Expert judgement)

Overall task setting	Real-life essay task	Real-life report task	Test Task A	Test Task B
1. Clarity of purpose (1=unclear; 5=clear)	3.6	4.4	4.6	4.8
2. Topic Domain (1=not at all; 5=definitely)	Professional (3.8) Academic (3)	Academic (4.2) Professional (3.8)	Academic (3.6) Social (3.2)	Professional (3.8) Social (3.2)
3. Genre	Essay	Report	Essay	Essay
4. Cognitive demands (1=telling/retelling content; 2=organising/reorganising content; 3=transforming content)	2.8	3.0	2.2	2.6
5. Language functions to perform (agreed by 2 or more pairs of judges)	Reasoning Express personal views Cite sources Evaluate Persuade Synthesise Describe Summarise Define	Describe Define Reasoning Illustrate visuals Cite sources Evaluate Predict Recommend Synthesise Express personal views	Summarise Express personal views Cite sources Evaluate Recommend Reasoning Synthesise Describe	Reasoning Summarise Express personal views Evaluate Recommend Synthesise Illustrate visuals
6. Clarity of intended reader	2.8	3.0	3.2	4.3
7. Clarity of marking criteria	4.6	4.5	3.8	4.6

4.4.2 Input text features

The results regarding the input text features analysed by expert judgement are summarised in Table 4.9. The features of the input texts provided on Test Task A and Test Task B were largely comparable to the real-life input texts. As shown in Table 4.8, both real-life tasks required students to write upon multiple external reading resources. Test Task A resembled the context by requiring the test takers to write upon two passages. Test Task B resembled the real-life context by requiring the test takers to write upon two passages containing non-verbal information. Nevertheless, the real-life input texts contained a variety of genres, such as *news / magazine articles, journal*

articles and *book chapters*. All input texts on Test Task A were regarded as belonging to a simplified version of the *essay* genre. Test Task B, on the other hand, contained texts belonging to simplified versions of the *report* and *news/magazine article* genre. Regarding the discourse mode, the real-life input texts were dominantly expository and argumentative / evaluative texts. Test Task A required test takers to process argumentative texts while Test Task B required test takers to process expository texts. In addition, the ideas in the test task input texts were considered more concrete than those in the real-life input texts, and the textual organisation of the test task input texts was more explicitly organised than the real-life input texts. With respect to the cultural specificity of the input texts, all input texts were rated towards the specific end of the cultural specificity scale.

Table 4.9 Summary of the results of the input text features (Expert judgement)

Input text features	Real-life essay task	Real-life report task	Test Task A	Test Task B
8. Input format	verbal (80%) verbal and non-verbal (20%)	verbal (30%) verbal and non-verbal (70%)	2 passages	2 passages with non-verbal information
9. Verbal input genre	book Chapter (60%) report (10%) journal article (10%) news article (10%) case study (10%)	news article (50%) journal article (30%) review (10%) report (10%)	essay (100%)	report (50%) news article (50%)
10. Non-verbal input	pictures (50%) diagrams (50%)	graphs (71%) tables (29%)	Nil	diagrams (100%)
11. Discourse mode	argument/ evaluation (60%) exposition (40%)	exposition (70%) argument/ evaluation (30%)	argument/ evaluation (100%)	exposition (100%)
12. Concreteness of ideas	3.3	2.9	4.1	4
13. Explicitness of textual organisation	3.5	3.15	3.9	4
14. Cultural specificity	3.3	3.4	4.2	3.6

4.4.3 Difficulty level of the input texts

Generally speaking, the difficulty level between the real-life input texts and undergraduate input texts (Green et al, 2010) was similar in terms of most

lexical, syntactic and cohesion automated indices investigated in the study. The only discrepancies were that

(1) the real-life input texts contained more high frequency words (the first 1000 and the first 2000) but less low frequency words than the undergraduate course texts.

(2) the real-life input texts contained slightly higher number of *modifiers per noun phrase* and *words before the main verbs of the main clauses* than the undergraduate texts, and had a slightly lower *logical operator incidence score* than the undergraduate texts.

In other words, the real-life input texts were apparently easier than the undergraduate course book texts in terms of the proportion of word frequency bands. However, it would be slightly more demanding to work out the meaning and syntactic structure embedded in the noun phrases and main clauses in the real-life-input texts, as well as more demanding to build the textual representation of the real-life-input texts than the undergraduate texts.

Regarding the comparison of the difficulty level between the real-life input texts and the test task input texts, the results again showed more similarities than discrepancies. The major discrepancies are summarised below:

(1) Out of 17 indices, only 3 indices obtained significant differences between Test Task A and real-life input texts. Test Task A input texts had a significantly greater density of *low frequency words*, mostly proper nouns, than the real-life input texts. This would probably increase the difficulty of processing the texts if the test takers were not familiar with these proper nouns. However, Test Task A input texts had a significantly higher *sentence syntax similarity index* than the real-life input texts, and contained significantly fewer *modifiers per noun-phrase* than the real-life input texts. This suggests that it would be less demanding to build the textual representation of the Test Task A input texts than that of the real-life-input texts. The degree of text cohesion in Test Task A and the real-life input texts was similar. There was no significant difference in the cohesion indices obtained between Test Task A and the real-life input texts.

(2) Due to the small number of testlets available for Test Task B, only descriptive statistics analysis was performed. Test Task B input texts contained more high frequency words (the first 1000 and 2000) than the real-life input texts. However, Test Task B input texts had a higher proportion of academic words and low frequency words and a higher type-token ratio of all content words than the real-life input texts. This indicates that the lexical complexity of Test Task B input texts was seemingly more demanding than the real-life input texts. Regarding the syntactic complexity, Test Task B had a much lower *logical operator incidence score* and a higher mean number of *modifiers per noun-phrase* than the real-life input texts. This indicates a higher demand on reader to process the noun-phrases and to sort out the logical connections between ideas in Test Task B input texts. Lastly, Test Task B input texts had higher proportions of adjacent sentences sharing one or more arguments, word stems and content words, and a higher adjacent semantic similarity score than the real-life input texts. All these indicated that Test Task B input texts were more cohesive than the real-life input texts.

As summarised above, the results of this study showed that the linguistic complexity of real-life input texts and those used in the two test tasks was largely comparable with only a few discrepancies. Nevertheless, it is worth investigating the issue of whether the apparent major disjunctions, such as *proportion of academic words*, *proportion of low frequency words*, *sentence syntax similarity index*, and *modifiers per noun-phrase*, were a result of test writers modifying or developing input reading texts for test design purposes. For example, in addition to requirements such as style and genre, test writers are often required to include a certain number of 'idea units' into an input text with a certain number of words. It is interesting to investigate how test writers select and edit real-life texts for test design purposes and the effects of such practice in further studies. Recommendations for test takers on developing appropriate input texts for reading-into-writing test tasks for academic purposes are provided in Chapter Seven.

According to the literature review, the difficulty level of a test task is largely determined by its contextual features (Khalifa & Weir, 2009; Shaw & Weir,

2007; Weigle, 2002; Weir & Wu, 2006; Wu, 2012). Many studies were conducted to survey the common writing tasks in the real-life academic context and these studies concluded that most academic writing tasks involved integration of reading materials (e.g. Bridgeman & Carlson, 1983; Carson, 2001; Horowitz, 1986; Johns, 1993; Weir, 1983). However, while revealing the general features of academic writing tasks, very few studies in the literature to date systematically provided detailed information of the contextual features of these academic writing tasks. With the use of expert judgement and automated textual analysis, this study analysed the features of two selected real-life academic writing tasks (the essay task and the report task) and the two reading-into-writing task type in terms of 7 parameters of overall task setting, 7 parameters of the features of input texts and 17 automated textual indices of the linguistic difficulty level of the input texts. In addition, researchers believe that the reading-into-writing task type has good context validity because such integrated task type can arguably reflect the characteristics of real-life academic writing tasks (e.g. Johns, 1993; Read, 1990; Weir et al., 2013). To the knowledge of the researcher, this study was the first study to compare the characteristics of the authentic real-life academic writing tasks and operationalised reading-into-writing test tasks in terms of a range of contextual parameters. The results of this study showed that the two reading-into-writing test tasks largely resembled the contextual parameters of the real-life academic writing tasks.

This chapter has reported and discussed the contextual validity of the two reading-into-writing tests. Chapter Five will shift the attention to the cognitive validity, which concerns to what extent the selected real-life tasks and the two reading-into-writing tasks elicited from the participants in this study the same cognitive processes.

5 INVESTIGATING THE COGNITIVE VALIDITY OF READING-INTO-WRITING TESTS TO ASSESS ACADEMIC WRITING ABILITY

5.1 Introduction

Chapter Four presented and discussed the results of the context validity of the reading-into-writing tests to assess academic writing ability. This chapter presents and discusses the results of their cognitive validity which is concerned with the extent to which a test elicits from test takers cognitive processes that correspond to the processes that they have to use in the target language context (Glaser, 1991; Shaw & Weir, 2007). There are two major steps involved in investigating cognitive validity. First we need to define the target cognitive processes to be measured in a writing test by investigating the processes that students employ in a real-life context. Secondly we need to investigate the extent to which these target cognitive processes can be elicited by the test tasks.

This study investigated the cognitive processes employed by over 200 participants in both real-life academic and authentic test conditions through a carefully developed and validated Writing Process Questionnaire (See Section 3.4.1 for the procedures of developing the questionnaire). As explained in Chapters Two and Three, the processes of *translation* and *micro-planning*, which also play an important role in the writing production, were not investigated in this study based on the following reasons:

- When compared to other processes such as macro-planning and organising, writers tend to be less aware of the use of translation and micro-planning processes because these processes are usually not

taught explicitly in normal classroom settings. Methods such as think-aloud protocols would be more appropriate for the investigation of these processes. Previous studies tended to investigate these processes solely under experimental settings (see Kellogg, 1994 for a review)

- Previous studies have indicated that writing-only and reading-into-writing tasks each elicit a distinct set of processes from writers. For example, the processes of creating textual or intertextual representations are not addressed by writing-only tasks. The processes of translation and micro-planning, on the other hand, might not differ as much as other processes between the independent and integrated test types.

A total of 443 questionnaires were collected from the real-life and test conditions in the study - 70 questionnaires on the real-life essay task, 73 on the real-life report task, 160 on the reading-into-writing Test Task A, and 140 on the reading-into-writing Test Task B (See Section 3.4.2 for the details of data collection).

This chapter begins with the results pertaining to the cognitive processes performed by the participants in real-life conditions. Descriptive statistics of individual questionnaire items from each of the real-life tasks and the comparison of the cognitive processes employed on the two real-life tasks are reported in Section 5.2.1. Results from the exploratory factor analysis (EFA) of the number of distinct cognitive processes involved in five academic writing cognitive phases and the underlying structure of these processes are reported in Section 5.2.2. After defining the EFA-generated underlying structure of the real-life cognitive constructs, further comparisons of the two real-life tasks, to compare the extent to which each cognitive process was elicited, are presented in Section 5.2.3. In the context of language tests, it is important to collect evidence to show if the cognitive parameters can distinguish how more proficient writers and less proficient writers employ these processes. A comparison of the cognitive processes employed by high-achieving and low-achieving participants in the real-life context is presented in Section 5.2.4. Sections 5.2.5 summarises the results of the data relating to real-life academic writing processes.

After investigating the target cognitive constructs, Section 5.3 reports the results elicited by the two reading-into-writing test tasks. Section 5.3.1 compares the cognitive processes elicited by Test Task A and real-life tasks (Section 5.3.1.1), and by Test Task B and real-life tasks (Section 5.3.1.2). Section 5.3.2 discusses the comparison of the cognitive processes employed by high and low achieving groups on Test Task A (Section 5.3.2.1) and Test Task B (Section 5.3.2.2). Section 5.3.3 discusses the comparison of the processes, in groups of high-, medium- and low-achievement, elicited by Test Task A and real-life tasks (Section 5.3.3.1), and by Test Task B and real-life tasks (Section 5.3.3.2). A summary is given in Section 5.3.3.3. Section 5.3.4 discusses the underlying structure of the cognitive processes elicited by the test tasks (Section 5.3.4.1 – Section 5.3.4.5). A summary is given in Section 5.3.4.6. Section 5.4 provides a brief synopsis of the whole chapter.

5.2 Investigating the target cognitive constructs in the real-life context

The cognitive processes the participants used when they were completing writing tasks in the real-life academic conditions were measured by the Cognitive Process Questionnaire (See Appendix 3.5). A total of 143 questionnaires were collected on the two real-life tasks: *essay* and *report*. 70 participants who completed the real-life essay task and 73 who completed the report task filled out the Cognitive Process Questionnaire. The questionnaire consisted of 48 items, measuring five major phases (i.e. *conceptualisation, meaning and discourse construction, organising, low-level monitoring and revising,* and *high-level monitoring and* revising) and the cognitive processes within each phase (i.e. *task representation, macro-planning, high-level reading, connecting and generating, organising, low-level editing* and *high-level editing*) that writers most likely go through when writing from external sources. (A glossary of the phases and the cognitive processes was presented in Table 2.4 in Chapter Two).

5.2.1 Significant differences between the two real-life tasks in terms of individual questionnaire items

First, the means and standard deviations of the rating (*4=definitely agree; 3=mostly agree; 2=mostly disagree; 1=definitely disagree*) of the 48 items

from the two sets of real-life data were obtained. Mann-Whitney U tests were performed to test the differences between the means of the rating to investigate if the participants employed the individual items differently between the two real-life tasks. The results showed that apart from 3 items (i.e. Item 1.4, 4.1 and 4.14), the differences in all items between the two real-life tasks were non significant (The results of the items with significant difference are presented in Table 5.1; the results of all items are provided in Appendix 5.1). This indicates that the participants rated the extent to which they employed most items in a similar way on the two real-life tasks, and the actual differences were relatively slight.

Table 5.1 Significant differences between the two real-life tasks in terms of individual items

		Essay		Report		Mann-Whitney U	Wilcoxon W	z	p
		Mean	Std Dev	Mean	Std Dev				
1.4	I understood the instructions for this writing task very well.	3.03	.636	3.25	.703	2091.500	4576.500	-2.105	.035
4.1	While I was writing I sometimes paused to organise my ideas.	3.16	.673	2.86	.751	2040.500	4741.500	-2.292	.022
4.14	I checked the possible effect of my writing on the intended reader while I was writing.	3.06	.814	2.75	.846	2059.500	4760.500	-2.162	.031

As shown in Table 5.1, the participants reported that they understood the instructions for the report task better on a scale from 1 (*definitely disagree*) to 4 (*definitely agree*) than the essay task. This agrees with the results in the contextual analyses, reported in Section 4.2, where the expert judges regarded that the report task presented clearer information about the communicative purpose and the intended reader than the essay task. In addition, participants rated the extent to which they *paused to organise their ideas* and *checked the possible effect on the intended reader during writing* higher on a scale from 1 (*definitely disagree*) to 4 (*definitely agree*) on the essay task than on the report task.

The essay task seemed to have elicited from the participants a higher awareness of the needs of the intended reader than the report task. The results in the contextual analyses may offer an explanation. According to the judges, the essay task requires students to *persuade* whereas the report task does not. The marking schemes of the two real-life tasks may offer more details. The report task was scored based on four categories: (1) examination of the data and description of the nature of the dataset; (2) discussion and justification of the techniques chosen; (3) reasons for rejecting the inappropriate techniques; and (4) discussion of other relevant issues. The essay task was scored based on: (1) problem definition and structure of the text; (2) information identification (the number of sources, relevance to the topic, reliability of the sources); (3) critical reasoning; and (4) persuasion and influencing. In comparison to the report task, the quality of the essay is more dependent on the persuasiveness of the content. This may be why the participants checked the possible effect on the intended reader significantly more on the essay task than the report task.

For the remaining 45 items, the results of the Mann-Whitney U tests showed no significant difference in the participants' rating between the two real-life tasks (see Appendix 5.1). With regards to individual questionnaire items, the participants seemed to have employed the processes very similarly between the real-life essay and report tasks. Further comparisons of the cognitive processes employed on the two real-life tasks are reported in Section 5.2.3.

5.2.2 Factor analyses of real-life academic writing cognitive processes

As mentioned previously, the 48 questionnaire items were categorised (based on the literature review) to measure the cognitive processes, i.e. *task representation, macro-planning, high-level reading, connecting and generating, organising, low-level editing* and *high-level editing,* that writers go through during the five cognitive phases (i.e. *conceptualisation, meaning and discourse construction, organising, low-level monitoring and revising,* and *high-level monitoring and* revising) of academic writing when they write from external sources. The internal reliabilities of the questionnaire items measuring the same phase and the same cognitive process were checked in the pilot study (See Section 3.4.1). In the main study, exploratory factor analysis (EFA) was

conducted to investigate the number of distinct cognitive processes within each of the five cognitive phases and the underlying structure of these cognitive processes elicited by the two real-life academic writing tasks.

Considering the three reasons given below, it was decided to analyse the data collected from the two real-life tasks together in the subsequent factor analyses.

(1) The means of 97% of the questionnaire items showed no significant difference between the two tasks;

(2) The primary purpose of the study was to define the cognitive constructs measured by predominant real-life academic writing tasks, so that these constructs can be targeted in the test conditions. Provided that the participants reported using individual processes similarly between the two real-life tasks, analysing the data as a whole group would improve the generalisability of the results.

(3) Analysing the data together would increase the size of the data which is beneficial for factor analysis.

Based upon the literature review, writers are likely to go through several cognitive phases when they write from external sources, though the phases can be overlapping or looping back. This study aimed to measure the cognitive processes that the participants employed at these five hypothesised phases: (1) *conceptualisation*, (2) *meaning and discourse construction*, (3) *organising*, (4) *low-level monitoring and revising* and (5) *high-level monitoring and revising* (Field, 2004, 2008, 2011, 2013; Kellogg, 1996, 1999; Shaw & Weir, 2007). Exploratory factor analysis was performed to investigate number of distinct cognitive processes employed at these hypothesised phases and the underlying structures of these processes on the real-life tasks. In other words, the study investigated what cognitive processes were involved in each of these hypothesised academic writing phases. The findings provided statistical evidence of (1) how many distinct cognitive processes loaded on each academic writing phase, and (2) the extent to which each distinct cognitive process loaded on the academic writing phase.

First of all, the *Kaier-Meyer-Olkin (KMO) measure of sampling adequacy* and the *Bartlett's test of sphericity* were performed to test the real-life processing data's appropriateness for factor analysis (See Appendix 5.2). The results showed that the data passed both tests, indicating its suitability for factor analysis. In addition, the data was analysed by *Kolmogorov-Smirnov* test with regards to its normal distribution. The results showed that the real-life processing data was not normally distributed ($p<0.01$). Following Fabrigar et al's (1999) recommendation for non-normally distributed data in factor analysis, the *Principal Axis Factor Method* was performed to extract the initial factors. The *eigenvalues* [6] and *scree plot* [7] were examined for an initial indication of the possible number of factors extracted by the data. Rotated solutions of the factor loadings, which avoid maximising the variance accounted by the first factor, were obtained by an oblique promax rotation. Oblique rotation method is often used in social science and language studies because it allows the factors to be correlated. To determine the ultimate number of underlying factors to be extracted, the possibilities were interpreted and evaluated based on both statistical results and theoretical rational.

The underlying structure of the five hypothesised academic writing phases: (1) conceptualisation, (2) meaning and discourse construction, (3) organisation, (4) low-level monitoring and revising, and (5) high-level monitoring and revising elicited on the real-life tasks are presented and discussed below.

5.2.2.1 The underlying structure of the conceptualisation phase (real-life)

Conceptualisation is the first phase of productive skills (Kellogg, 1996, Field, 2004, 2011) where writer develops an initial task representation, i.e. an initial understanding of the rhetorical situation of the writing task (Flower, 1990) and where writer sets macro-plans, i.e., to establish writing goals in different aspects such as intended readership, genre, content and style (Shaw & Weir, 2007). Lower-level reading processes, i.e. decoding, lexical search, and parsing, were not sampled in the questionnaire of this study because students

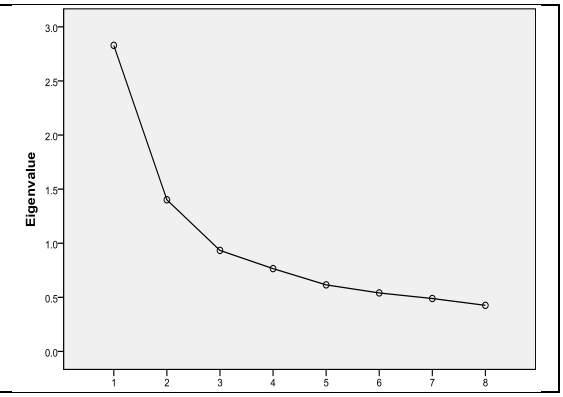
[6] Factors with an eigenvalues below 1 need to be dropped.
[7] The point (also called an elbow) where there is a sudden drop of the steepness of the curve indicates signals that the factors on its left are significant.

at the undergraduate level presumably have mastered high automaticity in these lower-level reading processes.

The conceptualisation phase was measured by 7 questionnaire items in this study. The initial factor extraction for the conceptualisation phase elicited in the real-life conditions yielded two factors with eigenvalues greater than 1.0. The scree plot also suggested a two-factor solution (See Table 5.2).

Table 5.2 Eigenvalues and scree plot for the conceptualisation phase (real-life)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.829	35.363	35.363
2	1.401	17.517	52.880
3	.934	11.670	64.550
4	.765	9.567	74.117
5	.615	7.690	81.806
6	.541	6.757	88.563
7	.489	6.112	94.675
8	.426	5.325	100.000



The two-factor suggestion was accepted. The rotated two-factor solution for the conceptualisation phase and the interacted correlations of the factors are presented in Table 5.3. The percentage in brackets indicates the extent to which each factor accounts for the variance. As shown in Table 5.3, Factor 1 contains the majority of the 8 items, which include three *macro-planning* processes (i.e. Item 1.2, 1.3, 1.5) with regards the relevance and adequacy of content, purpose of the task and effect on intended reader, and three *task representation* processes at different stages of the writing process (i.e. Item 1.4, 2.6, 4.4). The factor was named *task representation and macro-planning.* Factor 2 contains two items only, both relating to the process of changing macro plans at different stages of the writing production, one after reading the source texts (Item 2.13), another while writing the first draft (Item 4.6). The factor was named *revising macro plan.*

Table 5.3 Pattern and interfactor correlations matrix for the conceptualisation phase (real-life)

		F1 Task representation and macro-planning (34%)	F2 Revising macro plan (19.09%)
1.2	I thought of what I might need to write to make my text relevant and adequate to the task.	.806	
1.5	After reading the prompt, I thought about the purpose of the task.	.688	
1.4	I understood the instructions for this writing task very well.	.588	
1.3	I thought of how my text would suit the expectations of the intended reader.	.519	
4.4	I re-read the task prompt while I was writing.	.449	
2.6	I went back to read the task prompt again while I was reading the source texts.	.445	
2.13	I changed my writing plan while reading the source texts.		.804
4.6	I changed my writing plan (e.g. structure, content, etc) while I was writing.		.583
Interfactor correlations			
Factor 1 (Task representation and macro-planning)		1.000	
Factor 2 (Revising macro plan)		.061	1.000

The results suggested that, as indicated by Factor 1, task representation and macro-planning processes were employed by the participants together in a similar way when they conceptualised an understanding of the writing task and established their macro plans to complete the task in the real-life academic context. On the other hand, Factor 2 empirically supports Hayes & Flower's (1983) writing model that planning is a not a one-off process employed in the beginning of the writing process, but a recursive process employed throughout the writing production process. The results showed that participants revised their macro plans at different stages of the writing production in real-life academic writing. Interestingly, as shown in Table 5.3, the two factors correlated weakly (r=0.061). In other words, the process of *revising macro*

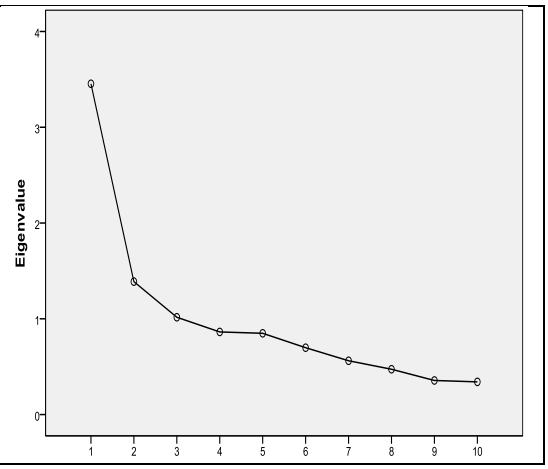
plan at later stages of the writing production is apparently an academic writing ability distinctive from the other conceptualisation processes.

5.2.2.2 The underlying structure of the meaning and discourse construction phase (real-life)

Meaning and discourse construction is a higher-level phase where the writer (1) contextualises abstract meanings based on the contextual clues provided in the writing task, (2) identifies what information (which could be retrieved from long-term memory or selected from input texts) is relevant to the context, and (3) identifies how information from different sources connects to each other and to the task (Field, 2013; Spivey, 1997). The *meaning and discourse construction* phase was measured by 11 questionnaire items in this study. The initial factor extraction for the *meaning and discourse construction* phase elicited in real-life academic conditions produced three factors with eigenvalues greater than 1.0. The scree plot suggested one- or two-factor solutions (See Table 5.4).

Table 5.4 Eigenvalues and scree plot for the meaning and discourse construction phase (real-life)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.454	34.541	34.541
2	1.388	13.880	48.420
3	1.016	10.161	58.582
4	.863	8.627	67.209
5	.849	8.485	75.694
6	.698	6.982	82.676
7	.562	5.619	88.295
8	.473	4.732	93.027
9	.356	3.560	96.587
10	.341	3.413	100.000



The one-factor solution was not examined. The rotated two-factor and three-factor solutions were compared. The two-factor solution (provided in Appendix 5.3 Table 1) showed that Factor 1 includes reading processes to select relevant ideas and some processes of *connecting and generate* while Factor 2 includes careful reading processes and a process of generating new

ideas or better understanding. However, three items loaded on both factors. Therefore the solution was rejected.

The three-factor solution, on the other hand, showed a clearer distinction between the factors. Based on the initial three-factor solution (see Table 5.5), Factor 1 includes mostly *connecting and generating* items, while Factor 2 and 3 include *reading* items. However, Item 4.5 and 1.1 loaded on more than one factor. Based on the reliability analyses presented in Section 3.4.1, both items are *reading* items. The fact that these two reading items loaded on more than one factor implies that these items might have contributed to other underlying constructs. Another reason could be that participants employed these reading processes differently from other reading processes. Future study should investigate into this issue. They were dropped from the factor analysis.

Table 5.5 Pattern matrix for the meaning and discourse construction phase (real-life): initial three-factor solution

Items		F1	F2	F3
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.	.739		
4.5	I selectively re-read the source texts while writing.	.588	.498	
2.9	I linked the important ideas in the source texts to what I know already.	.548		
4.2	I developed new ideas while I was writing.	.407		
4.3	I made further connections across the source texts while I was writing.	.274		
2.5	I read some relevant part(s) of the texts carefully.		.866	
2.7	I took notes on or underlined the important ideas in the source texts.		.569	
2.4	I searched quickly for part(s) of the texts which might help complete the task.		.462	
2.1	I read through the whole of each source text slowly and carefully.			.979
1.1	I read the whole task prompt (i.e. instructions) carefully.	.391		.451
2.2	I read the whole of each source text more than once.			.414

The final rotated three-factor pattern and interfactor correlations matrix is presented in Table 5.6.

Table 5.6 Pattern and the interfactor correlations matrix for the meaning and discourse construction phase (real-life)

		F1 Connecting and generating (33.27%)	F2 Selecting relevant ideas (12.53%)	F3 Careful global reading (9.57%)
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.	.665		
2.9	I linked the important ideas in the source texts to what I know already.	.528		
4.2	I developed new ideas while I was writing.	.514		
4.3	I made further connections across the source texts while I was writing.	.274		
2.5	I read some relevant part(s) of the texts carefully.		.767	
2.7	I took notes on or underlined the important ideas in the source texts.		.715	
2.4	I searched quickly for part(s) of the texts which might help complete the task.		.383	
2.1	I read through the whole of each source text slowly and carefully.			.846
2.2	I read the whole of each source text more than once.			.509
Interfactor correlations matrix				
	Factor 1 (Connecting and generating)	1.000		
	Factor 2 (Selecting relevant ideas)	.595	1.000	
	Factor 3 (Careful global reading)	.377	.223	1.000

Researchers argued that meaning and discourse construction involves high-level processes such as selecting relevant information, connecting ideas from different sources, and building a consistent discourse pattern (Field, 2003, 2008; Spivey, 1990, 1991, 1997). The results here showed that the *meaning and discourse construction* phase elicited on the real-life tasks in this study involved three distinct yet correlated cognitive processes. As shown in Table 5.6, Factor 1 includes four items of connecting ideas from different sources

and generating new representations. The factor was named *connecting and generating.*

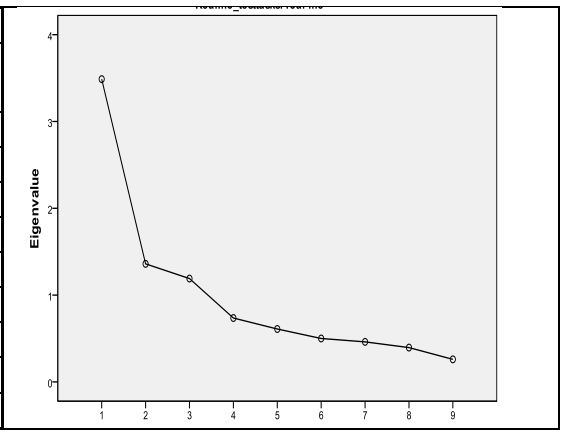
Factor 2 includes three reading items of identifying ideas which are relevant and important to the context of the writing task. The factor was named *selecting relevant ideas.* Factor 3 includes two global careful reading items. The factor was named *global careful reading.* According to the interfactor correlation matrix (See Table 5.6), Factor 2 (*selecting relevant ideas*) and Factor 3 (*global careful reading*) correlated weakly at 0.223. This supports the hypothesis that selective and search reading skills are different from global careful comprehension skills (Khalifa & Weir, 2009; Weir, Yang, & Jin, 2000). The results also reveal that Factor 1 (*connecting and generating*) correlated with Factor 2 (*selecting relevant ideas*) more than with Factor 3 (*global careful reading*). Both results indicate a need to test selective reading skills in EAP tests.

5.2.2.3 The underlying structure of the organisation phase (real-life)

Organisation is a phase 'where the writer provisionally organises the ideas, still in abstract form, (a) in relation to the text as a whole and (b) in relation to each other (Field, 2004, 329)'. The organisation construct was measured by 9 questionnaire items in this study. The initial factor extraction for the organisation construct produced three factors with eigenvalues greater than 1.0. The scree plot suggested one-, two- or three- factor solutions (See Table 5.7).

Table 5.7 Eigenvalues and scree plot for the organising phase (real-life)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.488	38.753	38.753
2	1.360	15.110	53.863
3	1.190	13.228	67.091
4	.736	8.178	75.268
5	.609	6.771	82.040
6	.500	5.557	87.596
7	.461	5.124	92.721
8	.395	4.393	97.113
9	.260	2.887	100.000



The rotated two- and three-factor solutions were compared. The three-factor solution (provided in Appendix 5.3 Table 2) showed that Factor 1 is related to organising main ideas. Factor 2 is related to organising the structure of the texts. However, Factor 3 is difficult to interpret. In addition, 2 items loaded on two factors. Therefore this solution was rejected.

The two-factor solution, on the other hand, provides a clearer distinction between the factors. Based on the initial two-factor solution (see Table 5.8), Factor 1 includes mostly the processes of organising ideas from the source texts while Factor 2 includes mostly the processes of organising ideas in relation to the writer's own text. However, Item 3.4 did not load on either of the factors at a level of 0.3 or above. It was dropped from the analysis.

Table 5.8 Pattern matrix for the organising phase (real-life): initial two-factor solution

Items		F1	F2
2.10	I worked out how the main ideas in each source text relate to each other.	.906	
2.11	I worked out how the main ideas relate across the source texts	.880	
2.8	I prioritised important ideas in the source texts in my mind.	.704	
2.3	I used my knowledge of how texts like the source texts are organised to find parts to focus on.	.482	
3.2	I recombined or reordered the ideas to fit the structure of my text.		.836
3.1	I organised the ideas for my text before starting to write.		.795
3.3	I removed some ideas I planned to write.		.688
4.1	While I was writing, I sometimes paused to organise my ideas.		.513
3.4	I tried to use the same structure as in the source texts to organise my text.		

After removal, the rotated two-factor solution was extracted (See Table 5.9).

Table 5.9 Pattern and interfactor correlations matrix for the organising phase (real-life)

		F1 Organising ideas in relation to input texts (34.73%)	F2 Organising ideas in relation to new text (16.60%)
2.11	I worked out how the main ideas relate across the source texts	.984	
2.10	I worked out how the main ideas in each source text relate to each other.	.761	
2.8	I prioritised important ideas in the source texts in my mind.	.564	
2.3	I used my knowledge of how texts like the source texts are organised to find parts to focus on.	.437	
3.3	I removed some ideas I planned to write.		.691
3.2	I recombined or reordered the ideas to fit the structure of my text.		.627
3.1	I organised the ideas for my text before starting to write.		.441
4.1	While I was writing, I sometimes paused to organise my ideas.		.314
Interfactor correlations matrix			
Factor 1 (Organising ideas in relation to input texts)		1.000	
Factor 2 (Organising ideas in relation to new text)		.533	1.000

Factor 1 involved four items of organising ideas in relation to a single input text or in relation to multiple input texts. The factor was named *organising ideas in relation to input texts.* Factor 2 involved four items of organising ideas in relation to the writer's own text. The factor was named *organising ideas in relation to own text.*

In the writing literature, the ability to organise ideas according to the communicative purpose of the task has been regarded as an important writing skill (e.g. Hayes & Flower, 1983; Shaw & Weir, 2007). In the reading-into-writing literature, the ability to organise has also been regarded as important when writers use external sources (e.g. Flower et al., 1990; Plakans, 2009; Segev-Miller, 2007; Spivey & King, 1989; Spivey, 1984).

The results of this study showed that the organisation phase elicited by the real-life tasks involved two distinct underlying cognitive processes - *organising ideas in relation to input texts* and *organising ideas in relation to own text.* This supports Spivey's notion that organising is a core process of writing from sources – a writer constructs their text by organising the ideas he/she selected from the input texts. The results of this study reveal that the process of organising ideas in relation to textual and intertextual representation of the input texts is distinctive from the process of organising ideas in relation to the writer's own text. The pattern matrix showed that the two types of organising processes correlate moderately at 0.533 (See Table 5.9). As both organising processes were elicited in the real-life academic conditions, it is important for EAP tests to sample both processes from the test takers. The process of organising in relation to textual and intertextual representation of the input texts has been neglected in most standardised academic writing tests which use the impromptu writing task type.

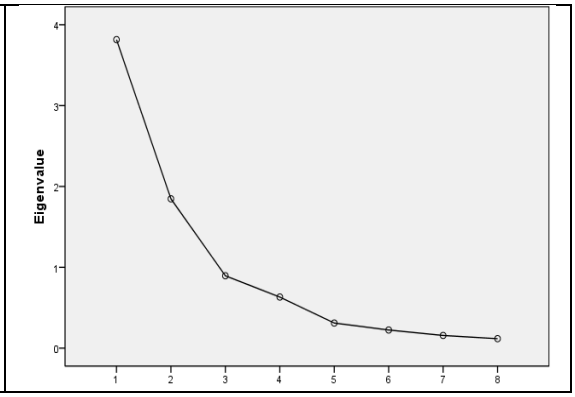
5.2.2.4 The underlying structure of the low-level monitoring and revising phase (real-life)

Low-level monitoring involves primarily checking the linguistic accuracy, e.g. spelling, grammar and sentence structure of the text. After monitoring, a writer will usually revise the unsatisfactory parts of the text (Field, 2004, 330). The low-level monitoring and revising processes can be done at any time during writing at the word, sentence or paragraph level, or after the whole draft has been completed. The low-level monitoring and revision construct was measured by 8 questionnaire items in this study.

The initial factor extraction for the low-level monitoring and revision construct produced two factors with eigenvalues greater than 1.0. The scree plot suggested two- or four- factor solutions (See Table 5.10).

Table 5.10 Eigenvalues and scree plot for the low-level monitoring and revision phase (real-life)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.816	47.703	47.703
2	1.847	23.087	70.790
3	.895	11.190	81.979
4	.632	7.899	89.879
5	.311	3.884	93.762
6	.225	2.808	96.571
7	.157	1.965	98.535
8	.117	1.465	100.000



The rotated two- and four- factor solutions were compared. The four-factor solution reflected to some extent the categorisation of different types of low-level monitoring and revising processes, e.g. grammar vs. use of own words. However, Factor 3 and Factor 4 consisted of one item only. In addition, two items loaded on two factors. Therefore, the four-factor solution (provided in Appendix 5.3 Table 3) was rejected.

On the other hand, the two-factor solution reflected clearly the distinction between the low-level monitoring and revising processes employed during the writing process and those employed after the whole draft has been completed (see Table 5.11).

Factor 1 contained four low-level monitoring and revising processes employed after the whole draft has been completed. The factor was named *low-level editing after writing.* Factor 2 contained four low-level monitoring and revising processes employed during the writing process. The factor was named *low-level editing during writing.*

Table 5.11 Pattern and interfactor correlations matrix for the low-level monitoring and revision phase (real-life)

		F1 Low-level editing after writing (47.70%)	F2 Low-level editing during writing (23.9%)
5.12	After I had finished the first draft, I checked that the quotations were properly made.	.898	
5.13	After I had finished the first draft, I checked that I had put the ideas of the source texts into my own words.	.872	
5.15	After I had finished the first draft, I checked the accuracy and range of the sentence structures.	.848	
5.16	After I had finished the first draft, I checked the appropriateness and range of vocabulary.	.847	
4.16	I checked the appropriateness and range of vocabulary while I was writing.		.908
4.15	I checked the accuracy and range of the sentence structures while I was writing.		.783
4.12	I checked that the quotations were properly made while I was writing.		.515
4.13	I checked that I had put the ideas of the source texts into my own words while I was writing.		.488
Interfactor correlations matrix			
Factor 1 (Low-level editing after writing)		1.000	
Factor 2 (Low-level editing during writing)		.347	1.000

In this study, the participants were asked to report the extent to which they revised different aspects of linguistic accuracy of their text while they were writing their text and after they had completed their first draft. Researchers (Field, 2004; Kellogg, 1996; Shaw & Weir, 2007) argued that monitoring and revising are highly demanding in terms of cognitive effect. Writers, especially L2 writers, tend to focus on one aspect of the text at a time due to short-term memory constraints. With attentional constraints, many writers would set aside the revising process to a later stage of the production. The results here confirm this notion. The results showed that the low-level editing processes employed by the participants in the real-life conditions clustered based on the

stages of the writing process when they employed these editing processes, i.e. *during the writing process* and *at the final stage after drafting the text.*

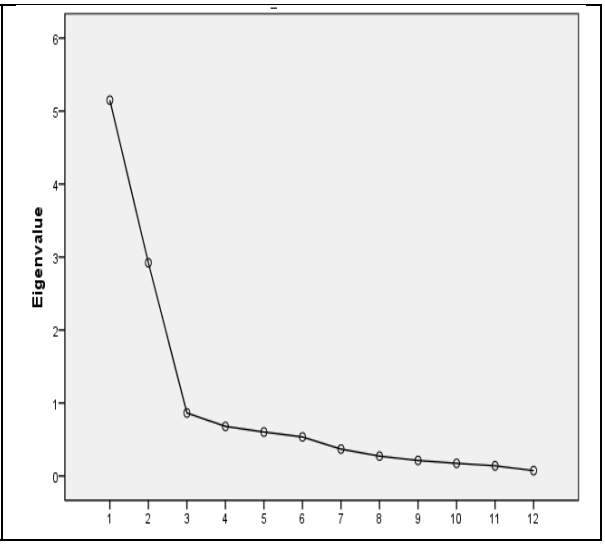
5.2.2.5 The underlying structure of the high-level monitoring and revising phase (real-life)

High-level monitoring involves primarily checking the effect of the text, such as clarity and appropriateness of ideas, coherence of arguments, style, and possible effect on reader. Similar to the low-level revising and monitoring construct, after high-level monitoring, a writer will usually revise the unsatisfactory parts of the text (Field, 2004, 330).

The high-level monitoring and revision phase elicited on the real-life tasks was measured by 12 questionnaire items in this study. The initial factor extraction for the high-level monitoring and revising construct produced two factors with eigenvalues greater than 1.0. The scree plot also suggested a two-factor solution (See Table 5.12).

Table 5.12 Eigenvalues and scree plot for the high-level monitoring and revising phase (real-life)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.150	42.920	42.920
2	2.922	24.351	67.271
3	.864	7.198	74.469
4	.681	5.673	80.141
5	.603	5.023	85.164
6	.536	4.464	89.628
7	.370	3.081	92.709
8	.272	2.271	94.980
9	.213	1.777	96.757
10	.175	1.454	98.211
11	.140	1.170	99.381
12	.074	.619	100.000



Similar to the low-level monitoring and revising phase, the rotated two-factor solution (See Table 5.13) produces simple and interpretable factor loadings which reflect the underlying distinction between the high-level editing processes employed while the participants were writing and those employed after they had completed the first draft.

Table 5.13 Pattern and interfactor correlations matrix for the high-level monitoring and revising phase (real-life)

		F1 High-level editing after writing (42.92%)	F2 High-level editing during writing (24.35%)
5.9	After I had finished the first draft, I checked that my text was coherent.	.903	
5.8	After I had finished the first draft, I checked that my text was well-organised.	.880	
5.11	After I had finished the first draft, I checked that I included my own viewpoint on the topic.	.877	
5.7	After I had finished the first draft, I checked that the content was relevant.	.865	
5.10	After I had finished the first draft, I checked that I included all appropriate main ideas from all the source texts.	.797	
5.14	After I had finished the first draft, I checked the possible effect of my writing on the intended reader.	.748	
4.9	I checked that my text was coherent while I was writing.		.747
4.7	I checked that the content was relevant while I was writing.		.618
4.8	I checked that my text was well-organised while I was writing.		.613
4.10	I checked that I included all appropriate main ideas from all the source texts while I was writing.		.581
4.11	I checked that I included my own viewpoint on the topic while I was writing.		.575
4.14	I checked the possible effect of my writing on the intended reader while I was writing.		.446
Interfactor correlations matrix			
Factor 1 (High-level editing after writing)		1.000	
Factor 2 (High-level editing during writing)		.208	1.000

Factor 1 contained six high-level monitoring and editing processes employed after the whole draft has been completed. The factor was named *high-level editing after writing.* Factor 2 contained six high-level monitoring and revising processes employed during the writing process. The factor was named *high-level editing during writing.*

5.2.2.6 Summary of the underlying structure of the cognitive processes (real-life)

The results of the underlying structure of the cognitive processes involved in the five writing phases elicited during real-life academic writing tasks have been discussed. Exploratory factor analysis proved invaluable in providing empirical information about the clustering of the individual processes within each writing phase elicited by the real-life tasks. The results showed that the hypothesised academic writing phases arising from the literature review were largely supported by the statistical analysis of the questionnaire data collected in this study. The majority of cognitive processes loaded on their corresponding writing phases at a level of 0.3 or above. Results of EFA here revealed that each hypothesised academic writing phase elicited under the real-life academic conditions in this study involved two or more distinct yet correlated underlying cognitive processes. Table 5.14[8] summarises the EFA-generated underlying structure of the real-life academic writing cognitive constructs.

Table 5.14 Summary of the EFA-generated underlying structure of the real-life academic writing processes

Academic writing phases	Cognitive processes
Conceptualisation	F1: Task representation and macro-planning (34%)
	F2: Revising macro plan (19.9%)
Discourse construction	F1: Connecting and generating (34.54%)
	F2: Selecting relevant ideas (13.88%)
	F3: Careful global reading (10.16%)
Organisation	F1: Organising ideas in relation to input texts (34.73%)
	F2: Organising ideas in relation to own text (16.60%)
Low-level monitoring and revising	F1: Low-level editing after writing (47.70%)
	F2: Low-level editing during writing (23.9%)
High-level monitoring and revising	F1: High-level editing after writing (42.92%)
	F2: High-level editing during writing (24.35%)

The percentage in brackets indicates the extent to which each factor (i.e. cognitive process) accounts for the variance (i.e. how participants reported employing the cognitive processes within each cognitive phase). In summary,

[8] The structure reflects the factor loadings of how individual processes clustered together based on the exploratory factor analyses, and not the original categorisation of the questionnaire items.

the conceptualisation phase in the real-life academic context involved the *task representation and macro-planning* processes to conceptualise an understanding of the writing task and establish their macro plans. In addition, the process of *revising macro plan* is particularly important in real-life academic writing. The two processes explained 53.9% of the variance of the conceptualisation phase. The meaning and discourse construction phase involved three underlying processes: *connecting and generating, selecting relevant ideas* and *global careful reading.* The three processes explained 58.58% of the variance of the meaning and discourse construction phase. The *organising* phase involved the processes of *organising ideas in relation to input texts* as well as *organising ideas in relation to writer's own text*. The two processes explained 51.33% of the variance of the organising phase. For both low-level monitoring and revising and high-level monitoring and revising phases, there was a clear distinction between the *editing processes employed while writing* and those employed *after the first draft has been completed*. The two low-level monitoring and revising processes explained 71.6% of the variance of the low-level monitoring and revising phase whereas the two high-level monitoring and revising processes explained 67.27% of the variance of the low-level monitoring and revising phase. While the results of this study identified 11 processes involved in the five cognitive phases elicited by the real-life academic writing tasks, the results indicated that the variance of each cognitive phase was not fully accounted. Future study should explore the additional cognitive processes involved in each phase, perhaps with a different research method, such as keystroke logging and stimulated recall.

5.2.3 Further comparisons of the cognitive process elicited by the two real-life tasks

Section 5.2.1 has showed that participants in this study rated the extent to which they employed the individual 48 questionnaire items similarly on the two real-life tasks (*essay* and *report*). Section 5.2.2 has identified the eleven underlying cognitive processes involved in five phases of academic writing. This sub-section further examines how the participants employed these eleven cognitive processes on the two real-life tasks. The means of the average rating (*4=definitely agree; 3=mostly agree; 2=mostly disagree; 1=definitely*

disagree) of the cognitive processes employed by the participants on the real-life essay and report tasks are presented in Figure 5.1.

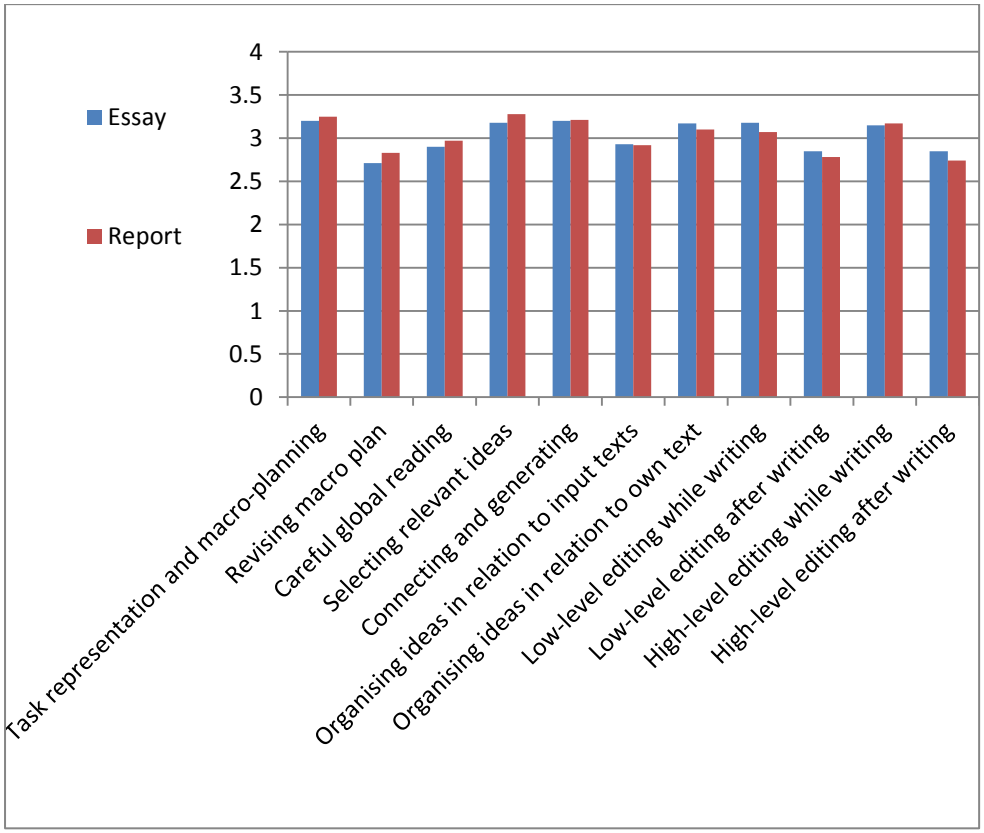


Figure 5.1 Comparison between the 2 real life tasks in terms of the cognitive processes employed

As indicated in Figure 5.1, with regards to the conceptualisation phase, the participants reported employing the processes of *task representation and macro-planning* slightly more on the report task than on the essay task. They also reported *revising macro plan* more on the report task than on the essay task. Regarding the meaning and discourse construction phase, the participants reported employing the processes of *connecting and generating* to a similar extent between the two real-life tasks. However, the report task seemed to have elicited the processes of *careful global reading* and *selecting relevant ideas* slightly more than the essay task did. With regards to the organising phase, the participants reported the extent they *organised ideas in relation to the input texts* similarly between the two real-life tasks. However, they reported *organising ideas in relation to own text* more on the essay task than the report task. Regarding the two monitoring and revising phases, the essay task seemed to have elicited from the participants the processes of *low-level*

editing while writing, *low-level editing after writing* and *high-level editing after writing* more than the report task did.

Shaw & Weir (2007) argued that among other task features, the presentation of a clear communicative purpose for completing the task and a clear intended reader would engage students in active macro planning. As previously presented, the expert judgement on the overall task setting showed that the report task presented the students with a clearer purpose and a clear intended reader. Their argument that a writing task with clear communicative purpose and intended reader would encourage macro-planning is supported by the results of this study.

The differences may also be due to the amount of reading required by the tasks. The participants' qualitative responses on the questionnaire suggested that they had to read more materials for the report task than for the essay task. The materials included textbooks about business theories and analysing techniques, some basic information about the business of a company provided by the lecturer, and some additional information about the company or other similar companies searched by them. On the other hand, for the essay task, they were given a single article to read. They were required to search for additional articles in order to complete the task but they commented that five articles would be enough. Due to the different amount of external reading required by the tasks, the participants seemed to have devoted their attention slightly differently. They put more attention on the source texts for the report task but more attention on their own text for the essay task.

The descriptive comparison of the cognitive processes employed on the two real-life tasks has been discussed so far. The difference obtained between the means of each process was analysed by Mann-Whitney U test (independent samples). The results from the Mann-Whitney U tests together with the means and standard deviations[9] of the average rating are presented in Table 5.15.

[9] Regarding the results of the Mann-Whitney U tests, median value is usually reported instead of mean and standard deviation. However, means and standard deviations are reported here for better consistency with other results in this thesis.

Table 5.15 Comparison of the cognitive processes employed between the two real-life tasks (inferential)

	Report (n=73)		Essay (n=70)		Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
	Mean	Std Dev	Mean	Std Dev				
Conceptualisation								
Task representation and macro-planning	3.25	.52	3.20	.47	2348.500	4833.500	-.839	.401
Revising macro plan	2.83	.68	2.71	.74	2325.500	4810.500	-.949	.342
Meaning and discourse construction								
Careful global reading	2.97	.73	2.90	.65	2379.000	4864.000	-.728	.467
Selecting relevant ideas	3.28	.58	3.18	.54	2189.000	4674.000	-1.507	.132
Connecting and generating	3.21	.57	3.20	.44	2396.000	4881.000	-.647	.518
Organisation								
Organising ideas in relation to source texts	2.92	.55	2.93	.43	2407.000	4892.000	-.601	.548
Organising ideas in relation to new text	3.10	.57	3.17	.45	2432.000	5133.000	-.502	.616
Low-level monitoring and revising								
Low-level editing while writing	3.07	.65	3.18	.56	2377.500	5078.500	-.729	.466
Low-level editing after writing	2.78	.88	2.85	.93	2360.500	5061.500	-.793	.428
High-level monitoring and revising								
High-level editing while writing	3.17	.59	3.15	.49	2413.500	4898.500	-.574	.566
High-level editing after writing	2.74	.89	2.85	.95	2270.000	4971.000	-1.158	.247

As shown in Table 5.15, the results from the Mann-Whitney U tests showed that all differences between the two real life tasks in terms of the mean ratings of each process were non significant ($p>0.05$). In other words, apart from the descriptive differences discussed above, the participants rated the extent to which they used the eleven cognitive processes similarly between the two real-life tasks. Even with considerable differences in contextual features as reported in the last chapter, the two real-life tasks elicited the same cognitive processes from the participants to a similar extent. The findings seem to suggest that factors other than immediate contextual features might have impacted on the way the participants employed the processes. Possible factors

include a common academic writing condition applied to both real-life tasks and the level of writer's academic writing experience. Future studies might take these factors into consideration. The findings, on the other hand, confirm that all eleven cognitive processes are important for completing real-life academic writing tasks which require the use of external reading materials.

In order to better define the target academic writing cognitive processes, it is worth discussing a few similar patterns in how participants employed the cognitive processes between the two real-life tasks.

In terms of the meaning and discourse construction phase, the participants reported employing the processes of expeditiously selecting relevant ideas more than careful reading processes on both real-life tasks. Urquhart & Weir (1998) argued for the importance of quick, efficient and selective reading, i.e. expeditious reading. Weir et al (2013) in their book reviewing the history of testing observed that reading tests in the first part of the 20th century tended to target careful local reading at the clause and sentence level rather than careful global reading, and rarely expeditious forms of reading beyond the United States (Urquhart & Weir 1998, Khalifa & Weir 2009; Moore, Morton & Price 2010).

Second, in terms of the organising phase, the participants reported organising ideas in relation to their own text more than the extent to which they organised ideas in relation to the input texts on both real-life tasks.

Lastly, in terms of the monitoring and revising phases, participants reported editing (at both low and high levels) while they were writing more than the extent to which they edited after the first draft had been completed.

In the later discussion of the processes elicited under test conditions, attention should be paid to these similar patterns observed on the two real-life tasks.

5.2.4 Comparisons between high-achieving and low-achieving participants

The purpose of RQ2 was to define the cognitive parameters which are appropriate for reading-into-writing tests for academic purposes. The above results of exploratory factor analysis have defined the eleven cognitive processes which students employed to complete the real-life academic writing tasks. A further step was to investigate whether these defined cognitive parameters can effectively distinguish how more proficient and less proficient writers employ these processes.

The 143 participants who completed either one of the real-life tasks were ranked according to their scores. Each performance can be scored from 0-16 (Participants' performances will be presented in detail in Chapter Six). The participants were divided into four groups (i.e. high, higher middle, lower middle and low) representing roughly 25% of the population of participants who completed the tasks (i.e. the exact number of participants in each group varies). As a result, 40 participants were classified as the high-achieving group (a score of 12 or above) and 39 participants were classified as the low-achieving group (a score of 7.5 or below). The means and standard deviations of the average rating (*4=definitely agree; 3=mostly agree; 2=mostly disagree; 1=definitely disagree*) of the eleven underlying cognitive process reported by the high-achieving and low-achieving groups are presented in Table 5.16. The means reported by the two groups were analysed by Mann-Whitney U tests (See Table 5.16).

Table 5.16 Comparisons between high-achieving and low-achieving participants

	High-achieving (n=40)		Low-achieving (n=39)		Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
	Mean	Std Dev	Mean	Std Dev				
Conceptualisation								
Task representation and macro-planning	3.29	.442	3.09	.474	585.500	1365.500	-1.917	.055
Revising macro plan	2.78	.847	2.80	.713	747.000	1567.000	-.330	.741
Meaning and discourse construction								
Careful global reading	3.09	.823	2.91	.648	639.500	1419.500	-1.412	.158
Selecting relevant ideas	3.47	.488	3.10	.525	474.500	1254.500	-3.058	.002
Connecting and generating	3.35	.432	3.12	.487	553.500	1333.500	-2.243	.025
Organising								
Organising ideas in relation to input texts	3.03	.443	2.86	.399	521.500	1301.500	-2.552	.011
Organising ideas in relation to own text	3.28	.507	3.01	.503	542.000	1322.000	-2.360	.018
Low-level monitoring and revising								
Low-level editing while writing	3.36	.543	3.00	.691	522.500	1302.500	-2.578	.010
Low-level editing after writing	2.83	.866	2.84	.920	747.500	1567.500	-.322	.747
High-level monitoring and revising								
High-level editing while writing	3.30	.467	3.01	.540	553.500	1333.500	-2.233	.026
High-level editing after writing	2.82	.857	2.84	.915	749.000	1569.000	-.306	.760

As shown in Table 5.16, the high achieving participants reported employing eight of the eleven cognitive processes (i.e. *task representation and macro-planning, careful global reading, selecting relevant ideas, connecting and generating, organising ideas in relation to input texts, organising ideas in relation to own text, low-level editing while writing* and *high-level editing while writing*) more than the low achieving groups. Apart from *task representation and macro-planning* and *careful global reading*, all differences were significant. There was not much difference in the average rating of the processes of *revising macro plan, low-level editing after writing* and *high-*

level editing after writing between the high-achieving and low-achieving groups (See Table 5.16).

For the conceptualisation phase, the high-achieving participants employed the processes of *task representation and macro-planning* more than the low-achieving participants, though the difference was not significant. On the other hand, there was not much difference in the extent to which they employed the processes of *revising macro plan* at later stages of the writing process. According to Scardamalia & Bereiter's (1987) continuum of writing expertise, mature writers are cautious about the rhetorical situation of the writing task and therefore actively establish writing goals to fulfil the communicative purpose of the task. In contrast, most immature writers might not be able to establish a complete representation the writing task (e.g. communicative purpose, intended readership, genre, content and style) and therefore would not establish comprehensive macro plans before they start to write. The results here seem to support this notion. Researchers (Field, 2004; Kellogg, 2001; Shaw & Weir, 2007) further argued that due to attentional constraints, L2 writers may have extra difficulty in building a comprehensive task representation and establishing macro plans because most of their of attention may be devoted to lower-level reading processes, e.g. lexical decoding and parsing (i.e. connecting words to meaning). Another challenge for weaker L2 writers is that they may not be able to hold their task representation and macro-plan in working memory while they are executing their plans (Field, 2013).

With regards to the meaning and discourse construction phase, the high-achieving participants reported employing the processes of *careful global reading, selecting relevant ideas* and *connecting and generating* more than the low-achieving group (See Table 5.16). Field's model of receptive skills (2004, 2011, 2013) argued that meaning construction and discourse construction are more cognitively demanding than decoding, lexical searching and parsing. He argued that less proficient language users would focus on the lower processes while the proficient language users would have high automaticity in executing these lower processes. The proficient language users would then be able to focus on meaning construction and discourse construction when they read.

Khalifa & Weir's (2007) model of reading skills similarly argued that lower proficiency L2 readers would focus on establishing understanding at local level (words, phases, sentences) while high proficiency readers would be able to establish understanding at global textual and intertextual level. In addition, researchers (e.g. Spivey 1991) argued that mature writers are able to connect ideas from different internal sources (e.g. topical knowledge and discourse knowledge) and external sources (e.g. input texts) to the context of writing task and generate new understanding of the rhetorical challenge presented by the writing task. Scardamalia & Bereiter (1987) argued that this is why mature writers are able to transform knowledge as they write. The results of this study showed that in the real-life academic context, these high-level processes of careful and selective reading skills at global textual and intertextual level and the processes of connecting and generating distinguish the high-achieving the low-achieving participants. In other words, these high level processes are important for success in academic writing. Therefore, it is important that these processes are targeted in academic writing tests.

With regards to the organising phase, the high-achieving participants reported organising ideas both in relation to input texts and the writer's own text significantly more than the low-achieving participants.

Regarding the low-level and high level monitoring and revising phases, the high-achieving participants reported employing the processes of *while writing* low-level editing and high-level editing significantly more than the low-achieving participants did. However, there was not much difference in the extent to which they reported the use of *after writing* low-level and high-level editing. Field (2004) argued that monitoring and revising processes are highly cognitively demanding, particularly for L2 writers. While monitoring and revising can be employed at any time during the writing process, most writers can only focus on one aspect of the editing at one time (e.g. grammatical accuracy at low level, or argument coherence at high level). Therefore, weaker writers who have not acquired high automaticity in the translating process tend not to be able spare attention on monitoring and revising. The results reported here confirm this notion. The high-achieving participants in this study seemed

to be more capable of performing editing at both low- and high- levels *while they were writing* than the low-achieving participants.

5.2.5 Summary of the results of real-life academic writing processes

This chapter has thus far discussed the results of the real-life academic writing processes. The main results can be summarised as follows.

The findings of the study have identified the cognitive processes employed at each of the five hypothesised writing phases when completing real-life academic tasks. Each phase involved two or more distinct yet correlated underlying cognitive processes. The conceptualisation phase involved the processes of *task representation and macro-planning* to conceptualise an understanding of the writing task and establish macro plans, and the processes of revising macro plans at later stages of the writing production. The meaning and discourse construction phase involved three underlying processes: *connecting and generating, selecting relevant ideas* and *global careful reading.* The organising phase involved the processes of *organising ideas in relation to the input texts* as well as *organising ideas in relation to writer's own text*. For both low-level monitoring and revising and high-level monitoring and revising, each phase involved *editing processes employed while writing* and those employed *after the first draft has been completed*. The findings suggested that these eleven cognitive processes would be the target cognitive processes for a valid academic writing test.

The findings showed that the participants reported the extent to which they employed these eleven cognitive processes on the two real-life tasks with some descriptive differences. For example, the report task, which was regarded by the judges to have clearer information about communicative purpose and intended reader than the essay task, seemed to have engaged the participants in macro-planning and revising macro plans slightly more than the essay task. In addition, the report task, which required the use of more extensive reading materials than the essay task, seemed to have engaged the participants in the processes of careful reading and selective reading more than the essay task. The essay task, on the other hand, seemed to have engaged the participants to organising ideas in relation to their own text and to edit their

text more than the report task. While individual contextual features seemed to have influenced to some extent how the participants employed the processes on the two real-life tasks, the differences described were non significant. Future studies which aim to explore in detail how individual contextual features might impact on the use of cognitive processes might need to consider other variables. However, for the purpose of this study, the findings support the importance of targeting these cognitive processes in academic writing tests.

In addition, the results have shown some similar patterns of how the participants employed the cognitive processes on both real-life tasks. These patterns need to be addressed in the discussion of cognitive processes elicited under conditions later in the chapter. The pattern included: (i) the participants reported employing the processes of *selecting relevant ideas* more than *careful reading*; (ii) the participants reported employing the processes of *organising ideas in relation to their own text* more than *organising ideas in relation to the input texts*, and (iii) the participants reported employing editing (at both low- and high-level) *while writing* more than *after the first draft had been completed.*

Last but not least, the results showed that the high achieving participants reported employing eight of the eleven cognitive processes more than low achieving participants, the difference in six processes was significant. Academic performance is known to be impacted by many factors other than academic writing ability. The results of this study, however, showed that there are significant differences in the extent to which the high-achieving and low-achieving students employed the cognitive processes. This argues strongly for the need to have cognitively valid academic writing tests which assess the same cognitive processes that are required in real-life conditions.

The chapter will next examine the cognitive processes elicited by the two reading-into-writing tests and discuss the cognitive validity of the task types.

5.3 Investigating the cognitive validity of reading-into-writing tasks

RQ2 aims to examine the extent to which the two types of reading-into-writing test tasks elicited the same cognitive processes as the real-life writing tasks did in this study. Therefore, after investigating the cognitive constructs elicited in real-life academic conditions, the next step was to investigate the cognitive processes that the participants used when they were completing the two types of reading-into-writing tasks under test conditions. 160 participants completed Test Task A (essay with multiple verbal inputs) and 140 participants completed Test Task B (essay with multiple verbal and non-verbal inputs) (see Appendix 3.1.3 and Appendix 3.1.4 for a sample of the two test tasks). They filled out the Cognitive Process Questionnaire (Appendix 3.4) immediately after they had completed the test tasks.

Based on the results from the exploratory factor analyses of the real-life processing data, this study identified eleven cognitive parameters which are appropriate for reading-into-writing tests of academic writing: (1) task representation and macro-planning, (2) revising macro plan, (3) connecting and generating, (4) selecting relevant ideas, (5) careful global reading, (6) organising ideas in relation to input texts, (7) organising ideas in relation to own text, (8) low-level editing during writing, (9) low-level editing after writing, (10) high-level editing during writing, and (11) high-level editing after writing. The results also showed that these parameters distinguished well how high-achieving and low-achieving students completed the real-life tasks.

The following Section 5.3.1 presents the results of the comparison of how these cognitive processes were employed by all participants as a whole group under the test and real-life conditions. Section 5.3.2 reports the investigation of how well the cognitive parameters distinguished the cognitive processes employed by high-achieving test takers and those employed by low-achieving test tasks on each of the reading-into-writing test tasks. Section 5.3.3 reports a further comparison of cognitive processes elicited under test and real-life conditions in groups of high-, medium- and low-achievement. The results reported in these three sub-sections (5.3.1-5.3.3) will indicate if these eleven cognitive parameters can a) distinguish how high-achieving and low-achieving

test takers completed the individual reading-into-writing test task; and b) sufficiently resemble the cognitive processes which the test takers (as a whole group as well as in groups of high-, medium, or low- achievement) would normally employ in non-test conditions.

Section 5.3.4 presents the results of the exploratory factor analysis of the cognitive processes elicited by the test tasks. The results will review the underlying structure of the cognitive processes elicited by the reading-into-writing test task types. The extent to which the two reading-into-writing test task types elicited the same underlying cognitive processes as those reviewed by the real-life data was discussed thoroughly.

5.3.1 Comparisons of the cognitive processes elicited under test and real-life conditions (whole group)

The means and standard deviations of the average rating (*4=definitely agree; 3=mostly agree; 2=mostly disagree; 1=definitely disagree*) of these processes reported by all participants as a whole group on Test Task A and the real-life tasks are presented in Table 5.17. The differences obtained between test and real-life conditions were then analysed by Wilcoxon signed ranks test (non-parametric related sample) (See Table 5.17). The corresponding results from Test Task B and real-life tasks are presented in Table 5.18.

5.3.1.1 Comparison of the cognitive processes employed on Test Task A and real-life tasks (Whole group)

As shown in Table 5.17, the participants reported employing all the eleven cognitive processes more on a scale from 1 (*definitely disagree*) to 4 (*definitely agree*) in the real-life conditions than under the test condition. The differences reported in six processes (which include *revising macro plan*, *organising ideas in relation to own text*, *low-level editing during writing*, *low-level editing after writing*, *high-level editing during writing*, and *high-level editing after writing*) were significant ($p<0.05$) (see Table 5.17).

Table 5.17 Comparison of the cognitive processes employed between Test Task A and real-life tasks (whole group)

	Test Task A (n=160)		Real-life tasks (n=142)		Z	Asymp. Sig. (2-tailed)
	Mean	Std Dev	Mean	Std Dev		
Conceptualisation						
Task representation and macro-planning	2.92	.66	3.17	.49	-1.814	.070
Revising macro plan	2.56	.71	2.78	.71	-2.890	.004
Meaning and discourse construction						
Careful global reading	2.86	.74	2.94	.69	-.577	.564
Selecting relevant ideas	3.22	.82	3.24	.56	-1.403	.161
Connecting and generating	2.89	.70	3.20	.51	-1.861	.063
Organising						
Organising ideas in relation to input texts	2.91	.62	2.95	.49	-1.069	.285
Organising ideas in relation to own text	3.01	.72	3.13	.52	-2.355	.019
Low-level monitoring and revising						
Low-level editing while writing	2.81	.74	3.12	.61	-3.562	.000
Low-level editing after writing	2.43	.99	2.80	.90	-3.595	.000
High-level monitoring and revising						
High-level editing while writing	2.86	.72	3.16	.54	-3.889	.000
High-level editing after writing	2.44	1.04	2.79	.92	-2.808	.005

In terms of the conceptualisation phase, the difference in the use of *task representation and macro-planning* processes between real-life and test conditions was non significant. However, the participants revised their macro plans significantly more on the real-life tasks than on Test Task A (See Table 5.17) though the differences were for the most part slight in real terms. In terms of the meaning and discourse construction phase, the participants reported using the processes of *careful global reading*, *selecting relevant ideas*, and *connecting and generating* more on the real-life tasks than on Test Task A. However, the differences were non significant (see Table 5.17). As mentioned earlier, there has been a call to test careful global reading and expeditious selective reading skills (Khalifa & Weir, 2009; Moore et al., 2010; Urquhart & Weir, 1998). The results suggested that the reading-into-writing test type would be a good format to test these reading skills. In addition, the ability to *connect and generate* is an important skill when writing from sources (Spivey,

1990; Spivey & King, 1989). Expert writers, who are able to transform existing knowledge into new knowledge as a result of their writing process, tend to purposefully connect ideas from different resources (Scardamalia & Bereiter, 1987). The process of *connecting and generating* is believed to facilitate the occurrence of knowledge transformation. One key function of academic writing is to create knowledge (Weigle, 2004). For knowledge transformation to take place, a writer has to *connect* the relevant ideas which have been *selected* from external input texts and their internal personal knowledge based on the communicative purpose of the writing task.

Regarding the organising phase, the participants reported *organising ideas in relation to the input texts* to a similar extent in both real-life and test conditions. However, the mean rating of *organising in relation to the writer's own text* was significantly higher on the real-life tasks than on Test Task A (See Table 5.17) though again in real terms the difference was slight. With respect to the monitoring and revising phase, the participants reported employing *low-level editing while writing, low-level editing after writing, high-level editing while writing*, and *high-level editing after writing* significantly more in the real-life conditions than on Test Task A (See Table 5.17).

It must be remembered of course that in real life students have considerably more time to monitor and revise and iteratively revisit their work. Few tests give dedicated time for either planning or monitoring and exam tasks are performed under serious time pressure. Time for monitoring and revising is more limited in the exam situation.

5.3.1.2 Comparison of the cognitive processes on Test Task B and real-life tasks (whole group)

The participants reported employing eight cognitive processes more on a scale from 1 (*definitely disagree*) to 4 (*definitely agree*) in the real-life conditions than on Test Task B (See Table 5.18) though the differences were not substantial.

Table 5.18 Comparison of the cognitive processes between Test Task B and real-life tasks (whole group)

	Test Task B (n=140)		Real-life tasks (n=142)		Z	Asymp. Sig. (2-tailed)
	Mean	Std Dev	Mean	Std Dev		
Conceptualisation						
Task representation and macro-planning	3.12	.55	3.17	.49	-1.108	.268
Revising macro plan	2.61	.71	2.78	.71	-1.450	.147
Meaning and discourse construction						
Careful global reading	3.17	.66	2.94	.69	-2.429	.015
Selecting relevant ideas	3.28	.63	3.24	.56	-.421	.674
Connecting and generating	2.87	.64	3.20	.51	-4.658	.000
Organising						
Organising ideas in relation to input texts	3.07	.57	2.95	.49	-2.118	.034
Organising ideas in relation to own text	3.01	.68	3.13	.52	-2.254	.024
Low-level monitoring and revising						
Low-level editing while writing	2.89	.73	3.12	.61	-3.063	.002
Low-level editing after writing	2.37	1.06	2.80	.90	-3.995	.000
High-level monitoring and revising						
High-level editing while writing	2.99	.66	3.16	.54	-2.892	.004
High-level editing after writing	2.40	1.06	2.79	.92	-3.503	.000

As shown in Table 5.18, the extent to which the participants reported employing the processes of *task representation and macro-planning* and *revising macro plan* were slighter higher for the conceptualisation phase in the real-life conditions than on Test Task B. However, the differences were non significant. Regarding the meaning and discourse construction phase, the participants reported employing the processes of *careful global reading* and *selecting relevant ideas* more on Test Task B than on the real-life tasks. The difference in careful global reading was significant. This could be because the input texts on Test Task B were comparatively short due to the inclusion of non-verbal inputs. The seeming over-eliciting of careful global reading on Test Task B should be noted and reviewed. On the other hand, the participants reported the use of the processes of *connecting and generating* significantly more on the real-life tasks than on Test Task B.

In terms of the organising phase, the participants reported *organising ideas in relation to the input texts* more on Test Task A than on the real-life tasks. In contrast, they reported *organising ideas in relation to writer's own text* more on the real-life tasks than Test Task B. Both differences were significant ($p<0.05$). With respect to the monitoring and revising phases, similar to Test Task A, the participants reported employing *low-level editing while writing, low-level editing after writing, high-level editing while writing*, and *high-level editing after writing* significantly ($p<0.05$) more in the real-life conditions than on Test Task B. However, the caveat expressed above about the time available for this applies equally here as well.

Comparison of the extent to which the participants (as a whole group) in this study employed the cognitive processes under test conditions compared to the real-life conditions has been reported so far. Further comparison will be reported and discussed in Section 5.3.3 with consideration of the resulting performances. The results so far showed that the participants tended to employ cognitive processes more on a scale from 1 (*definitely disagree*) to 4 (*definitely agree*) in real-life conditions than under test conditions (both Test Task A and Test Task B). The results might be expected because of the much tighter time constraints under test conditions when compared to real-life conditions.

The common discrepancies between real-life conditions and the two test conditions were that the participants reported employing the processes of *connecting and generating, organising ideas in relation to own text, low-level editing while writing, high-level editing while writing* and *high-level editing after writing* significantly more on the real-life tasks than on the Test Task A or Test Task B. Lastly, there seemed to be an over-eliciting of the processes of careful global reading on Test Task B possibly due to shortness of text.

The next section presents the results of the extent to which these cognitive parameters distinguished the cognitive processes employed by high- and low-achieving participants on the two test tasks. Unlike the other productive language skill - speaking, the processes of writing are not assessed in most standardised writing tests. The processes of writing are sometimes assessed

indirectly in some classroom-based assessments by requiring students to submit multiple drafts. These multiple drafts can reflect to some extent how writers produce a text, especially the processes of organising ideas and editing. It is beyond the scope of this thesis to discuss about classroom-based assessments. Nevertheless, even though most standardised writing assessments for academic purposes might aim to evaluate the products rather than the processes of writing, it is essential for test developers to demonstrate empirical evidence which indicates a relationship between the processes the test takers perform on the test task and the scores they eventually receive.

5.3.2 Comparisons of the cognitive processes employed by high- and low-achieving groups (test tasks)

As previously discussed, results on the real-life tasks showed that high-achieving participants reported employing eight of the eleven cognitive processes more than low-achieving students, six of the differences were significant (see Section 5.2.4). The next step was to investigate if these cognitive parameters distinguished the high-achieving and low-achieving test takers.

The 160 participants who completed Test Task A and the 140 participants who completed Test Task B were ranked according to their scores (Their performances will be presented and discussed in Chapter Six). They were divided into four groups (i.e. high, higher middle, lower middle and low achieving) representing roughly 25% of the population each. As a result, on Test Task A, 36 participants were classified as high-achieving (with a score of 10.5 or above out of 20) and 34 participants were classified as low-achieving (with a score of 7.5 or below). On Test Task B, 27 participants were classified as the high-achieving group (with a score of 7 or above out of 9), and 34 were classified as the low-achieving group (with a score of 3 or below)[10]. The comparisons between how the high-achieving and low-achieving participants employed the cognitive processes on Test Task A are shown in Table 5.19, the results on Test Task B in Table 5.20.

[10] There is a noticeable disparity between the cut of scores of different levels of achievement on the two test tasks because Test Task A is a level-specific criterion-referenced test whereas Test Task B is a university diagnostic test.

5.3.2.1 Comparison between high- and low- achieving groups on Test Task A

On Test Task A, the high-achieving participants reported employing six of the processes (which include *revising macro plan*, *selecting relevant ideas*, *organising ideas in relation to own text, low-level editing while writing, low-level editing after writing,* and *high-level editing after writing*) more than the low-achieving group. However, all differences were non-significant ($p>0.05$) (See Table 5.19).

Table 5.19 Comparison between high-achieving and low-achieving (Test Task A)

	High-achieving (n=36)		Low-achieving (n=34)		Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
	Mean	Std Dev	Mean	Std Dev				
Conceptualisation								
Task representation and macro-planning	3.04	.54	3.06	.54	609.500	1275.500	-.030	.976
Revising macro plan	2.56	.59	2.53	.82	586.000	1181.000	-.313	.754
Meaning and discourse construction								
Careful global reading	2.74	.91	3.16	.59	461.500	1127.500	-1.813	.070
Selecting relevant ideas	3.51	.44	3.33	.70	545.500	1140.500	-.803	.422
Connecting and generating	2.92	.65	2.95	.68	589.000	1255.000	-.273	.785
Organisation								
Organising ideas in relation to input texts	3.07	.46	3.08	.53	599.500	1265.500	-.148	.882
Organising ideas in relation to own text	3.20	.68	3.00	.69	510.000	1105.000	-1.235	.217
Low-level monitoring and revising								
Low-level editing while writing	3.02	.67	2.82	.78	528.000	1123.000	-.994	.320
Low-level editing after writing	2.43	.90	2.23	1.15	553.500	1148.500	-.698	.485
High-level monitoring and revising								
High-level editing while writing	3.02	.68	3.06	.55	608.000	1203.000	-.047	.962
High-level editing after writing	2.56	1.09	2.20	1.11	496.500	1091.500	-1.381	.167

On the other hand, the low-achieving group reported using the processes of *task representation and macro-planning, careful global reading, connecting and generating, organising ideas in relation to source texts* and *high-level editing while writing* more than the high-achieving group. All differences were non significant ($p>0.05$). Apart from the process of careful global reading, the differences were very slight. The high achieving test takers reported an average rating of 2.74 for their use of careful global reading on Test Task A whereas the low achieving test takers reported an average rating of 3.16. The low achieving test takers might have relied too much on careful global reading under the test conditions. The results showed no significant difference in the extent to which high-achieving and low-achieving participants employed the processes on Test Task A. The finding is perhaps unexpected. However, while interpreting the results, it should be noted that Test Task A was a level-specific test targeting at C2 level while most of the participants in this study were at B2 level. In addition, the range of the participants' performances on Test Task A was the narrowest among all tasks. (The participants' performances on all tasks will be discussed in detail in Chapter Six). Therefore, the range of performance elicited on Test Task A in this study might have limited the difference which can be obtained in the use of cognitive processes between high-achieving and low-achieving participants.

5.3.2.2 Comparison between high and low achieving groups on Test Task B

The results of the comparison of the cognitive processes employed by the high- and low- achieving groups on Test Task B are presented in Table 5.20. Regarding Test Task B, the high-achieving participants reported employing nine of the cognitive processes more than the low-achieving group (see Table 5.20).

Table 5.20 Comparison between high-achieving and low-achieving (Test Task B)

	High-achieving (n=27)		Low-achieving (n=34)		Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
	Mean	Std Dev	Mean	Std Dev				
Conceptualisation								
Task representation and macro-planning	3.31	.39	3.04	.50	300.000	895.000	-2.331	.020
Revising macro plan	2.28	.92	2.94	.56	243.000	621.000	-3.199	.001
Meaning and discourse construction								
Careful global reading	3.20	.76	3.10	.66	406.500	1001.500	-.784	.433
Selecting relevant ideas	3.69	.31	3.17	.56	201.500	796.500	-3.840	.000
Connecting and generating	3.00	.80	3.00	.60	441.500	1036.500	-.257	.797
Organising								
Organising ideas in relation to input texts	3.44	.48	2.98	.46	220.000	815.000	-3.499	.000
Organising ideas in relation to own text	3.16	.55	3.10	.58	423.000	1018.000	-.546	.585
Low-level monitoring and revising								
Low-level editing while writing	3.28	.64	2.80	.74	279.500	874.500	-2.633	.008
Low-level editing after writing	2.84	1.14	2.01	.99	251.000	846.000	-3.097	.002
High-level monitoring and revising								
High-level editing while writing	3.37	.47	2.92	.63	249.500	844.500	-3.063	.002
High-level editing after writing	2.88	1.10	2.00	.89	218.500	813.500	-3.555	.000

The differences of seven processes (which include *task representation and macro-planning, selecting relevant ideas, organising ideas in relation to source texts, low-level editing while writing, low-level editing after writing, high-level editing while writing* and *high-level editing after* writing) between the two groups were significant ($p<0.05$). The results largely resemble the results of the same analysis on the real-life tasks.

Similar to the results reported on the real-life tasks (see Table 5.16), the low-achieving group reported that they revised macro plans more than the high-achieving group. The difference here was significant. The results seem to

suggest that the low-achieving test takers changed their minds more at the later stages of the writing process than the high-achieving test takers. As argued previously, this is probably because they were less capable of establishing a complete representation of the writing task (e.g. communicative purpose, intended readership, genre, content and style) and comprehensive macro plans before they started to write than the high-achieving test takers.

Valid cognitive parameters should not only distinguish the processes employed by high-achieving and low-achieving test takers on the test task, but also sufficiently resemble the cognitive processes which a test taker with high-, medium, or low- achievement would normally employ in non-test conditions. The following section reports a further comparison of cognitive processes elicited under test and real-life conditions in groups of high-, medium- and low-achievement.

5.3.3 Comparisons between the cognitive processes elicited under test conditions and the real-life conditions (in groups of high-, medium- and low-achievement)

Section 5.3.1 reported the comparison of the extent to which the participants, as a whole group, employed the cognitive processes under test conditions compared to real-life. The results showed that the participants, as a whole group, rated the extent they employed most cognitive processes significantly more in real-life conditions than under test conditions (both Test Task A and Test Task B). In order to further explore how test tasks elicited the processes from the participants when compared to the real-life tasks, the comparative analyses between the real-life and test conditions were performed again, taking the level of performance into consideration. The participants were divided into the following three groups:

(1) High-achieving group – the participants who were identified as high-achieving on both the test task (which is to be analysed) and the real-life tasks

(2) Low-achieving group – the participants who were identified as low-achieving on both the test task (which is to be analysed) and the real-life tasks

(3) Medium-achieving group – the participants who were identified as either higher-medium or lower-medium on both the test task (which is to be analysed) and the real-life tasks

Results of the inferential analysis using the Wilcoxon signed ranks tests between the processes employed on Test Task A and real-life data are reported according to the three levels of performance in Table 5.21, and results on Test Task B and real-life data in Table 5.22. The means and standard deviations of the rating (1=*definitely disagree*; 2=*mostly disagree*; 3=*mostly agree*; 4=*definitely agree)* per cognitive process are reported.

5.3.3.1 Comparison of the processes elicited on Test Task A and real-life tasks (in groups of high-, medium- and low-achievement)

As shown in Table 5.21 below, the high-achieving participants, those whose performances were ranked the top 25% on both Test Task A and real-life tasks, reported employing all cognitive processes similarly in both conditions. Any differences obtained were not significant. The low-achieving group (those whose performances were ranked the bottom 25% on both Test Task A and real-life tasks) reported employing most of the cognitive processes similarly in both conditions. However, they employed the low-level and high-level editing processes after writing significantly more on the real-life tasks than on Test Task A. The low-achieving participants reported an average rating of 1.61 (4=definitely agree; 3=mostly agree; 2=mostly disagree; 1=definitely disagree) for low-level editing after writing and 1.74 for high-level editing after writing on Test Task A. This implies that they did not employ these after writing editing processes on Test Task A, probably because the low-achieving test takers simply did not have the processing capacity to deal with these editing processes as less was available for this activity than in the case of the higher proficiency group.

Table 5.21 Comparisons between Test Task A and real-life cognitive processing data

	High-achieving group						Middle-achieving group						Low-achieving group					
	Test Task A (n=14)		Real-life tasks (n=14)				Test Task A (n=85)		Real-life tasks (n=85)				Test Task A (n=13)		Real-life tasks (n=13)			
	Mean	Std Dev	Mean	Std Dev	Z	Sig.	Mean	Std Dev	Mean	Std Dev	Z	Sig.	Mean	Std Dev	Mean	Std Dev	Z	Sig.
Conceptualisation																		
Task representation and macro-planning	3.21	.58	3.14	.47	-.566	.572	2.94	.64	3.15	.525	-2.005	.045	2.98	.56	2.98	.28	-.039	.969
Revising macro plan	2.39	.59	2.75	.80	-1.650	.099	2.3	.75	2.69	.90	-2.248	**.025**	2.53	.74	2.76	.64	-1.08	.280
Meaning and discourse construction																		
Careful global reading	2.42	1.17	2.89	1.07	-1.767	.077	2.93	.65	2.83	.65	-.947	.344	3.15	.51	2.81	.63	-1.218	.223
Selecting relevant ideas	3.66	.36	3.50	.53	-.654	.513	3.21	.82	3.17	.55	-1.031	.303	3.43	.58	3.05	.67	-1.620	.105
Connecting and generating	2.92	.74	3.28	.44	-1.446	.148	3.00	.68	3.15	.53	-1.261	.207	3.00	.47	2.96	.46	-.356	.722
Organising																		
Organising ideas in relation to source texts	3.10	.62	3.00	.46	-.594	.552	2.92	.60	2.89	.49	-.705	.481	3.06	.39	2.78	.46	-1.505	.132
Organising ideas in relation to own text	3.11	.62	3.41	.51	-1.259	.208	2.97	.65	3.10	.53	-1.732	.083	2.76	.52	2.86	.33	-.768	.443
Low-level monitoring and revising																		
Low-level editing while writing	3.10	.82	3.35	.55	-.595	.552	2.75	.72	3.08	.59	-3.262	**.001**	2.59	.62	2.88	.67	-.994	.320
Low-level editing after writing	2.37	1.01	2.66	.95	-.971	.331	2.37	1.01	2.81	.90	-3.140	**.002**	1.61	.92	2.63	1.06	-2.558	**.011**
High-level monitoring and revising																		
High-level editing while writing	3.15	.79	3.23	.51	-.118	.906	2.81	.65	3.11	.54	-3.904	**.000**	2.98	.53	2.93	.38	-.223	.823
High-level editing after writing	2.61	1.25	2.76	1.00	-.070	.940	2.36	1.02	2.76	.92	-2.69	**.010**	1.74	1.04	2.61	1.02	-2.572	**.010**

In comparison to the other two groups, the middle-achieving group showed greater discrepancy in how the participants employed the processes when comparing Test Task A and real-life tasks (See Table 5.21). There was no significant difference in the extent to which they reported employing the processes of the meaning and discourse construction and organising phases (i.e. *careful global reading*, *selecting relevant ideas*, *connecting and generating*, *organising ideas in relation to input texts*, and *organising ideas in relation to own text*) similarly between the two conditions. However, they reported employing the processes of the conceptualisation phase (i.e. *task representation and macro-planning* and *revising macro* plan) and of the monitoring and revising phases (i.e. *low-level editing during writing, low-level editing after writing, high-level editing during writing,* and *high-level editing after writing*) significantly more in the real-life conditions than under the test conditions (See Table 5.21).

In short, the results provided empirical evidence supporting the cognitive validity of this reading-into-writing task type (essay task with multiple verbal inputs). There was no significant difference in the extent to which the high-achieving participants reported employing all processes on Test Task A and the real-life tasks. The high-achieving group seemed to be able to utilise the processes under both conditions. This suggests that provided that the test takers were proficient in academic writing, Test Task A was able to elicit the same processes from test takers to a similar extent as they employed them on the real-life tasks. Apart from the processes of after writing low-level and high-level editing, the low-achieving participants reported employing all cognitive processes on Test Task A in a similar manner as they employed them in the real-life academic contexts. The middle group reported employing five of the processes on Test Task A in a similar manner as they did on the real-life tasks. The middle group seemed to show greater discrepancy in how they employed the processes under the test and real-life conditions. This is probably because they were in transitional state of developing their academic writing skills and perhaps more affected by the performance conditions, e.g. stricter time allowance.

In addition, it is encouraging that all three groups of participants employed all processes of the meaning and discourse construction phase and the organising phase *(i.e. careful global reading, selective reading, connecting and generating, organising ideas in relation to input texts,* and *organising ideas in relation to own text*) on Test Task A and in real-life conditions in a statistically similar manner. This evidence supports the literature that the reading-into-writing task type is a valid task type to test the process of discourse synthesis (Spivey, 1984, 2001; Spivey & King, 1989), which is a core set of academic writing skills. These processes are also believed to play an important role in critical academic literacy (Flower et al, 1990).

5.3.3.2 Comparison of the processes elicited on Test Task B and real-life tasks (in groups of high-, medium- and low-achievement)

As presented in Table 5.22, there was no significant difference in the extent to which the high-achieving participants (those whose performances were ranked in the top 25% on both Test Task B and real-life tasks) reported employing eight of the eleven cognitive processes on Test Task B and the real-life tasks. Nevertheless, the high-achieving participants employed the processes of revising macro plan and low-level editing while writing significantly more on the real-life tasks than on Test Task B whereas they reported organising ideas in relation to input texts significantly more on Test Task B than on real-life tasks.

Table 5.22 Comparisons between Test Task B and real-life cognitive processing data

	High-achieving group						Middle group						Low-achieving group					
	Test Task B n=11)		Real-life tasks n=11)		Z	Sig.	Test Task B (n=65)		Real-life tasks (n=63)		Z	Sig.	Test Task B (n=13)		Real-life tasks (n=13)		Z	Sig.
	Mean	Std Dev	Mean	Std Dev			Mean	Std Dev	Mean	Std Dev			Mean	Std Dev	Mean	Std Dev		
Conceptualisation																		
Task representation and macro-planning	3.25	.31	2.95	.45	-1.782	.075	3.13	.54	3.15	.55	-.642	.521	3.06	.518	3.11	.35	-.035	.972
Revising macro plan	1.63	.60	2.5	1.01	-2.102	**.036**	2.85	.90	2.92	.49	-.259	.796	2.62	.77	2.73	.66	-.815	.415
Meaning and discourse construction																		
Careful global reading	3.04	.61	2.81	1.12	-.850	.396	3.21	.59	2.88	.65	-2.769	**.006**	3.03	.47	2.92	.81	-.171	.864
Selecting relevant ideas	3.66	.33	3.57	.51	-.680	.497	3.31	.52	3.13	.62	-.959	.338	3.12	.64	3.33	.43	-.738	.461
Connecting and generating	2.78	.95	3.02	.42	-.045	.964	2.90	.58	3.20	.57	-3.539	**.000**	3.10	.516	3.12	.51	-.275	.783
Organising																		
Organising ideas in relation to input texts	3.54	.50	2.83	.44	-2.762	**.006**	3.08	.54	2.86	.56	-1.947	.052	3.06	.45	2.92	.36	-1.434	.152
Organising ideas in relation to own text	2.86	.50	3.00	.68	-.106	.916	3.06	.69	3.08	.54	-.485	.628	3.00	.40	3.11	.59	-.677	.498
Low-level monitoring and revising																		
Low-level editing while writing	2.95	.75	3.25	.79	-2.081	**.037**	2.87	.78	2.97	.67	-.872	.383	2.82	.56	3.17	.49	-1.740	.082
Low-level editing after writing	2.52	1.15	2.34	1.06	-.340	.734	2.40	1.14	2.86	.90	-2.996	**.003**	1.86	.927	2.71	.92	-2.336	**.019**
High-level monitoring and revising																		
High-level editing while writing	3.21	.54	3.06	.57	-1.137	.256	2.92	.70	3.10	.61	-1.683	.092	3.07	.45	3.21	.515	-.666	.505
High-level editing after writing	2.62	1.12	2.65	1.12	-.422	.673	2.43	1.14	2.85	.94	-2.695	**.007**	1.97	.95	2.65	.88	-1.887	.059

It is interesting to explore why the high-achieving participants did not revise their macro plans as they did in the real-life context. One participant in the high-achieving group commented on the questionnaire that 'I planned the writing following closely the instructions' (P37). The task prompt of Test Task B lists the key points need to be covered in the text (See Appendix 3.1.4). This might have offered too much help to the high-achieving participants and therefore they did not seem to revise their macro plans as much as they did on the real-life tasks. This should not be a big concern to the cognitive validity because the discussion below will show that the other two groups reported the extent to which they revised macro-plans similarity on Test Task B and in real-life conditions. Another interesting finding is that the high-achieving participants edited at low level while writing on Test Task B significantly less than they did in the real-life conditions. Low-level revisions primarily concern grammatical accuracy.

For the low-achieving group (those whose performances were ranked the bottom 25% on both Test Task B and real-life tasks), the participants reported employing most cognitive processes similarly between the two conditions. Except for the processes of *low-level editing after writing*, all differences reported were insignificant.

The middle-achieving participants reported employing seven of the cognitive processes similarly between Test Task B and the real-life tasks. They employed the processes of *connecting and generating, low-level editing after writing* and *high-level editing after writing* significantly more in the real-life conditions than on Test Task B. On the other hand, they employed the processes of *global reading* significantly more on Test Task B than in the real-life conditions. As discussed earlier, the seemingly over-eliciting of global reading processes could be due to the fact that length of the input texts on Test Task B were comparatively short. While these participants might have over-employed careful global reading processes, the results showed that they employed selective reading processes on Test Task B similarly as they did in the real-life conditions. The task requires the use of multiple verbal and non-verbal inputs within a limited time. To solve the

issue, further evidence is needed to determine the most appropriate length of the verbal texts on Test Task B.

In short, the results reveal that Test Task B (essay task with multiple verbal and non-verbal inputs) elicited use of the majority of the cognitive processes by the participants in a similar manner as the processes were employed in the real-life academic contexts. The results also reveal that a slightly increased difference in use of cognitive processes on Test Task B compared to real-life tasks was reported by the middle-achieving participants. As discussed earlier, the middle group was perhaps more affected by the different performance conditions. Under a stricter time allowance performance condition, they employed the processes of *connecting and generating*, *low-level editing after writing* and *high-level editing after writing* on Test Task B significantly less than the extent to which they did under the real-life conditions which provided much longer time-allowance.

5.3.3.3 Summary of the cognitive processes employed by the proficiency groups between real-life and test conditions

The results comparing the extent to which the eleven cognitive processes were employed by the high-, middle- and low-achieving groups between real-life and test conditions have been discussed.

It is encouraging that both reading-into-writing test tasks were able to elicit from high-achieving and low-achieving participants most of the cognitive processes to a similar extent as participants employed the processes on the real-life tasks. The middle group showed greater discrepancy in how they employed the processes under the test and real-life conditions. They tended to employ some processes more in the real-life conditions than the test conditions. They employed the processes of the conceptualisation, low-level monitoring and revising, and high-level monitoring and revising phases significantly less on Test Task A than on real-life tasks. They employed the processes of *connecting and generating*, *low-level editing after writing* and *high-level editing after writing* significantly less on Test Task B than on real-life tasks. When compared to the high-achieving group, the middle-achieving group might not be able to employ all processes with full automaticity as they were at the transitional stage of developing their academic

writing ability. Due to limited cognitive capacity, they may need more time to complete the processes. Many participants commented on the questionnaire that they did not have sufficient time for different processes, such as 'read the passages and understand better', 'proofread the mistakes', 'improve the writing', 'think more carefully' etc. It would be essential in future research to investigate why, under test conditions, the middle-achieving participants employed some of the real life processes but not others on these two types of reading-into-writing tasks.

This next sub-section will continue to examine the number of distinct cognitive processes involved in the five cognitive phases and the underlying structure of these cognitive processes elicited under the test conditions.

5.3.4 Factor analyses of the cognitive processes elicited by the test tasks

The last step of investigating the cognitive validity of the reading-into-writing test was to examine the underlying structure of the cognitive processes elicited by the two reading-into-writing test task types. Exploratory factor analysis was used to examine the underlying structure of the cognitive processes elicited at each hypothesised writing phases in real-life conditions (Section 5.2.2). The results showed that each writing phase involved two or more distinct yet correlated underlying cognitive processes. Here, exploratory factor analysis was used again to examine the underlying structure of the five hypothesised writing phases elicited by two test tasks (Test Task A and Test Task B). It was thought that exploratory factor analysis would be more appropriate for the analysis here rather than confirmatory factor analysis. Confirmatory factor analysis is usually used to test whether a particular data set fit a measurement model, i.e. a model of cognitive processes in this case. Since there is apparently insufficient empirical evidence of the cognitive processes elicited by reading-into-writing tests in the literature, exploratory factor analysis seems most appropriate to examine the underlying structure of the cognitive processes elicited by the two reading-into-writing task types. Future studies can then use confirmatory factor analysis to evaluate the results of this study.

Kaier-Meyer-Olkin (KMO) test and Bartlett's test of sphericity were performed on the data of each of the five academic writing phases to show the data appropriateness for factor analysis. The results showed that the processing data on both Test Task A and Test Task B (See Appendix 5.4) were appropriate for factor analysis. Similar to the real-life data, the data of the cognitive processes on the test tasks were not normally distributed (K-S test: sig<0.01). The principal axis factor method with the promax rotation procedure was performed. Eigenvalues of the factors (i.e. factors below the value of 1) and scree plot (i.e. the factors on the left of the point of the curve where there is a sudden change of steepness are significant) were consulted for initial factor solutions. The possible factor structures were evaluated to determine the final underlying structure of cognitive processes at each writing phase elicited by the two test tasks.

5.3.4.1 The underlying structure of the conceptualisation phase (Test Task A and Test Task B)

Conceptualisation is the first phase of the writing process where writers conceptualise the writing task and set macro plans. As presented in Section 5.2.2.1, the conceptualisation phase elicited by the real-life tasks in this study involved the processes of *task representation and macro-planning* (which was measured by six questionnaire items) and *revising macro plan* (which was measured by two items).

On Test Task A data, the initial factor extractions yielded two factors with eigenvalues greater than 1.0 and the scree plot also suggested a two-factor solution (See Table 5.23). The two-factor suggestion was accepted. The pattern matrix and the interfactor correlations for the conceptualisation construct on Test Task A are presented in Table 5.24. The percentage in brackets indicates the extent to which each factor accounts for the variance.

Table 5.23 Eigenvalues and scree plot for the conceptualisation phase (Test Task A)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.720	33.996	33.996
2	1.527	19.090	53.086
3	.970	12.123	65.210
4	.757	9.457	74.667
5	.658	8.228	82.895
6	.557	6.968	89.863
7	.460	5.751	95.614
8	.351	4.386	100.000

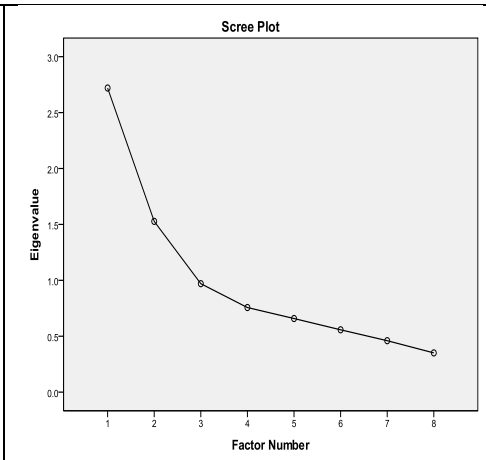


Table 5.24 Pattern and interfactor correlations matrix for the conceptualisation phase (Test Task A)

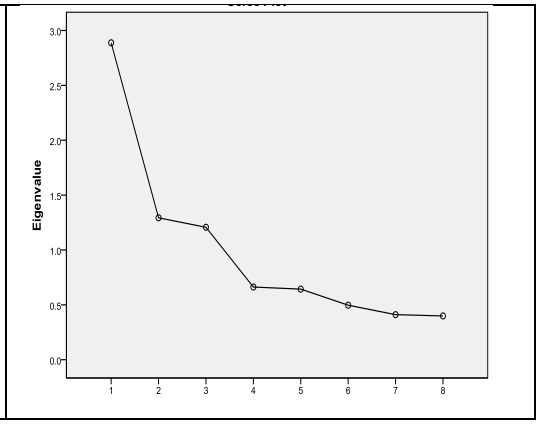
		F1 Task representation and macro-planning (33.98%)	F2 Revising macro plan (19.04%)
1.2	I read the whole task prompt (i.e. instructions) carefully.	.805	
1.5	I thought about the purpose of the task.	.663	
1.4	I understood the instructions for this writing task very well.	.584	
1.3	I thought of how my text would suit the expectations of the intended reader.	.510	
2.6	I went back to read the task prompt again while I was reading the source texts.	.465	
4.4	I re-read the task prompt while I was writing.	.445	
2.13	I changed my writing plan while reading the source texts.		.799
4.6	I changed my writing plan (e.g. structure, content etc) while I was writing.		.584
Interfactor correlations			
Factor 1 (Task representation and macro-planning)		1.000	
Factor 2 (Revising macro plan)		.042	1.000

The two-factor solution presented in Table 5.24 resembles the two factors generated by the real-life data. In comparison to the conceptualisation phase elicited by the real-life tasks, the same six items loaded on Factor 1 (which was named *task representation and macro-planning*) and the same two items loaded on Factor 2 (which was named *revising macro plan*).

On Test Task B data, the initial factor extraction for the processes at the conceptualisation phase yielded three factors with eigenvalues greater than 1.0. The scree plot suggested two- or three- factor solutions (See Table 5.25). The rotated two- and three- factor solutions were compared. The two-solution (provided in Appendix 5.7 Table 1) was rejected because one item (i.e. Item 2.13) did not load on either of the factors. Besides, Factor 2 was difficult to interpret.

Table 5.25 Eigenvalues and scree plot for the conceptualisation phase (Test Task B)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.888	36.101	36.101
2	1.293	16.164	52.265
3	1.207	15.086	67.350
4	.662	8.278	75.629
5	.643	8.041	83.670
6	.497	6.210	89.880
7	.411	5.134	95.014
8	.399	4.986	100.000



The three-factor solution was accepted because it provides additional information about the conceptualisation phase elicited by Test Task B (See Table 5.26). The four items loaded on Factor 1 were mainly related to setting macro plans. Factor 1 was named *macro-planning.* Compared to the real-life data, Factor 2 involved the same two items of revising macro plan employed at later stages of the writing process. It was named *revising macro plan.* However, the conceptualisation phase elicited by Test Task B showed an additional Factor 3, which involved two items of reading the task prompt again at later stages of the writing process. It was named *rereading task prompt.* As shown in Table 5.26, the processes of *macro-planning* correlated with the processes of *revising macro plan* at the level of 0.462.

However, the processes of *macro-planning* barely correlated with the processes of rereading task prompt.

Table 5.26 Pattern and interfactor correlations matrix for the conceptualisation phase (Test Task B)

		F1 Macro-planning (31.98%)	F2 Revising macro plan (16.16%)	F3 Rereading task prompt (15.09%)
1.2	I thought of what I might need to write to make my text relevant and adequate to the task.	.710		
1.5	I thought about the purpose of the task.	.705		
1.4	I understood the instructions for this writing task very well.	.680		
1.3	I thought of how my text would suit the expectations of the intended reader.	.560		
4.6	I changed my writing plan (e.g. structure, content etc) while I was writing.		.584	
2.13	I changed my writing plan while reading the source texts.		.583	
4.4	I re-read the task prompt while I was writing.			.793
2.6	I went back to read the task prompt again while I was reading the source texts.			.467
	Interfactor correlations			
	Factor 1 (Macro-planning)	1.000		
	Factor 2 (Revising macro plan)	.462	1.000	
	Factor 3 (Rereading task prompt)	.152	.132	1.000

The conceptualisation phase elicited by Test Task B showed an additional Factor 3 (rereading task prompt). The processes of reading task prompt again were not identified as a distinct factor in real-life conditions. Based on the qualitative data collected on the questionnaire, the most commonly mentioned reasons why the participants reported reading their task prompt again at later stages of the writing production on Test Task B were as follows:

- I wanted to check whether I am following the instructions (P24)
- I was checking the key points I have to finish (P154)
- I checked the marking criteria (P64)

According to Scardamalia & Bereiter (1987), writers who use a knowledge transforming approach are highly aware of different rhetorical problems (e.g. what to write, how to write, to whom to write) during the entire writing process. They would constantly evaluate their progress against available resources (which can be obtained from long-term memory and/or external reading materials) and constraints (e.g. remaining time). The results from the contextual analysis of task setting reported in Chapter Four showed that Test Task B was rated the highest for the *clarity of purpose*, *clarity of intended reader* and *clarity of the knowledge criteria* in a scale of 1 (unclear) to 5 (clear), when compared to the real-life tasks and Test Task A. It seems that a clear presentation of these contextual features in the task prompt would encourage test takers to monitor their progress by checking the task prompt from time to time during the writing process.

5.3.4.2 The underlying structure of the meaning and discourse construction construct (Test Task A and Test Task B)

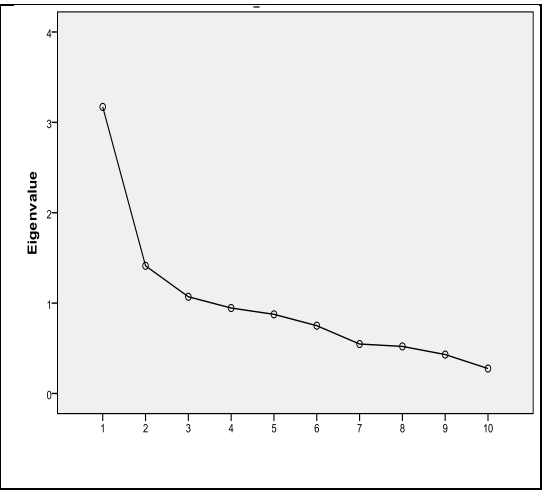
Meaning and discourse construction is a higher-level phase where students contextualise meaning and establish discourse representations from different sources. As presented in Section 5.2.2.2, the meaning and discourse construction phase elicited on the real-life tasks involved three distinct underlying cognitive processes. The first process was careful reading at global level (*careful global reading)*. The second one was to select ideas which are relevant to the writing task (*selecting relevant ideas).* The last one was to generate links between ideas or new meaning by connecting ideas/discourse features provided in the source texts (*connect and generate).*

On Test Task A, the initial factor extraction for the meaning and discourse construction construct produced three factors with eigenvalues greater than 1.0. The scree plot suggested two- or three- factor solutions (See Table 5.27). The rotated two- and three- factor solutions were compared. The three-factor solution

(provided in Appendix 5.6 Table 1) was rejected because factor 3 includes one item only. Besides, Item 4.2 did not load on any factors at the level of 0.3 or above.

Table 5.27 Eigenvalues and scree plot for the discourse and meaning construction phase (Test Task A)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.171	31.712	31.712
2	1.413	14.128	45.840
3	1.070	10.702	56.542
4	.945	9.455	65.996
5	.876	8.757	74.753
6	.750	7.502	82.256
7	.547	5.472	87.728
8	.521	5.207	92.935
9	.431	4.307	97.242
10	.276	2.758	100.000



The two-factor solution was accepted (See Table 5.28). Compared to the meaning and discourse construction phase elicited by real-life tasks, Factor 1 (*selecting relevant ideas*) included the same three items of search reading for ideas which are relevant to the task. However, Factor 2 on Test Task A included four items, which were loaded as separate factors in real-life conditions. Factor 2 on Test Task A was named *connecting and generating with careful global reading.*

Table 5.28 Pattern matrix and interfactor correlations for the discourse and meaning construction phase (Test Task A)

		F1 Selecting relevant ideas (33.38%)	F2 Connecting and generating with careful global reading (14.09%)
2.5	I read some relevant part(s) of the texts carefully.	.892	
2.4	I searched quickly for part(s) of the texts which might help complete the task.	.807	
2.7	I took notes on or underlined the important ideas in the source texts.	.607	
2.9	I linked the important ideas in the source texts to what I know already.		.649
4.3	I made further connections across the source texts while I was writing.		.531
2.1	I read through the whole of each source text slowly and carefully.		.489
4.2	I developed new ideas while I was writing.		.476
2.2	I read the whole of each source text more than once.		.447
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.		.402
Interfactor correlations matrix			
Factor 1 (Selecting relevant ideas)		1.000	
Factor 2 (Connecting and generating with careful global reading)		.175	1.000

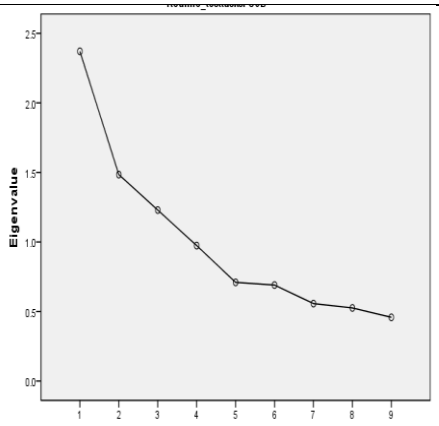
It is interesting to explore why the processes of careful global reading employed on Test Task A did not represent a stand-alone factor. Descriptive results reported in Section 5.3.1.1 showed that the mean rating of the careful reading was 2.86 (on a scale of 1 to 4) whereas the use of selecting relevant ideas was 3.22 on Test Task A. When compared to real-life conditions, Test Task A imposes a tighter time constraint on test takers. The contextual analysis reported in Chapter four may also offer an explanation why the participants employed less global careful reading on Test Task A than real-life tasks. When compared to real-life tasks, according to the judges, Test Task A was placed more towards to the lower end of the cognitive demands of transforming content from source texts to writer's own

text. In addition, the content of Test Task A source texts was more concrete and the textual organisation of Test Task A source texts was clearer than real-life source texts.

On Test Task B, the initial factor extraction for the meaning and discourse construction phase produced three factors with eigenvalues greater than 1.0. The scree plot suggested one- or four-factor solutions (see Table 5.29). The rotated three- and four-factor solutions were compared. The four-factor solution (provided in Appendix 5.7 Table 2) reflected interesting results of how reading processes interact with the processes of connecting and generating. The solution was, however, rejected because Factor 4 did not reach an eigenvalue of 1.0 or above, which indicates that the last factor is non significant.

Table 5.29 Eigenvalues and scree plot for the discourse and meaning construction phase (Test Task B)

	Initial Eigenvalues		
Factor	Total	% of Variance	Cumulative %
1	2.371	26.341	26.341
2	1.485	16.495	42.836
3	1.230	13.661	56.497
4	.974	10.823	67.320
5	.710	7.886	75.206
6	.690	7.671	82.877
7	.557	6.190	89.067
8	.526	5.844	94.911
9	.458	5.089	100.000



The three-factor solution was taken. Based on the initial three-factor solution, Item 4.3 did not load on any factors at the level of 0.3 or above (see Table 5.30). It was therefore dropped from the analysis.

Table 5.30 Pattern matrix for the meaning and discourse construction phase (Test Task B): initial three-factor solution

Items		F1	F2	F3
2.4	I searched quickly for part(s) of the texts which might help complete the task.	.766		
2.7	I took notes on or underlined the important ideas in the source texts.	.585		
2.5	I read some relevant part(s) of the texts carefully.	.501		
4.3	I made further connections across the source texts while I was writing.			
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.		.713	
4.2	I developed new ideas while I was writing.		.597	
2.9	I linked the important ideas in the source texts to what I know already.		.448	
2.2	I read the whole of each source text more than once.			.723
2.1	I read through the whole of each source text slowly and carefully.			.469

After removal, the rotated three-factor solution is presented in Table 5.31. In common to the meaning and discourse construction phase elicited by real-life tasks, Factor 1 (*selecting relevant ideas)* involved the same three items of search reading, and Factor 3 (*global careful reading)* involved the same two items as identified by the real-life data. Factor 2 (*connecting and generating*) involved three instead of four items.

Table 5.31 Pattern and interfactor correlations matrix for the meaning and discourse construction phase (Test Task B)

		F1 Selecting relevant ideas (28.20%)	F2 Connect and generate (18.45%)	F3 Global careful reading (15.07%)
2.4	I searched quickly for part(s) of the texts which might help complete the task.	.723		
2.5	I read some relevant part(s) of the texts carefully.	.564		
2.7	I took notes on or underlined the important ideas in the source texts.	.555		
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.		.758	
4.2	I developed new ideas while I was writing.		.546	
2.9	I linked the important ideas in the source texts to what I know already.		.499	
2.2	I read the whole of each source text more than once.			.678
2.1	I read through the whole of each source text slowly and carefully.			.504
Interfactor correlations matrix				
	Factor 1 (Selecting relevant ideas)	1.000		
	Factor 2 (Connecting and generating)	.284	1.000	
	Factor 3 (Global careful reading)	.273	.304	1.000

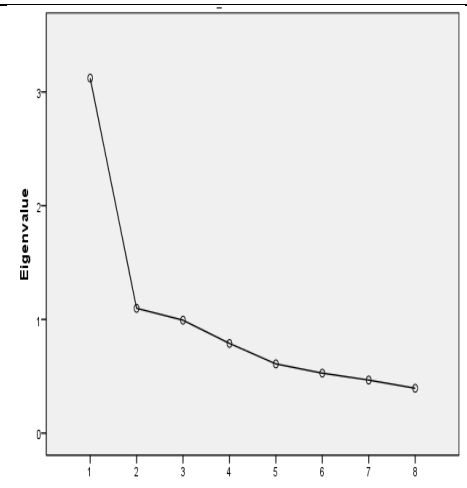
5.3.4.3 The underlying structure of the organising phase (Test Task A and Test Task B)

The use of organising processes is an important academic writing construct which provides evidence of distinguishing different levels of writing expertise. Scardamalia & Bereiter (1987) argued that immature writers tend to translate the ideas from their long-term memory to their text in the same order as the idea retrieval. In contrast, mature writers would explicitly organise the ideas which have been retrieved from long-term memory according to their macro writing plans. Section 5.2.2.3 showed the organising phase elicited by the real-life tasks involved two distinct cognitive processes. The first one (*organising ideas in relation to input texts)* was related to the processes of organising the textual and intertextual representations of the input texts. The second one (*organising ideas in relation to new text*) was related to the processes of organising the ideas to be put in the writer's own text.

On Test Task A, the initial factor extraction for the organisation construct produced three factors with eigenvalues greater than 1.0. The scree plot, however, suggested one or two-factor solutions (See Table 5.32). The one-factor solution was not investigated. Rotated solutions with two or three factors were compared.

Table 5.32 Eigenvalues and scree plot for the organising phase (Test Task A)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.040	38.001	38.001
2	1.047	13.092	51.093
3	1.014	12.672	63.765
4	.956	11.948	75.713
5	.562	7.021	82.733
6	.533	6.659	89.392
7	.452	5.649	95.041
8	.397	4.959	100.000



The three-factor solution (provided in Appendix 5.6 Table 2) was rejected because Factor 2 and Factor 3 included one primary item only. Besides, Item 3.1 loaded on two factors. The two-factor solution was accepted. The initial results showed that Item 3.2 loaded on both factors while Item 3.3 loaded on neither at a level of 0.3 or above (See Table 5.33). They were dropped from the analysis. The process of reordering and recombining ideas (Item 3.2) loaded on both factors. This implies that while the participants reordered and recombined ideas, they might have focused on both organising their representation of the input texts as well as their own text. Further evidence is needed to confirm this. Besides, the process of removing ideas (Item 3.3) did not load on either of the factors at the level of 0.3 or above. This suggests that the participants did not employ the process of removing ideas from their plans in the same way as they employed other organising processes.

Table 5.33 Pattern matrix for the organising phase (Test Task A): initial two-factor solution

Items		F1	F2
2.11	I worked out how the main ideas across the source texts	.858	
2.10	I worked out how the main ideas in each source text relate to each other.	.614	
2.8	I prioritised important ideas in the source texts in my mind.	.587	
3.2	I recombined or reordered the ideas to fit the structure of my essay.	.494	.309
2.3	I used my knowledge of how texts like these are organised to find parts to focus on.	.493	
3.3	I removed some ideas I planned to write.		
4.1	While I was writing I sometimes paused to organise my ideas.		.633
3.1	I organised the ideas for my text before starting to write.		.426

After the removal, the two-factor solution was extracted again (See Table 5.34). Similar to the real-life data, the underlying structure of the organising processes elicited on Test Task A yields two distinct cognitive processes. Factor 1 (*organising ideas in relation to input texts*) consisted of the same four items of

organising the textual or intertexual representations of the input texts, as identified by the real-life data. However, Factor 2 (*organising ideas in relation to own texts*) elicited on Test Task A included less items than the same factor elicited by the real-life tasks. The two factors correlated at a level of 0.61 (See Table 5.34).

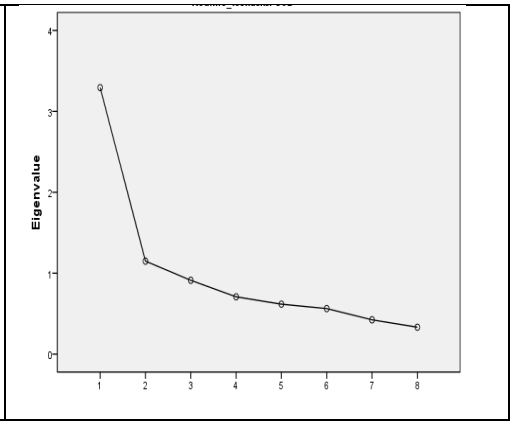
Table 5.34 Pattern and interfactor correlations matrix for the organising phase (Test Task A)

		F1 Organising ideas in relation to input texts (41.70%)	F2 Organising ideas in relation to own text (17.65%)
2.11	I worked out how the main ideas across the source texts	.842	
2.8	I prioritised important ideas in the source texts in my mind.	.604	
2.10	I worked out how the main ideas in each source text relate to each other.	.545	
2.3	I used my knowledge of how texts like these are organised to find parts to focus on.	.447	
3.1	I organised the ideas for my text before starting to write.		.637
4.1	While I was writing I sometimes paused to organise my ideas.		.505
Interfactor correlations matrix			
Factor 1 (Organising ideas in relation to input texts)		1.000	
Factor 2 (Organising ideas in relation to own text)		.612	1.000

On Test Task B, the initial factor extraction for the organisation phase produced two factors with eigenvalues greater than 1.0. The scree plot, however, suggested a one-factor solution (see Table 5.35). The one-factor solution was rejected because the two-factor solution provided more information about the underlying structure of the organisation construct.

Table 5.35 Eigenvalues and scree plot for the organising phase (Test Task B)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.293	41.168	41.168
2	1.149	14.366	55.534
3	.912	11.401	66.935
4	.708	8.855	75.790
5	.617	7.717	83.507
6	.563	7.038	90.545
7	.424	5.301	95.846
8	.332	4.154	100.000



The initial solution showed that Item 3.1 loaded on both factors (See Table 5.36). It was dropped from the analyses. The item reads as: I organised the ideas for my text before starting to write.

Table 5.36 Pattern matrix for the organising phase (Test Task B): initial two-factor solution

Items		F1	F2
2.10	I worked out how the main ideas in each source text relate to each other.	.916	
2.11	I worked out how the main ideas across the source texts	.688	
2.8	I prioritised important ideas in the source texts in my mind.	.567	
3.1	I organized the ideas for my text before starting to write.	.466	.323
2.3	I used my knowledge of how texts like these are organised to find parts to focus on.	.408	
4.1	While I was writing I sometimes paused to organise my ideas.		.636
3.2	I recombined or reordered the ideas to fit the structure of my text.		.622
3.3	I removed some ideas I planned to write.		.430

After removal, the rotated two-factor solution was extracted again (See Table 5.37). The underlying structure of the organisation phase elicited on Test Task B largely resembles the two distinct processes identified in the real-life data. Factor 1 (*organising ideas in relation to input texts*) consisted of the same four items. However, Factor 2 (*organising ideas in relation to own text*) included three instead of four items. The two factors correlated at the level of 0.53 (See Table 5.37).

Table 5.37 Pattern and interfactor correlations matrix for the organising phase (Test Task B)

		F1 Organising ideas in relation to input texts (40%)	F2 Organising ideas in relation to own text (16.41%)
2.10	I worked out how the main ideas in each source text relate to each other.	.874	
2.11	I worked out how the main ideas across the source texts	.649	
2.8	I prioritised important ideas in the source texts in my mind.	.590	
2.3	I used my knowledge of how texts like these are organised to find parts to focus on.	.431	
3.2	I recombined or reordered the ideas to fit the structure of my text.		.671
4.1	While I was writing I sometimes paused to organise my ideas.		.484
3.3	I removed some ideas I planned to write.		.460
Interfactor correlations matrix			
Factor 1 (Organising ideas in relation to input texts)		1.000	
Factor 2 (Organising ideas in relation to own texts)		.534	1.000

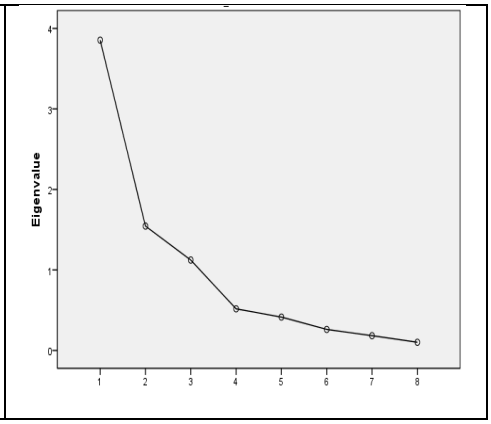
5.3.4.4 The underlying structure of the low-level monitoring and revising phase (Test Task A and Test Task B)

Low-level monitoring and revising is a phase where the writer monitors the quality of their own text (mainly in terms of grammatically accuracy) and revises the unsatisfactory parts of the text (Field, 2004). Section 5.2.2.4 showed that the low-level monitoring and revising phase elicited by the real-life tasks consisted of two distinct processes. The first one was low-level editing employed during writing, and the other one was low-level editing employed after the first draft has been produced.

On Test Task A, the initial factor extraction for the low-level monitoring and revising processes produced three factors with eigenvalues greater than 1.0. The scree plot suggested one-, two- or three- factor solutions (See Table 5.38).

Table 5.38 Eigenvalues and scree plot for the low-level revising phase (Test Task A)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.855	48.187	48.187
2	1.545	19.308	67.495
3	1.123	14.032	81.527
4	.517	6.460	87.987
5	.415	5.181	93.168
6	.261	3.267	96.436
7	.183	2.291	98.727
8	.102	1.273	100.000



The one-factor solution was not analysed. Rotated solutions with two- and three-factor solutions were compared. The first factors extracted by both solutions were the same. According to the three-factor solution (provided in Appendix 5.6 Table 3), Factor 2 focused on the linguistic accuracy whereas Factor 3 focused on the appropriate use of source texts. The two-factor solution (See Table 5.39) was taken because it resembled more closely the underlying structure extracted from the real-life data than the three-factor solution. As per the real-life data, the low-level monitoring and revising phase elicited by Test Task A involved two distinctive cognitive processes: *low-level editing after writing* and *low-level editing during writing.* The items loaded on each factor were the same as identified by the real-life tasks. The two processes correlated moderately at a level of 0.45.

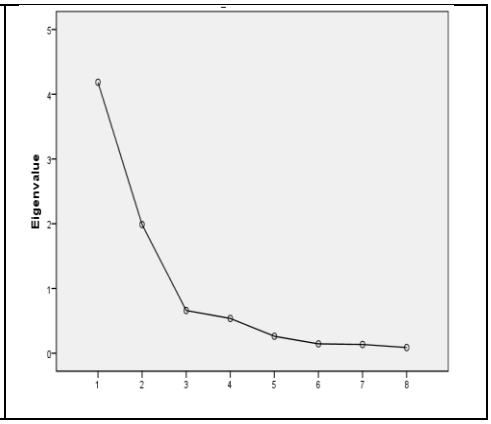
Table 5.39 Pattern and interfactor correlations matrix for the low-level revising phase (Test Task A)

		F1 Low-level editing after writing (48.18%)	F2 Low-level editing while writing (19.31%)
5.12	After I had finished the first draft, I checked that the quotations were properly made.	.857	
5.15	After I had finished the first draft, I checked the accuracy and range of the sentence structures.	.832	
5.16	After I had finished the first draft, I checked the appropriateness and range of vocabulary.	.809	
5.13	After I had finished the first draft, I checked that I had put the ideas of the source texts into my own words.	.801	
4.15	I checked the accuracy and range of the sentence structures while I was writing.		.930
4.16	I checked the appropriateness and range of vocabulary while I was writing.		.724
4.12	I checked that the quotations were properly made while I was writing.		.501
4.13	I checked that I had put the ideas of the source texts into my own words while I was writing.		.421
Interfactor correlations matrix			
	Factor 2 (Low-level editing after writing)	1.000	
	Factor 1 (Low-level editing while writing)	.450	1.000

On Test Task B, the initial factor extraction for the low-level monitoring and revising construct produced two factors with eigenvalues greater than 1.0. The scree plot also suggested a two-factor solution (see Table 5.40).

Table 5.40 Eigenvalues and scree plot for the low-level revising phase (Test Task B)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.185	52.310	52.310
2	1.988	24.846	77.156
3	.660	8.248	85.404
4	.537	6.716	92.120
5	.263	3.282	95.402
6	.145	1.808	97.210
7	.135	1.690	98.901
8	.088	1.099	100.000



The low-level monitoring and revising phase elicited by Test Task B yielded the same underlying factors: *low-level editing after writing* and *low-level editing during writing*, as extracted by the real-life tasks. The two factors correlated at a level of 0.371 (See Table 5.41).

Table 5.41 Pattern and interfactor correlations matrix for the low-level monitoring and revising phase (Test Task B)

		F1 Low-level editing after writing (52.31%)	F2 Low-level editing while writing (24.85%)
5.16	After I had finished the first draft, I checked the appropriateness and range of vocabulary.	.951	
5.15	After I had finished the first draft, I checked the accuracy and range of the sentence structures.	.923	
5.13	After I had finished the first draft, I checked that I had put the ideas of the source texts into my own words.	.893	
5.12	After I had finished the first draft, I checked that the quotations were properly made.	.807	
4.15	I checked the accuracy and range of the sentence structures while I was writing.		.902
4.16	I checked the appropriateness and range of vocabulary while I was writing.		.750
4.12	I checked that the quotations were properly made while I was writing.		.651
4.13	I checked that I had put the ideas of the source texts into my own words while I was writing.		.622
Interfactor correlations matrix			
	Factor 2 (Low-level editing after writing)	1.000	
	Factor 1 (Low-level editing while writing)	.371	1.000

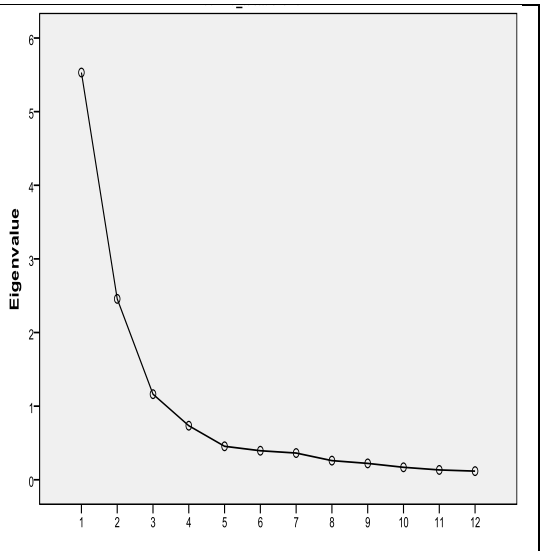
5.3.4.5 Underlying structure of the high-level monitoring and revising construct (Test Task A and Test Task B)

Similar to the low-level monitoring and revising phase, Section 5.2.2.5 showed that the high-level monitoring and revising phase elicited by the real-life tasks consisted of two distinct processes. Factor 1 included the processes of high-level editing employed during writing, and Factor 2 included the processes of high-level editing employed after the first draft has been produced.

On Test Task A, the initial factor extraction for the high-level monitoring and revising phase produced three factors with eigenvalues greater than 1.0. The scree plot suggested a two-factor or four-factor solutions (See Table 5.42). Rotated two-, three- and four-factor solutions were compared. The three-factor (provided in Appendix 5.6 Table 4) was rejected because Factor 3 included only one primary item. The four-factor solution (provided in Appendix 5.6 Table 5) was rejected because Factor 3 included only one item and Factor 4 had no primary factor (i.e. all its items loaded more heavily on another factor).

Table 5.42 Eigenvalues and scree plot for the high-level monitoring and revising phase (Test Task A)

	Initial Eigenvalues		
Factor	Total	% of Variance	Cumulative %
1	5.531	46.094	46.094
2	2.457	20.477	66.571
3	1.163	9.688	76.258
4	.734	6.119	82.377
5	.455	3.789	86.166
6	.394	3.286	89.452
7	.363	3.025	92.477
8	.260	2.167	94.644
9	.222	1.853	96.497
10	.169	1.412	97.909
11	.134	1.113	99.022
12	.117	.978	100.000



The two-factor solution was accepted. The initial two-factor solution (See Table 5.43) showed that Item 4.14 did not load on any factor at the level of 0.3 or above. The item was dropped from the analysis.

Table 5.43 Pattern matrix for the high-level monitoring and revising phase (Test Task A): initial two -factor solution

Items		F1	F2
5.7	After I had finished the first draft, I checked that the content was relevant.	.914	
5.10	After I had finished the first draft, I checked that I included all appropriate main ideas from all the source texts.	.878	
5.8	After I had finished the first draft, I checked that my text was well-organised.	.861	
5.9	After I had finished the first draft, I checked that my text was coherent.	.861	
5.11	After I had finished the first draft, I checked that I included my own viewpoint on the topic.	.830	
5.14	After I had finished the first draft, I checked the possible effect of my writing on the intended reader.	.806	
4.10	I checked that I included all appropriate main ideas from all the source texts while I was writing.		.789
4.7	I checked that the content was relevant while I was writing.		.780
4.8	I checked that my text was well-organised while I was writing.		.704
4.11	I checked that I included my own viewpoint on the topic while I was writing.		.701
4.9	I checked that my text was coherent while I was writing.		.686
4.14	I checked the possible effect of my writing on the intended reader.		

After removal, the rotated two-factor solution was extracted again (See Table 5.44). Similar to the real-life data, the high-level monitoring and revising phase elicited on Test Task A consisted of two distinctive processes: *high-level editing after writing* (F1) and *high-level editing while writing* (F2). However, Factor 2 elicited by Test Task A did not include the process of checking the possible effect on the intended reader (Item 4.14).

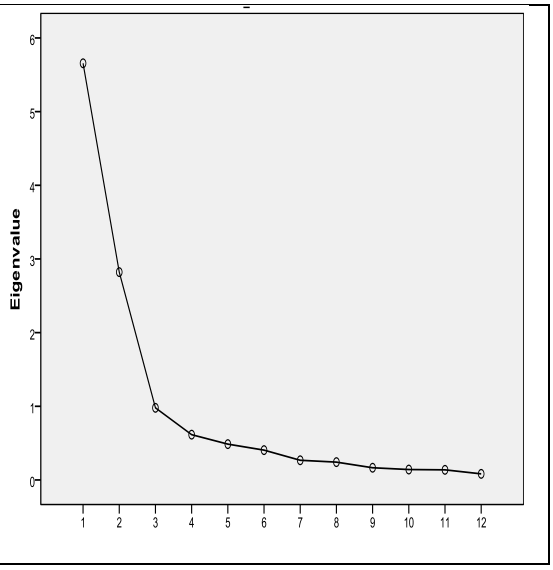
Table 5.44 Pattern and interfactor correlations matrix for the high-level monitoring and revising phase (Test Task A)

		F1 High-level editing after writing (47.42%)	F2 High-level editing while writing (19.46%)
5.7	After I had finished the first draft, I checked that the content was relevant.	.914	
5.10	After I had finished the first draft, I checked that I included all appropriate main ideas from all the source texts.	.879	
5.8	After I had finished the first draft, I checked that my text was well-organised.	.866	
5.9	After I had finished the first draft, I checked that my text was coherent.	.865	
5.11	After I had finished the first draft, I checked that I included my own viewpoint on the topic.	.832	
5.14	After I had finished the first draft, I checked the possible effect of my writing on the intended reader.	.800	
4.10	I checked that I included all appropriate main ideas from all the source texts while I was writing.		.802
4.7	I checked that the content was relevant while I was writing.		.783
4.11	I checked that I included my own viewpoint on the topic while I was writing.		.707
4.8	I checked that my text was well-organised while I was writing.		.692
4.9	I checked that my text was coherent while I was writing.		.669
Interfactor correlations matrix			
Factor 2 (High-level editing after writing)		1.000	
Factor 1 (High-level editing after writing)		.373	1.000

On Test Task B, the initial factor extraction for the high-level monitoring and revising construct produced two factors with eigenvalues greater than 1.0. The scree plot also suggested a two-factor solution (Table 5.45).

Table 5.45 Eigenvalues and scree plot for the high-level monitoring and revising phase (Test Task B)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.656	47.137	47.137
2	2.821	23.511	70.647
3	.979	8.159	78.806
4	.615	5.124	83.930
5	.487	4.054	87.985
6	.405	3.375	91.360
7	.267	2.227	93.587
8	.242	2.018	95.605
9	.166	1.382	96.986
10	.141	1.179	98.165
11	.137	1.144	99.309
12	.083	.691	100.000



The high level monitoring and revising phase elicited by Test Task B yielded the exactly same underlying factors: *high-level editing after writing* (F1) and *high-level editing during writing* (F2), as identified by the real-life data. The two factors correlated weakly at a level of 0.292 (See Table 5.46).

Table 5.46 Pattern and interfactor correlations matrix for the high-level monitoring and revising phase (Test Task B)

		F1 High-level editing after writing (47.14%)	F2 High-level editing while writing (23.51%)
5.7	After I had finished the first draft, I checked that the content was relevant.	.940	
5.8	After I had finished the first draft, I checked that my text was well-organised.	.920	
5.10	After I had finished the first draft, I checked that I included all appropriate main ideas from all the source texts.	.903	
5.11	After I had finished the first draft, I checked that I included my own viewpoint on the topic.	.893	
5.9	After I had finished the first draft, I checked that my text was coherent.	.847	
5.14	After I had finished the first draft, I checked the possible effect of my writing on the intended reader.	.842	
4.8	I checked that my text was well-organised while I was writing.		.839
4.7	I checked that the content was relevant while I was writing.		.817
4.11	I checked that I included my own viewpoint on the topic while I was writing.		.742
4.10	I checked that I included all appropriate main ideas from all the source texts while I was writing.		.673
4.9	I checked that my text was coherent while I was writing.		.610
4.14	I checked the possible effect of my writing on the intended reader while I was writing.		.470
Interfactor correlations matrix			
	Factor 2 (High-level editing after writing)	1.000	
	Factor 1 (High-level editing while writing)	.292	1.000

5.3.4.6 Summary of the underlying structure of the cognitive processes (real-life and test tasks)

Table 5.47 summarises the findings of the underlying structure of the cognitive processes employed at the five writing phases elicited by the real-life tasks and reading-into-writing test tasks (The order of the factor follows the results of the explanatory factor analyses).

Table 5.47 Summary of the underlying structure of the cognitive processes of the five cognitive phases elicited between the real-life and test conditions

	Real-life tasks	Test Task A	Test Task B
Conceptualisation phase			
F1:	Task representation and macro-planning (34%)	Task representation and macro-planning (33.98%)	Macro-planning (31.98%)
F2:	Revising macro plan (19.9%)	Revising macro plan (19.04%)	Revising macro plan (16.16%)
F3:			Rereading task prompt (16.16%)
Meaning and discourse construction phase			
F1:	Connecting and generating (34.54%)	Selecting relevant ideas (33.38%)	Selecting relevant ideas (28.20%)
F2:	Selecting relevant ideas (13.88%)	Connecting and generating with careful global reading (14.09%)	Connecting and generating (18.45%)
F3:	Careful global reading (10.16%)	-	Careful global reading (15.07%)
Organising phase			
F1:	Organising ideas in relation to input texts (34.73%)	Organising ideas in relation to input texts (41.70%)	Organising ideas in relation to input texts (40%)
F2:	Organising ideas in relation to own text (16.60%)	Organising ideas in relation to own text (17.65%)	Organising ideas in relation to own text (16.41%)
Low-level monitoring and revising phase			
F1:	Low-level editing after writing (47.70%)	Low-level editing after writing (48.18%)	Low-level editing after writing (52.31%)
F2:	Low-level editing while writing (23.9%)	Low-level editing while writing (19.31%)	Low-level editing while writing (24.85%)
High-level monitoring and revising phase			
F1:	High-level editing after writing (42.92%)	High-level editing after writing (47.42%)	High-level editing after writing (47.14%)
F2:	High-level editing while writing (24.35%)	High-level editing while writing (19.46%)	High-level editing while writing (23.51%)

As presented in detail in Section 5.3.4.1 - Section 5.3.4.5 above, the underlying structures of the four out of five phases of academic writing, which included

conceptualisation, **organising**, **low-level organising and revising**, and **high-level monitoring and revising**, elicited on Test Task A and the real-life tasks were identical. And seven factors within these phases, which included *task representation and macro-planning, revising macro plan, selecting relevant ideas, organising ideas in relation to input texts, low-level editing after writing, low-level editing while writing* and *high-level editing after writing,* elicited by Test Task A contained the same individual questionnaire items as the corresponding factors identified by the real-life tasks.

Besides, the underlying structures of the cognitive processes of four phases of academic writing, which included **discourse and meaning construction**, **organising**, **low-level organising and revising**, and **high-level monitoring and revising**, elicited on Test Task B and the real-life tasks were identical, though the order of the factors of the discourse and meaning construction was different between Test Task B and the real-life tasks. And eight factors, which included *revising macro plan, selecting relevant ideas, careful global reading, organising ideas in relation to input texts, low-level editing after writing, low-level editing while writing, high-level editing after writing* and *high-level editing while writing,* elicited by Test Task B contained the same individual questionnaire items as the corresponding factors identified by the real-life tasks. Confirmatory factor analysis should be employed in future studies to test the underlying structures extracted by the real-life and test data in this study.

The results of the explanatory factor analysis revealed that the two reading-into-writing test tasks were able to elicit most of the academic writing processes in the same manner as they were employed on the real-life tasks, especially for the cognitive processes of the organising, low-level organising and revising, and high-level monitoring and revising phases. Nevertheless, some discrepancies were shown on the underlying structure of the cognitive processes of the meaning and discourse construction phase between Test Task A and the real-life tasks, and of the conceptualisation phase between Test Task B and the real-life tasks.

Regarding the conceptualisation phase, both real-life tasks and Test Task A elicited two factors: *task representation and macro-planning* and *revising macro*

plan. However, the processes of *reading task prompt again* clustered as an additional factor on Test Task B. As discussed earlier, this is probably because Test Task B was rated the highest among all tasks in this study for its *clarity of purpose, clarity of intended reader* and *clarity of the knowledge criteria.* It seems that a clear presentation of these features would encourage test takers to monitor their progress by checking the task prompt from time to time during the writing process. Further evidence is needed to confirm this finding.

Regarding the meaning and discourse construction phase, the processes of *connecting and generating* was identified as the first factor on the real-life tasks, *selecting relevant ideas* the second, and *careful global reading* the third. The order of the factors indicates the percentage of the variance explained by each factor. In other words, the process of *connecting and generating* accounted for the largest percentage (i.e. 34.54%) of the variance of all cognitive processes within the meaning and discourse construction phase elicited by the real-life tasks. In contrast, both reading-into-writing test tasks had the processes of *selecting relevant ideas* (search reading) as the first factor of the meaning and discourse construction phase. This implies that the process of *selecting relevant ideas* was most important within the meaning and discourse construction phase elicited by the test tasks. In addition, the process of careful global reading employed on Test Task A (essay task with multiple verbal inputs) did not yield a stand-alone factor. The processes of careful global reading clustered with the processes of connecting and generating as the second factor. Test Task B (essay task with multiple verbal and non-verbal inputs) elicited careful global reading as a distinct factor but there was a seemingly over-eliciting of careful global reading on Test Task B. Further studies should investigate into these issues. Nevertheless, although there were some discrepancies in the underlying structure of the meaning and discourse construction phase activated by the real-life academic writing tasks and the two test tasks, the participants reported employing the process of expeditiously selecting relevant ideas more than careful reading processes on the two reading-into-writing test tasks as they did on the real-life tasks.

The organising phase elicited by the real-life tasks involved two distinct cognitive processes. The findings showed that the participants distinctively employed the processes to organise their representation of the input texts and those to organise their own text, with a stronger attention on the latter, under real-life conditions. Generally speaking, Test Task A and Test Task B were able to elicit these two distinct organising processes from the participants. However, the factors of *organising ideas in relation to own text* elicited on both test tasks involved less process items than the corresponding factor identified by the real-life data. It requires further evidence to discuss why the participants did not organise their own text in the same way they did in real-life conditions. For example, the process of removing ideas (Item 3.3) did not load on any of the factors at the level of 0.3 or above on Test Task A. It might be helpful to provide guidelines on task prompts to encourage test takers to devote equal attention to both organising processes.

Both reading-into-writing tasks did well in eliciting the same underlying processes of the low-level and high-level monitoring and revising phases. Both test tasks were able to elicit the processes of *while writing* and *after writing* low-level editing and high-level editing from test takers in the same manner as they were employed on the real-life tasks.

5.4 Summary of the chapter

Regarding the cognitive processes elicited by the real-life tasks, the results reported in this chapter revealed eleven cognitive processes: (1) task representation and macro-planning, (2) revising macro plan, (3) connecting and generating, (4) selecting relevant ideas, (5) careful global reading, (6) organising ideas in relation to input texts, (7) organising ideas in relation to own text, (8) low-level editing while writing, (9) low-level editing after writing, (10) high-level editing while writing, and (11) high-level editing after writing. There is a good case for considering these as the target cognitive processes for a valid academic writing test. The results also showed that high achieving participants reported

employing most of the eleven processes significantly more than low achieving participants in real-life conditions.

The analysis comparing the extent to which these eleven processes were employed by the high-, middle- and low-achieving groups between real-life and test conditions revealed positive results for the cognitive validity of the reading-into-writing test tasks. Both Test Task A (essay task with multiple verbal inputs) and Test Task B (essay task with multiple verbal and non-verbal inputs) were able to elicit from high-achieving and low-achieving participants most of the cognitive processes in a similar manner to the way the participants employed the processes on the real-life tasks. In comparison, the middle-achieving group showed greater discrepancy in the way they employed some processes on test tasks and real-life tasks. It would be interesting in future research to consider why, under test conditions, the middle-achieving participants employed some of the real life processes but not others on Test Task A and Test Task B.

In addition, the results of the explanatory factor analysis provided positive evidence for the cognitive validity of the reading-into-writing test tasks as a tool to assess academic writing ability. Both Test Task A and Test Task B were able to elicit from the participants most of the underlying factors of cognitive processes yielded by the real-life data. Common factors of cognitive processes yielded by the test tasks and real-life tasks included the processes of *revising macro plan, selecting relevant ideas, organising ideas in relation to input texts, low-level editing while writing, low-level editing after writing* and *high-level editing after writing.*

Chapter Six will shift the attention to an a posteriori component of test validity – criterion-related validity, to explore the extent to which the participants' reading-into-writing test scores correlate with the scores on their real-life academic writing tasks.

6 ESTABLISHING THE CRITERION-RELATED VALIDITY OF READING-INTO-WRITING TESTS TO ASSESS ACADEMIC WRITING ABILITY

6.1 Introduction

Chapter Four and Chapter Five have reported and discussed the results of two key a priori components of the socio-cognitive test validation framework (Weir, 2005): context validity and cognitive validity. Context validity concerns the contextual parameters of the reading-into-writing test tasks in terms of overall task setting and input text features. Cognitive validity concerns the cognitive processes elicited by the reading-into-writing test tasks. Chapter Six focuses on an a posteriori component: criterion-related validity, which concerns 'the extent to which test scores correlate with a suitable external criterion of performance with established properties' (Weir, 2005:35). This study investigated the correlations between the participants' reading-into-writing test scores and their real-life academic performances on different writing tasks in their degree course work and examinations.

Section 6.2 presents the results of the participants' performances on the two reading-into-writing test tasks and their performances on a range of real-life writing tasks. Some details of the participants' proficiency level as measured by IELTS reading and writing scores are provided in Section 6.2.1 as a reference. Results on Test Task A and Test Task B are presented first in Sections 6.2.2 and 6.2.3 respectively. After that, results on the selected real-life writing tasks in their degree course work and examinations are presented in Section 6.2.4. A summary is provided in Section 6.2.5.

Section 6.3 presents the results from the correlational analysis between the two reading-into-writing test scores and the real-life scores, and discusses the extent to which the reading-into-writing test scores relate to the test takers' ability to perform on real-life academic writing tasks. The results of the correlations between test scores and individual real-life scores are presented in Section 6.3.1 whereas the results of the correlations between test scores and overall real-life scores are presented in Section 6.3.2. Section 6.3.3 further discusses the patterns of the correlations between the test scores and overall real-life scores. A summary of the chapter is provided in Section 6.3.4.

6.2 Participants' performances

Chapter Three reported the procedures of selecting the suitable writing tasks in the real-life academic context as the external criterion of the participants' performances on the two reading-into-writing test tasks (Test Task A and Test Task B). Four points of reference were selected for the analysis of the criterion-related validity. In addition to the *essay* task and the *report* task used in the previous analysis of context validity and cognitive validity, an in-class *question-and-answer test* task and an end-of-term *case study examination* task were selected (For the details of the procedures, see Section 3.5.2.1). Table 6.1 summarises the features of these four selected real-life tasks as well as the two reading-into-writing test tasks.

Table 6.1 The 4 selected real-life tasks and 2 reading-into-writing test tasks for the correlational analysis

Condition	Task	Task instructions	Input	Time	Output
Real-life academic context	Essay	Write an essay on a given topic - Summarise salient issues - Discuss the issues with justified personal views	A stimulus article with non-verbal, e.g. diagrams, pictures (Students are expected to make use of other input texts of their choice)	N/A	5000 words
	Report	Write a report to forecast the business of a company - Describe the data - Discuss and justify ways of analysis - Make recommendations	A passage (less than 200 words) plus a numeric dataset (Students are expected to make use of other input texts of their choice)	N/A	2000 words
	In-class question and answer test	Demonstrate understanding of core concepts and theories: Examples: - Critically examine *X*. Justify your answer using appropriate examples. - Discuss the potential value of *X*. Give examples to support your arguments.	4-5 questions (about 20-30 words each)	1 hour	No specific word limits
	End-of-term case study examination	Write an essay based on a case study (provided in advance) - Critically analyse the issues presented in the case study - Make recommendations and justify with reasons	A case study with non-verbal input (2500 words)	2 hours	No specific word limits
Reading-into-writing language tests	Test Task A	Write a comparative essay summarising the main ideas from verbal input and stating own viewpoints	2 articles without non-verbal input	1 hour	At least 250 words
	Test Task B	Summarise the main ideas from both verbal input and non-verbal input and express opinions	2 articles with a non-verbal input each	1 hour	180-200 words

6.2.1 Participants' proficiency level in English (measured by IELTS reading and writing)

As reported in Chapter Three, the mean scores of the 291 participants' IELTS *reading* and *writing* were 5.88 and 5.54 respectively. As both reading-into-writing test tasks (Test Tasks A and B) and the selected real-life writing tasks involve considerable reading, the average of the participants' IELTS reading and writing bands were computed for analysis (See Table 6.2).

Table 6.2 Participants' IELTS bands

	Minimum	Maximum	Mean	Std. Deviation
IELTS Reading	5.00	7.50	5.88	0.60
IELTS Writing	4.50	7.00	5.54	0.49
Average IELTS Reading and Writing	5.00	7.00	5.81	0.43

The distribution of the participants' IELTS *average reading and writing* bands is presented in Table 6.3. All IELTS scores were effective at the time of the study in line with the University's admission policy, i.e. within 2 years for effectiveness. Students who score IELTS 6.0 can be offered places on 3-year Bachelor's; students who score IELTS 5.0 can be offered places on 4-year extended Bachelor's programmes, which include compulsory extra English classes in the first year. In this study, 54.4% of the participants had an IELTS average reading and writing band of 6.0 or above whereas 45.6% had an IELTS overall band of 5.0 or 5.5. According to Figure 6.1, which showed the indicative IELTS bands at CEFR levels, 12.8% of the participants were at C1 level and 87.2% were at B2 level. As described in Chapter Three, Test Task A is a level-specific test targeting CEFR C1 level whereas Test Task B is a University's diagnostic test at B2 level.

Table 6.3 Frequency table of the participants' average IELTS reading and writing bands

Average IELTS reading and writing band	Frequency	Per cent
7.0	5	2.3
6.5	23	10.5
6.0	91	41.6
5.5	83	37.9
5.0	17	7.8
Total	291	100

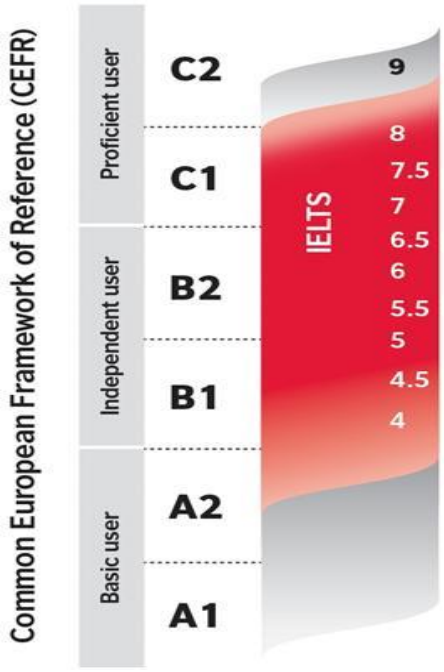


Figure 6.1 Indicative IELTS band scores at CEFR levels

(http://www.ielts.org/researchers/common_european_framework.aspx)

Although most participants in this study did not reach the target level of Test Task A, i.e. C1, the sampled proficiency range is still considered to be appropriate for the purpose of this thesis. This is because the sampled range reflects a typical IELTS score distribution of overseas undergraduate students admitted to study at UK universities, and the results and implications drawn from this study therefore provide 1) a more realistic picture of the criterion-related validity of Test Task A,

when the test is used as an admission test for UK universities; and 2) necessary criterion-related evidence of Test Task B as the University's diagnostic test of academic writing needs.

6.2.2 Participants' performance on Test Task A

160 performances on Test Task A were marked by the test provider, i.e. LTTC, following their standard operationalised procedures (see Appendix 6.1 for the marking scheme of Test Task A). All scripts were double marked and 5% were marked by a third rater. Each script was scored using four analytical marking categories: (1) relevance and adequacy, (2) coherence and organisation, (3) lexical use, and (4) grammatical use (The public marking scheme is provided in Appendix 6.1). *Relevance and adequacy* concerns whether the text is relevant to the writing task, and whether all parts of the writing task are fully addressed. *Coherence and organisation* concerns whether the text shows coherence and cohesion, and whether the organisational structure of the text at different levels is clear. *Lexical use* and *grammatical use* concern the range and appropriateness of the lexical use and of grammatical use of the text respectively. Each category can be scored from 1 to 5 with 3 being the threshold. An overall band 3 on all of the four analytical categories is required to pass Test Task A. In real-life operationalised contexts, LTTC reports only the overall band (1-5) to candidates.

In this study, 17 participants (10.6%) passed Test Task A (i.e. obtained a minimum of total analytical scores of 12 with a minimum score of 3 on all analytical categories). The low passing rate on Test Task A was expected because only about 12 % of the participants were at C1 level based on their IELTS band. 10 of the 17 participants who passed Test Task A had an average reading and writing band 6.5 or above. In order to generate more insightful information about the participants' performances on Test Task A, this chapter focuses the discussion on the participants' individual analytical scores and total analytical scores instead of the overall band. Descriptive statistics of the 160 participants' scores on individual analytical categories and their total analytical scores on Test Task A are presented in Table 6.4. The participants' mean total analytical score on Test

Task A was 8.72 with a standard deviation of 2.10. The means of the four analytical scores were seemingly close, ranging from 2.01 to 2.34. The mean score on *coherence and organisation* was highest whereas the mean score on *grammatical use* was lowest.

Table 6.4 Descriptive statistics on Test Task A scores

	Max. possible score	Min.	Max.	Mean	Std. Dev.
Test Task A total scores	20	4.0	13.5	8.72	2.10
Relevance and adequacy	5	1.0	4.0	2.27	0.65
Coherence and organisation	5	1.0	3.5	2.34	0.60
Lexical use	5	1.0	3.0	2.15	0.58
Grammatical use	5	1.0	3.0	2.01	0.57

In order to understand more comprehensively the participants' performances on Test Task A, Table 6.5 presents the frequency of the four analytical scores on Test Task A. Test Task A is a criterion-referenced test which is aimed specifically at the C1 level. It reports how well candidates are doing relative to a pre-determined performance level on a specified set of goals. 12.8% of the participants in this study were presumably at C1 level according to their IELTS bands. As indicated in Table 6.5, more participants achieved Band 3, i.e. the pass band on the categories *relevance and adequacy* (30%) and *coherence and organisation* (31.2%) than on the categories of *lexical use* (19%) and *grammatical use* (15%).

Table 6.5 Frequency table of analytical scores on Test Task A

		Relevance and adequacy		Coherence and organisation		Lexical use		Grammatical use	
		Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
Fail	1.0	17	10.6	13	8.1	15	9.4	19	11.9
	1.5	10	6.3	6	3.8	15	9.4	20	12.5
	2.0	52	32.5	51	31.9	67	41.9	85	53.1
	2.5	33	20.6	40	25.0	32	20.0	12	7.5
Pass	3.0	47	29.4	49	30.6	31	19.4	24	15.0
	3.5	0	0	1	0.6	0	0	0	0
	4.0	1	0.6	0	0	0	0	0	0
	5.0	0	0	0	0	0	0	0	0

	Total	160	100.0	160	100.0	160	100.0	160	100.0

The distribution of the total analytical scores on Test Task A is presented in Figure 6.2. The curve in the histogram represents the distribution curve. The curve is skewed towards the lower end of the range of total analytical scores. As mentioned before, a low pass rate on Test Task A was expected. However, even though 17 participants passed Test Task A, no participants in this study scored above 14 out of 20.

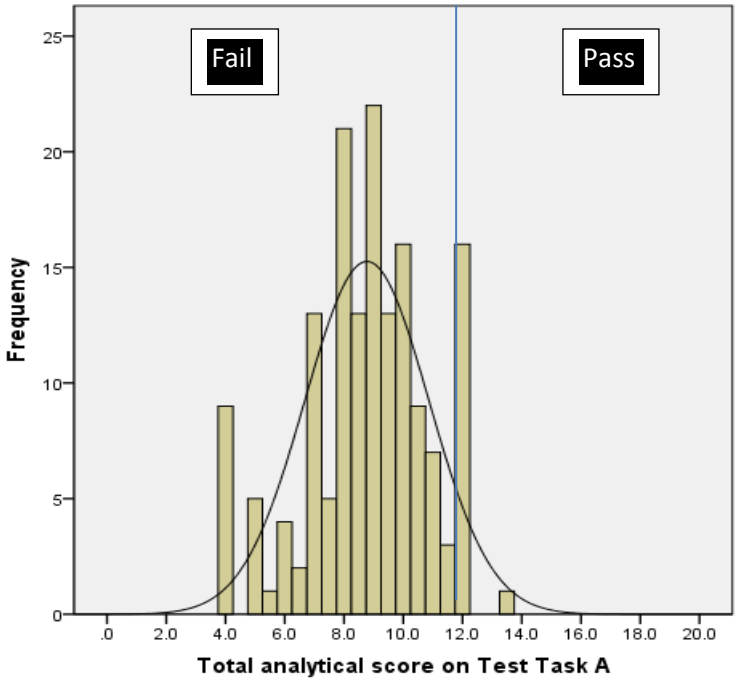


Figure 6.2 Distribution of the total analytical scores on Test Task A

The next sub-section presents the participants' performances on Test Task B.

6.2.3 Participants' performance on Test Task B

140 performances on Test Task B (essay task with multiple verbal and non-verbal inputs) were marked by the test provider, i.e. CRELLA (see Appendix 6.2 for the marking scheme of Test Task B). . 30% of the scripts were double marked. Each script was scored by three analytical marking categories, i.e., content, organisation and language (The marking scheme is provided in Appendix 6.2).

The *content* category refers to the extent to which the writer has responded appropriately to the task and the specific instructions given about the relationship between the input reading material and the written output. It covers the inclusion of all essential key points, as well as communicative effect on the reader. The *organisation* category refers to the way in which the written production has been structured and organised in terms of the overall format, the grouping and sequencing of ideas in paragraphs, and the coherence of the argumentation. It covers the notion of cohesion and coherence at levels of sentences and paragraphs. The *language* category refers to the clarity of linguistic expression in English, including the selection and control of grammar and vocabulary items. It also includes stylistic choices relating to academic register.

Each category can be scored from 1 to 3. Score 1 indicates a significantly weak performance, score 2 indicates a below adequate performance and score 3 indicates an adequate performance. Texts that are too short, completely off topic, illegible or plagiarised are scored 0. Test Task B was still at a trial stage when the study was conducted. Test scores were proposed to report both at the *overall* (i.e. the total analytical scores) and *analytical* levels. The trial grade boundaries were as follows:

A score of 8-9	=	Grade A (no intervention required)
A score of 6/7	=	Grade B (low-level intervention needed)
A score of 5 or <	=	Grade C (high-level intervention needed)

Descriptive statistics on the 140 performances on Test Task B are presented in Table 6.6. The participants' mean total analytical score on Test Task B was 4.99 out of 9 with a standard deviation of 1.76. The mean score on *organisation* was higher than the mean scores on *content* and *language* (See Table 6.6).

Table 6.6 Descriptive statistics on Test Task B scores

	Max possible score	Min.	Max.	Mean	Std. Dev.
Test Task B total scores	9	0	9	4.99	1.76
Content (coverage of key points)	3	0	3	1.64	0.66
Organisation (cohesion and coherence)	3	0	3	1.78	0.68
Language (choice and control of lexis and grammar)	3	0	3	1.60	0.70

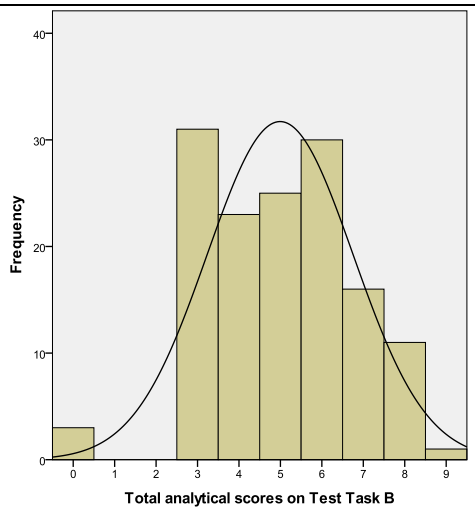
In order to understand more comprehensively the participants' performances on Test Task B, Table 6.7 presents the frequency of the three analytical scores on Test Task B. Test Task B is a university diagnostic test of test takers' academic writing ability. 8.6%, 12.9% and 10.7% of participants scored band 3 (which indicates an adequate or above adequate performance) on the categories of *content*, *organisation* and *language* respectively. Most participants were scored band 2 (which indicates a below adequate performance) on *content* and *organisation* but most participants were scored band 1 (which indicates a significantly weak performance) on *language.* Similar to the results indicated by Test Task A, the participants seemed to perform less successfully on the *language* category than other analytical categories.

Table 6.7 Frequency table of analytical scores on Test Task B

	Content (coverage of key points)		Organisation (cohesion and coherence)		Language (choice and control of lexis and grammar)		
	Freq	Percent	Freq	Percent	Freq	Percent	Implication of score
0	2	1.4	2	1.4	2	1.4	too short, completely off topic, illegible or plagiarised
1.0	59	42.1	45	32.1	67	47.9	a significantly weak performance
2.0	67	47.9	75	53.6	56	40.0	a below adequate performance

3.0	12	8.6	18	12.9	15	10.7	an adequate or above adequate performance
Total	140	100.0	140	100.0	140	100.0	

The distribution of the total analytical scores on Test Task B is presented in Figure 6.3. The distribution curve reaches both ends of the score range. The results indicated that 8.6%, who scored a total of 8 or above, did not require any intervention to their academic writing ability. 32.8% of the participants, who scored 6 or 7, required low-level intervention whereas 58.6% of the participants, who scored 5 or below, required high-level intervention.



	Freq	Percent	Cumulative Percent	Implication of scores
0	3	2.1	58.6	high-level intervention needed
3	31	22.1		
4	23	16.4		
5	25	17.9		
6	30	21.4	32.8	low-level intervention needed
7	16	11.4		
8	11	7.9	8.6	no intervention needed
9	1	0.7		
Total	140	100.0		

Figure 6.3 Distribution of the total scores on Test Task B

6.2.4 Participants' performance on the real-life tasks

The participants' performances on the two test tasks have been discussed so far. This sub-section discusses their performances on the selected real-life tasks. Each real-life performance (i.e. the essay task, the report task, the in-class question-and-answer test and the end-of-term case-study examination), can award a score from 0 to 16, representing 5 bands (See Table 6.8).

Table 6.8 Real-life scores and the corresponding grades

Score	Band	Pass/Fail
16-14	A+/A/A-	Pass
13-11	B+/B/B-	
10-8	C+/C/C-	
7-5	D+/D/D-	
4-0	E	Fail

Participants' performances on the four selected real-life tasks were scored by the lecturers at the University's Business School. As reported in Chapter Three, the four tasks were selected from four different modules. The four tasks were scored by different module teams. All marking followed university departmental marking procedures. Lecturers who marked the real-life performances were not informed of the present study. In addition, they were not informed of the students' IELTS scores or their performances on the two reading-into-writing test tasks.

The essay task was scored based on four categories: (1) examination of the data and description of the nature of the dataset; (2) discussion and justification of the techniques chosen; (3) reasons for rejecting the inappropriate techniques; and (4) discussion of other relevant issues. The report task was scored based on: (1) problem definition and structure of the text; (2) information identification (the number of sources, relevance to the topic, reliability of the sources); (3) critical reasoning; and (4) persuasion and influencing (See Appendices 3.1.1 and 3.1.2). The marking scheme of the in-class test and the end-of-term examination was not available to the researcher. In addition, the sub-scores of the tasks were not available to the researcher. All marking was conducted by the lecturers following the university departmental marking procedures. The real-life data reported in this chapter were the final standardised marks submitted to the University.

Table 6.9 presents the descriptive statistics of the scores for each selected real-life task. Participants in this study, as a whole group, performed best on the report task (mean: 9.72) while scoring lowest on the exam (mean: 6.03). The standard deviations of the four performances were moderate.

Table 6.9 Descriptive statistics of real-life performances

Task	N[11]	Max possible score	Minimum	Maximum	Mean	Std. Deviation
Essay	161[12]	16	2	15	9.27	2.99
Report	136[13]	16	2	14	9.72	2.60
Test	145	16	2	15	8.84	2.98
Exam	143	16	2	14	6.03	2.56

The distribution of the participants' scores on the four real-life tasks is presented in Figure 6.4. According to the histograms, participants' scores on the essay task and the in-class test largely follow the normal distribution curve. However, participants' scores on the report task clustered on the score of 10 (i.e. grade C+) and their scores on the end-of-term examination clustered between the score 4 to 6 (i.e. E, D- and D correspondingly). A score of 4 or below indicates a fail on the task. 7.4% participants in this study failed the essay task, 3.7% failed the report task, 8.3% failed the question-and-answer test and 33.6% failed the case-study examination. It is perhaps not surprising that the pass rate on the case-study examination task was much lower than the other three tasks. According to one of the lecturers, one major purpose of the examination is for the participants to demonstrate the subject knowledge they had learnt on the module. The examination task was presumably more challenging than the other three tasks.

[11] For the 219 participants in this study, 56 enrolled one of the modules, 21 enrolled two of the modules, 80 enrolled three of the modules and 62 enrolled all the modules.
[12] 73 out of the 161 participated in the investigation of cognitive processes.
[13] 70 out of the 136 participated in the investigation of cognitive processes.

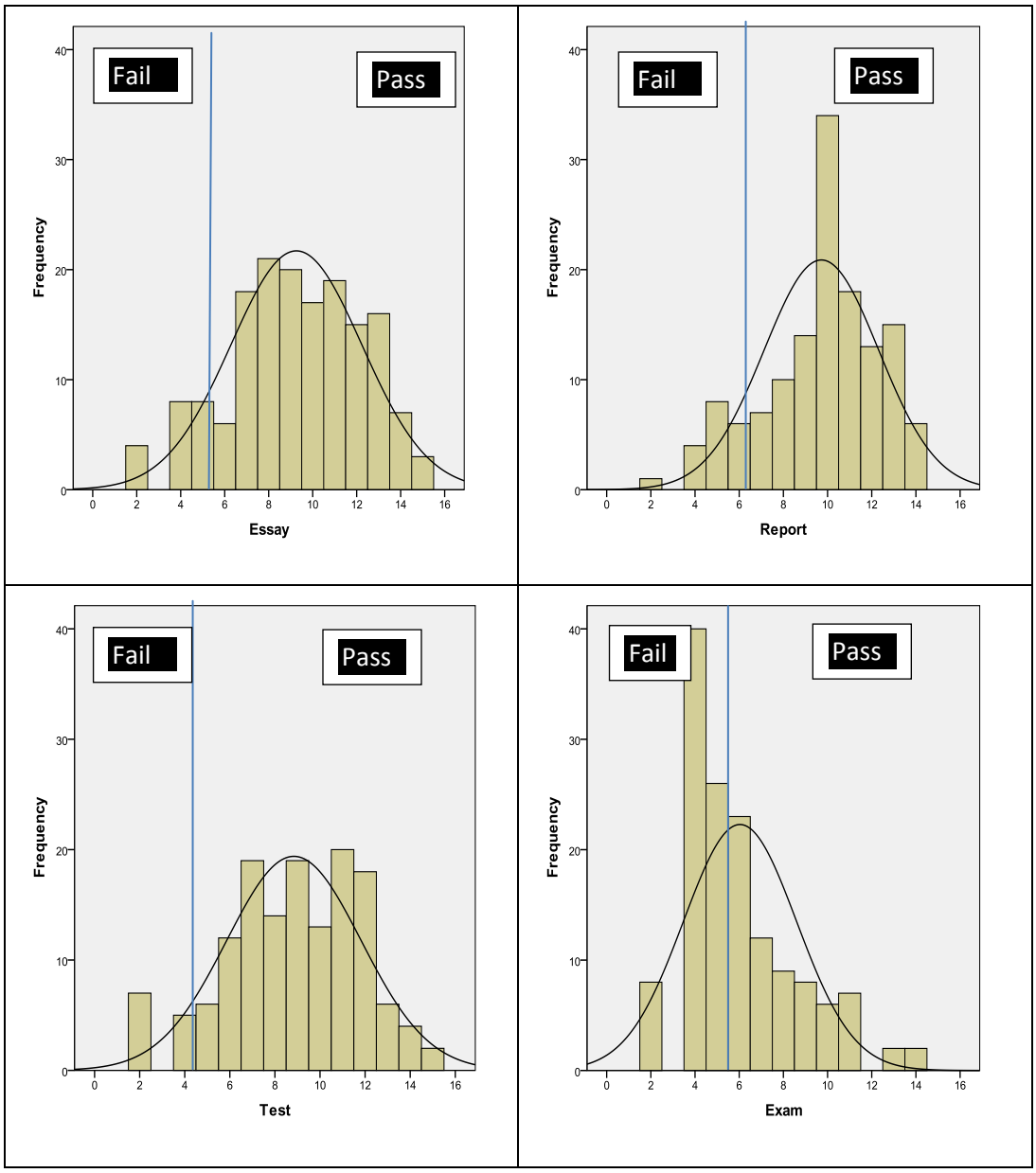


Figure 6.4 Score distribution of the four real-life tasks

For the purpose of correlational analysis, the participants' mean scores of the four selected tasks i.e. the essay, report, question-and-answer test, case study examination tasks, were calculated. As shown in Table 6.10, the participants' mean real-life score was 8.59.

Table 6.10 Descriptive statistics of mean real-life performances

N	Max possible score	Minimum	Maximum	Mean	Std. Deviation
219	16	2	14	8.59	2.24

Figure 6.5 below shows the frequency of participants at each corresponding grade based on their mean real-life scores. Only 1 participant (0.5%) in this study got an average Grade A. 21% of the participants got an average Grade B and about 30% got Grade D. Most participants (45.2%) got an average Grade D and 3.7% got an average Grade E (Fail).

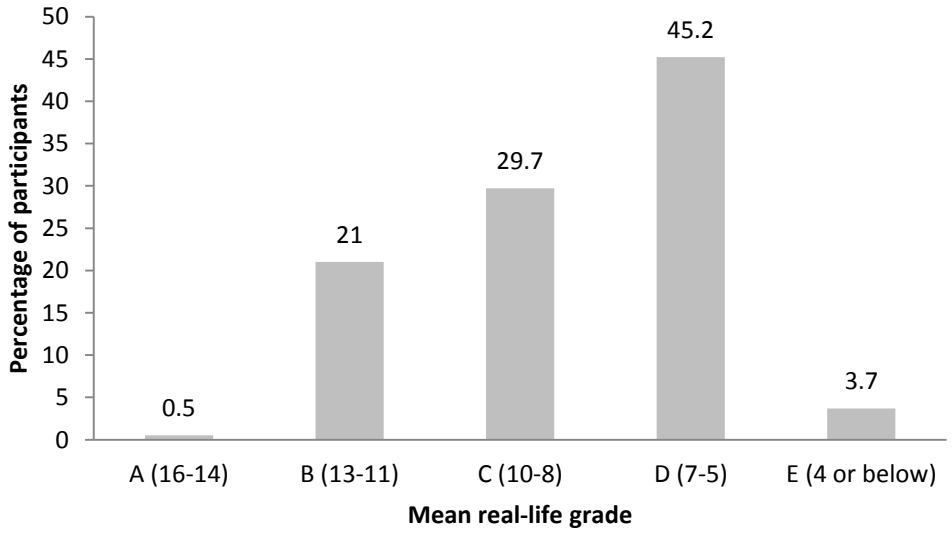


Figure 6.5 Mean real-life grade

6.2.5 Summary

There are reasons to be somewhat cautious in interpreting the participants' performances on the two reading-into-writing tests. The participants may have

been less familiar with reading-into-writing test tasks which involve multiple reading materials, even though they had been briefed about the format of Test Task A (with multiple verbal inputs) and Test Task B (with multiple verbal and non-verbal inputs) one or two weeks before they did the test. In addition, the participants might not have perceived Test Tasks A and B as high-stakes for them because their scores on Test Tasks A and B would not affect their university grades. Although Test Task B was a university diagnostic test, the participants completed the task solely for the research purpose and they did not receive any corresponding support based on their Test Task B scores. The functions of Test Task A (which is part of a level-specific proficiency test at C1 level) and Test Task B (which is part of an academic writing diagnostic test) are different. Therefore, it is not appropriate to make direct comparison of the test results between two test tasks.

All participants in this study provided information of their IELTS bands. 12.8% of the participants in this study had an average IELTS reading and writing score between 6.5 and 8.0, a range which indicates a proficiency level at C1, and 87.2% between 5.0 and 6.5, a range which indicates a proficiency level at B2 (See Figure 6.1).

The results showed that 10.6% of the 160 participants who completed Test Task A passed the test task, which is at C1 level. Regarding Test Task B, 8.6% of the 140 who completed the task got Grade A, which indicates that the test taker did not require any intervention to his/her academic writing ability. 32.8% got Grade B, which indicates the test taker's need of low-level intervention to his/her academic writing ability, and 58.6% got Grade C, which indicates the test taker's need of high-level intervention.

Regarding the participants' performance on real-life academic writing tasks, 0.5% of the whole population in this study (i.e. 219 participants) got an average Grade A, 21% Grade B, and 29.7% Grade C. Most participants (45.2%) got an average Grade D whereas 3.7% got an average Grade E (Fail).

The results showed some similarities in the participants' performances on the two reading-into-writing test tasks. In terms of the analytical scores, the results showed that the participants in this study were scored highest for the organisation category, followed by the content category, and lowest on the linguistic category on both reading-into-writing test tasks. It appears that both reading-into-writing test tasks provided a similar picture of the participants' strengths and weaknesses in terms of analytical categories.

The major difference found between Test Task A and Test Task B scores was the range of scores. On Test Task A, almost no performance was scored band 4 or band 5 from a scale of 5 on all the four marking categories. On Test Task B, the full range of scores was achieved. This is most likely because Test Task A was set at C1 level whereas Test Task B was set at B2 level.

The results so far have shown the percentage of the participants who did not reach CEFR C1 level (based on Test Task A) or the percentage of those who were identified to have significantly weak academic writing ability (based on Test Task B). The percentage of the participants who got an overall real-life grade from A to E, which indicate the different degree of academic success, has also been reported. The correlations between reading-into-writing test scores and real-life writing task scores are discussed in the following section to provide insights of the extent to which academic performance could be predicted by these test scores.

6.3 Correlations between reading-into-writing test scores and real-life writing task scores

The previous sections presented the participants' performances on the two reading-into-writing test tasks (Test Tasks A and B), and their performances on the four selected real-life academic writing tasks (i.e. essay, report, question-and-answer test and case-study examination). This section examines the relationships of these scores obtained between the test conditions and the real-life academic conditions, and discusses the extent to which the performances on the two types of reading-into-writing test tasks accounted for the participants' writing performance on their course work, test and examination. Section 6.3.1 discusses the correlations between test scores and individual real-life scores whereas Section 6.3.2 discusses the correlations between test scores and the average scores of all real-life tasks.

6.3.1 Correlations between test scores and individual real-life scores

For the 160 participants who completed Test Task A, 96 completed the report task, 111 completed the essay task, 109 completed the question-and-answer test, and 99 completed the end-of-term case study examination as part of their degree programme. As indicated in Table 6.11 below, the correlations between Test Task A and individual real-life scores ranged from 0.126 to 0.343. Test Task A scores correlated moderately with report scores at $r=0.343$ ($p<0.001$), weakly with question-and-answer test scores at $r=0.212$ ($p=0.027$) and essay scores at $r=0.187$ ($p=0.050$). Test Task A scores did not correlate significantly with the end-of-term examination scores.

Table 6.11 Correlation between Test Task A scores and individual real-life scores

		Essay (n=111)	Report (n=96)	Test (n=109)	Exam (n=99)
Test Task A	Pearson Correlation[14]	.187*	.343**	.212*	.126

[14] Non-parametric correlation tests were also performed. The significant results were not affected by the use of parametric tests.

	Sig. (2-tailed)	.050	.001	.027	.215

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

For the 140 participants who completed Test Task A, 96 completed the report task, 111 completed the essay task, 109 completed the question-and-answer test, and 99 completed the end-of-term case study examination as part of their degree programme. The correlations between Test Task B and individual real-life writing tasks ranged from 0.082 to 0.438 (See Table 6.12). Test Task B scores correlated moderately with question-and-answer test scores ($r=0.438$, $p<0.001$) and with essay scores ($r=0.386$, $p<0.001$), and weakly with report scores at $r=0.283$ ($p=0.005$). Test Task B scores correlated weakly with case study examination at $r=0.082$ but, similar to Test Task A, the correlation between Test Task B scores and case-study examination scores was non significant.

Table 6.12 Correlation between Test Task B scores and individual real-life scores

		Essay (n=93)	Report (n=95)	Test (n=75)	Exam (n=69)
Test Task B	Pearson Correlation	.386**	.283**	.438**	.082
	Sig. (2-tailed)	.000	.005	.000	.484

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Most test tasks are not designed to predict test taker's writing ability on a single task type, but a representative range of task types that test takers are likely to encounter in the real-life context. The weak to moderate correlations between test scores and individual real-life scores are not surprising. Nevertheless, the results of the two test scores seem to have predicted performances on particular real-life tasks better than the others. Comparing between two test scores, Test Task B scores have better correlations with the real-life *essay* and *in-class test* scores than Test Task A, whereas Test Task A scores have better correlations with the real-life *report* and *end-of-term* scores than Test Task B.

As discussed in Chapter Two, based on the results of contextual analysis, Test Task A seems to be more comparable to the real-life essay task in terms of task features whereas Test Task B seems to be more comparable to the real-life report task. Therefore, the results that Test Task A correlated with the real-life report task better whereas Test Task B correlated with the real-life essay task better are to some extent unexpected. It appears that task difficulty level might have had a bigger impact on the degree of correlation than task features. The real-life report task was regarded to be more challenging than the real-life essay task by the judges For example, the report task was regarded to be more challenging than the essay task in terms of cognitive demands, number of language functions to perform, and explicitness of textual organisation of the input texts (See Section 4.2 and Section 4.3). This may be one of the reasons why Test Task A (which is at C1 level) correlated with the real-life report task better whereas Test Task B (which is at a lower B2 level) correlated with the real-life essay task better.

Another interesting finding is that both Test Task A and Test Task B scores did not correlate significantly with the end-of-term examination scores. The first reason could be due to time effect. The participants completed the test tasks at the beginning of the term. The report task, the essay task, and the question-and-answer task were assigned to the participants during the term whereas the case-study examination was assigned towards the end of the term. The correlations between test scores and the real-life scores might have dropped due to the time effect. An increase of the time length between the two events means that many other factors may have interfered with the correlations. For example, the participants' proficiency might have improved. Their knowledge of academic writing might have improved. The amount of the subject knowledge required by the task could be another reason why the test scores did not have significant correlations to the examination scores. The case-study examination requires the students to a) critically analyse the issues presented in the case study, b) make recommendations and justify with reasons (See Table 6.1). According to one of the lecturers, one major purpose of the examination was for the participants to demonstrate the subject knowledge they had acquired on the module. Therefore,

the variable of subject knowledge might as well have contributed to the examination scores more than the participants' writing ability.

As the results showed that Test Task A and Test Task B scores correlated with individual real-life tasks at different levels, it is recommended to collect evidence from more than one task type, especially in high stakes writing tests. As argued in the literature, such practice would help to generate a more comprehensive evaluation of test takers' academic writing ability (Weigle, 2002; Shaw & Weir, 2007).

6.3.2 Correlations between test scores and mean real-life scores

As test scores are used to infer test takers' ability in performing different writing tasks in the target context, the next step was to examine the extent to which the test scores (total scores and analytical scores) relate to the average scores of the four real-life writing tasks (See Table 6.13).

Table 6.13 Correlation between Test Task A scores and mean real-life scores

		Mean real-life score			Mean real-life score
Test Task A total scores (n=160)	Pearson Correlation	.306**	Test Task B total scores (n=140)	Pearson Correlation	.379**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
Relevance and adequacy	Pearson Correlation	.160*	Content	Pearson Correlation	.300*
	Sig. (2-tailed)	.043		Sig. (2-tailed)	.000
Coherence and organisation	Pearson Correlation	.226**	Organisation	Pearson Correlation	.365**
	Sig. (2-tailed)	.004		Sig. (2-tailed)	.000
Lexical use	Pearson Correlation	.306**	Language	Pearson Correlation	.307**
	Sig. (2-tailed)	.000		Sig. (2-tailed)	.000
Grammatical use	Pearson Correlation	.391**			
	Sig. (2-tailed)	.000			

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

As shown in Table 6.13, Test Task A total scores correlated significantly with their mean real-life scores at a level of 0.31, explaining 9.36% variance of the real-life performances. Test Task B total scores correlated significantly with their mean real-life scores at a level of 0.38, explaining 14.36% variance of the real-life performances. The results of this study are better than most previously reported figures in the literature. As reviewed in Chapter Two, for correlations between overall test scores and academic outcome, some studies found no significant correlations between the overall IELTS test scores and academic scores (e.g. Cotton & Conrow, 1998; Ingram & Bayliss, 2007; Dooey, 1999). Some other studies found low correlations between overall test scores and academic scores (e.g. Cho & Bridgeman, 2012; Kerstjens & Nery, 2000). Some found low to medium correlations between overall test scores and academic outcome. For example, Davies and Criper (1988) reported a correlation of 0.3 between ELTS scores and academic outcome. They concluded that language proficiency can explain about 10% of the variance of academic outcome, which is frequently quoted as a benchmark level of predictive power of test scores. Feast (2002) reported 0.39 between IELTS scores and academic outcome. Yen and Kuzma (2009) reported that IELTS scores correlated significantly with the first semester academic outcomes at 0.46 and the second semester's outcomes at 0.25 (For details, see Section 2.5.3.2). It is important to notice that these studies studied the relationships between overall test scores and academic outcomes whereas this study investigated the predictive power of two single reading-into-writing tasks. Overall test scores are expected to have a better predictive power than individual task scores. Nevertheless, both reading-into-writing test tasks reported a better correlation with the academic outcome than the results of most of the above studies.

Most previous studies investigated the predictive power of the overall test scores rather than individual paper scores. Of the individual paper scores, writing test scores tend to have no or low correlations with academic success. For example, Cotton & Conrow (1998) and Humphreys et al. (2010) found no significant

correlation between the participants' IELTS writing scores (which include an independent writing task with input texts and an integrated writing task with non-verbal inputs) and their academic achievement. Kerstjen & Nery (2000), on the other hand, reported a significant correlation of 0.25 between their participants' IELTS writing test and academic scores (For details, see Section 2.5.3.2). The results of this study showed that both the reading-into-writing test tasks (Test Task A: essay task with multiple verbal inputs and Test Task B: essay task with multiple verbal and non-verbal inputs) reported a better correlation with the academic outcome than the figure reported in Kerstjen & Nery's (2000) study. Results of RQ1 and RQ2 of this study showed that the two reading-into-writing test tasks represented a range of salient features of real-life academic writing tasks in terms of overall task setting and input texts features, and elicited most of the eleven identified target real-life academic cognitive processes from test takers. Weir (2005) argued that the more valid the context and cognitive parameters a test task possesses, the more accurate the estimate of the test takers' performance in the real-life target conditions the test can provide. The results here showed that the two reading-into-writing test tasks had a better predictive power than the other independent writing-only or reading-into-writing tasks with only non-verbal inputs investigated in the literature.

The four analytical scores on Test Task A correlated with the mean real-life scores from 0.160 to 0.391 whereas the three analytical scores on Test Task B correlated with the mean real-life scores from 0.300 to 0.379 (See Table 6.14). It is encouraging that most categories on both test tasks correlated with the mean real-life scores at a level of 0.3 or above, except the categories of content (relevance and adequacy) and organisation (coherence and organisation) on Test Task A. Comparing the descriptors of both marking schemes, the descriptors of Test Task B appear to be more task specific than the descriptors of Test Task A (See Appendix 6.1 and 6.2). This could be because of the fact that the marking scheme of Test Task A was developed to assess Test Task A and another task (essay task with single non-verbal inputs) of the same test. In contrast, the marking scheme of Test Task B was developed specifically for Test Task B. The

results seem to suggest that a more task-specific marking scheme, especially for the categories of content and organisation, is beneficial. However, further evidence is needed to confirm this.

6.3.3 Patterns of the correlations between test scores and mean real-life scores

The correlations between the two reading-into-writing test scores and the real-life scores have been discussed. The last step was to examine the pattern of the correlations. The graphic representation of the correlations between the test scores and the mean real-life scores is presented in Figure 6.6 (Test Task A) and Figure 6.7 (Test Task B). The size and density of the plots indicate the strength of the correlation between the test scores and the real-life scores. The bigger and more concentrated the plots are, the stronger the correlation between the test scores and real-life scores is. The scatterplots in Figure 6.6 and Figure 6.7 are divided into five sections (Grade E to Grade A) by dotted lines. The relationships between the test scores and real-life scores are discussed with reference to the corresponding grade of the real-life scores.

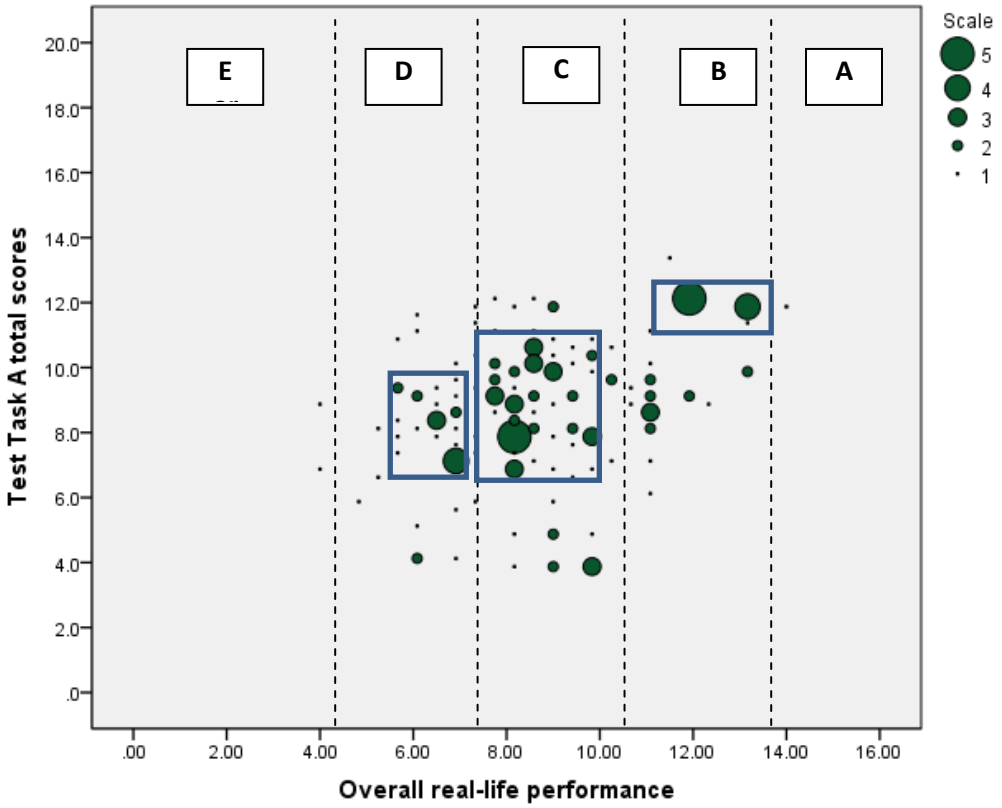


Figure 6.6 Relationships between Test Task A and real-life performance

As reported previously, the participants' Test Task A scores (n=160) correlated with their mean real-life scores at 0.306. According to Figure 6.6, the correlation appears to be strongest/clearest between Test Task A score 12 and real-life Grade B. The plots at the real-life Grade B cluster most densely at the score of 12 on Test Task A. In other words, most participants who obtained a total score of 12 on Test Task A achieved an overall Grade B on the selected real-life tasks.

The plots at the real-life Grade C spread densely along a wide range of Test Task A scores whereas the plots at the real-life Grade D cluster densely between Test Task A scores 7 and 9. It appears that for those participants who obtained an overall real-life Grade C, other factors might have had a bigger impact on their academic achievement than their writing ability. Criper & Davies (1998) argued that individual non-linguistic characteristics might have a stronger impact on one's academic success than language proficiency.

The results also showed that for the participants who scored between 7 and 9 on Test Task A, they were most likely to achieve an overall Grade D and Grade C on the real-life tasks. It is interesting for future study to explore why, with a similar academic writing ability, some of these participants (Grade C or even Grade B) achieved better in the target context than the others (Grade D).

As mentioned previously, the passing requirement of Test Task A, which is a level-specific test at C1 level, requires a minimum of band 3 on all 4 categories. Test Task A is part of the GEPT Advanced, a test which has been recognised by many universities as a means to indicate if applicants have reached the minimum of English language requirement. Nevertheless, the latest UKBA regulations require students who wish to apply to study at the level of bachelor's degree in the UK under Tier 4 (General) of points-based system to have a minimum of level B2 of the CEFR (UKBA, 2013). Results on Test Task A using a pass/fail dichotomy might no longer be entirely appropriate for university entry purposes. Although GEPT has a High-Intermediate test at B2 level, the test does not seem to target at the academic domain. It would seem more appropriate to adjust the score reporting method on the GEPT Advanced, which was developed in the academic domain, to indicate the test takers who have reached the minimum threshold of English requirement for higher education in the UK, i.e. B2 of CEFR. The results of this study showed that academic performance, especially at the levels of Grades D and B, could be predicted by different score ranges on Test Task A.

The patterns of correlation between Test Task B scores and real-life scores are discussed next (See Figure 6.7). As reported previously, the participants' Test Task B scores (n=140) correlated with their mean real-life scores at 0.379. As the range of the total analytical scores on Test Task B is comparatively limited, i.e. from 0 to 9, the patterns of plots are expected to be less clear than those obtained on Test Task A.

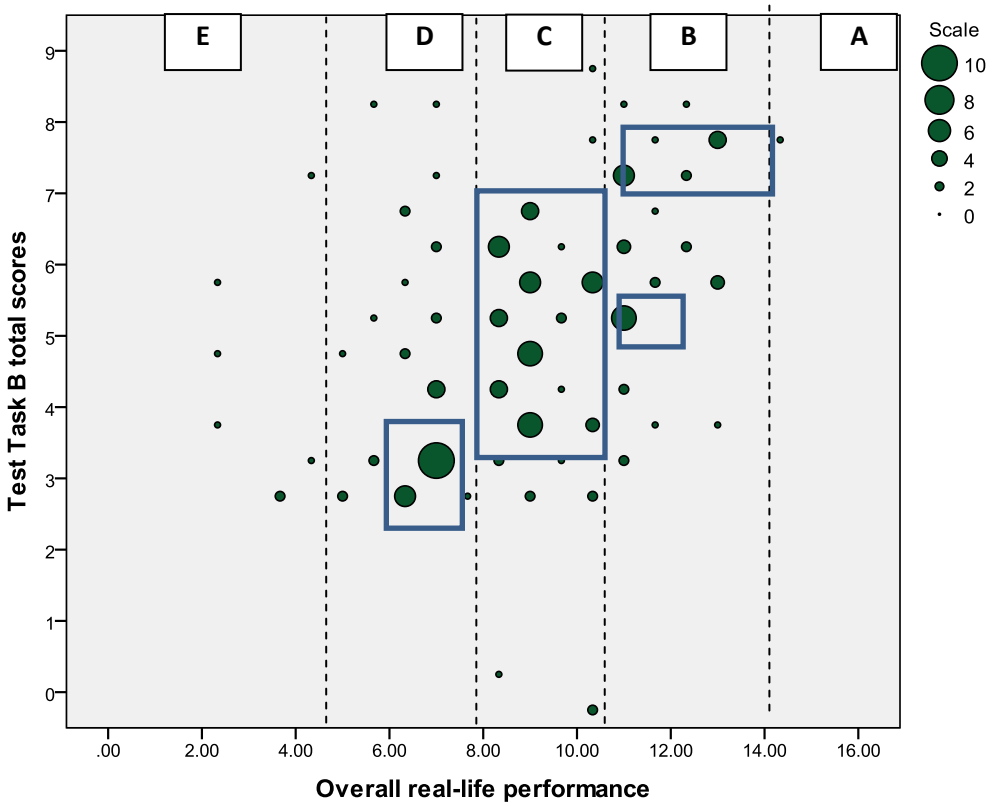


Figure 6.7 Relationships between Test Task B and real-life performance

According to Figure 6.7, the correlation pattern between Test Task B scores and real-life Grade D is the clearest. The plots at the real-life Grade D cluster most densely at the score of 3 on Test Task B. In other words, most participants who got a total score of 2.5 to 3.5 on Test Task B achieved an overall Grade D on the selected real-life tasks.

Like the pattern obtained on Test Task A, the plots at the real-life Grade C bunch densely along a wide range of Test Task B scores. Presumably other factors, such as subject knowledge and motivation, might have impacted the participants' academic achievement more than their academic writing ability. As mentioned previously, it is unfortunately beyond the scope of this study to investigate what other non-linguistic factors might have impacted on their performances on the selected real-life tasks.

Lastly, the plots at the real-life Grade B bunch most densely at the Test Task B score 5, and between 7 and 8. In other words, participants who scored 7 or 8

tended to achieve an overall Grade B on the real-life writing tasks. In addition, some participants who scored 5 on Test Task B obtained a better overall grade (Grade B) than the others (Grade C). It would be interesting in future studies to explore the reasons.

Test Task B was designed to indicate the test takers' need for academic writing support. As mentioned previously, Grade A (a score of 8 and 9) indicates no need for academic writing support, Grade B (a score of 6 and 7) indicates a low-level need for academic writing intervention, whereas Grade C (a score of 5 or below) indicates a high-level need for academic writing intervention. Considering the variable of academic writing proficiency solely, the results of this study seem to suggest that the cut-off scores of Test Task B which indicate different levels of academic writing support required might not be the most effective. Further evidence is needed to confirm the results and to investigate the most appropriate cut-off scores for the need of different levels of academic writing support.

6.3.4 Summary

The major results of the correlations between test scores and real-life scores are summarised as follows:

In terms of individual real-life tasks, both Test Task A and Test Task B scores correlated significantly with individual writing tasks (report, essay and question-and-answer test) which were assigned during the term, but not with the end-of-term case-study examination scores. The reasons could be because a) the time between the language test event and the examination event was comparatively longest, and b) the demand for subject knowledge imposed by the examination task was comparatively highest.

Both the two types of reading-into-writing tests (Test Task A with multiple verbal inputs and Test Task B with multiple verbal and non-verbal inputs) show significant moderate correlations to the participants' mean real-life academic scores at r=0.306 and r=0.369 respectively. When compared to similar studies in the relevant literature which investigated the predictive power of writing test

scores, most of which involved an independent essay writing task and an integrated reading-into-writing task with single non-verbal input, the results of the correlations between the two reading-into-writing test scores and real-life outcome reported in this study are encouraging (cf Cho & Bridgeman, 2012; Cotton & Conrow, 1998; Davies & Criper, 1988; Dooey, 1999; Feast, 2002; Humpreys et al., 2010; Ingram & Bayliss, 2007; Kerstjens & Nery, 2000; Yen & Kuzma, 2009).

In terms of the effectiveness of predictive power, the reading-into-writing test scores seem to be able to 'predict' performance in the target context better at high and low levels than at the medium level. In other words, participants who achieved comparatively well (Grade B) on real-life tasks tended to have scored comparatively high on the reading-into-writing test scores. For participants who achieved comparatively poorly (Grade D) on real-life tasks tended to have scored comparatively poorly on the reading-into-writing test scores. However, for participants who achieved at the medium level (Grade C), their scores on the reading-into-writing test tasks ranged widely. It appears that academic writing ability might have limited impact on the medium-level academic achievement in the context of this study. Therefore, any high-stakes decisions for these medium-level test takers need to be made with extra caution, and supported by other forms of evidence.

The results showed that the reading-into-writing task type has not only valid contextual validity and cognitive validity, as demonstrated in Chapters Four and Five respectively, but also has moderate predictive power for test takers' real-life academic writing performance. This study has provided empirical evidence of the three major components of validity in using reading-into-writing test tasks to assess academic writing ability.

The results of the study presented in Chapters Four, Five and Six are further discussed in Chapter Seven to identify implications for test development of valid reading-into-writing tests to assess academic writing ability, test validation of such integrated task type and directions for further research.

7 CONCLUSIONS AND LIMITATIONS

7.1 Introduction

Regarding EAP academic writing tests, integrated reading-into-writing tasks are considered to have better context and cognitive validity than independent writing-only tasks (Carson, 2001; Feak & Dobson, 1996; Hamp-Lyons & Kroll, 1997b; Johns, 1993; Plakans, 2009a, 2010; Weigle, 2004). However, when compared to the independent writing-only task type, there is insufficient empirical evidence of the context and cognitive validity of the integrated reading-into-writing task type in the literature to date (Plakans, 2008, 2010; Segev-Miller, 2007; Weigle, 2004). This study, building on Weir's (2005) socio-cognitive framework for test validation, developed a new framework for validating reading-into-writing tests to assess academic writing ability. Building on Khalifa & Weir's (2009) framework of reading tests and Shaw & Weir's (2007) framework of writing tests, this study aimed to establish an empirical framework for the development and validation of EAP reading-into-writing test tasks, and to identify parameters that are useful to explicitly describe the context and cognitive validity of such reading-into-writing test tasks. In addition, the study aimed to investigate the extent to which reading-into-writing test scores predict real-life academic writing performance. The three research questions that this study set out to answer were:

1. What are the contextual characteristics of the academic writing tasks that students would normally encounter in real life? To what extent do the reading-into-writing test tasks resemble these contextual features under test conditions?

2. What are the cognitive processes that students would normally employ to complete the real-life academic writing tasks? To what extent do the reading-into-writing test tasks elicit these cognitive processes from test takers?
3. To what extent can performances on the reading-into-writing tests predict test takers' ability to perform on real-life academic writing tasks?

In order to answer these research questions, this study investigated the target construct of academic writing ability involving integration of reading materials by investigating the contextual parameters of two real-life academic writing tasks and the cognitive parameters activated by these real-life tasks. After defining the target construct, the study investigated the contextual parameters of two reading-into-writing test task types and the cognitive parameters activated by these reading-into-writing test tasks. The findings revealed the extent to which the real life contextual and cognitive parameters were represented by the two reading-into-writing task types. Lastly, the study investigated the correlations between the reading-into-writing test scores and real-life academic writing performance to shed light on the criterion-related validity of reading-into-writing test tasks.

The results of RQ1, RQ2, and RQ3 of this study are summarised in Sections 7.2.1, 7.2.2 and 7.2.3 respectively. The limitations of this study and areas for future research are discussed in Section 7.3 in terms of sampling (Section 7.3.1) and research instruments (Section 7.3.2). Section 7.4 discusses the implications and contributions of this study (Section 7.4.1 - 7.4.5).

7.2 Conclusions concerning the validity of EAP reading-into-writing tests

The results of this study showed that the two reading-into-writing task types (*essay task with multiple verbal inputs* and *essay task with multiple verbal and non-verbal inputs*) have promising context validity, cognitive validity and criterion-related validity. This was supported by validity evidence collected based

on three major components of the Weir's (2005) socio-cognitive validity framework:

(1) context validity (reported in Chapter Four): analysis of the contextual parameters of the two real-life academic writing tasks and the reading-into-writing test tasks;

(2) cognitive validity (reported in Chapter Five): analysis of the cognitive processing activated by the two real-life academic writing tasks and the reading-into-writing test tasks; and

(3) criterion-related validity (reported in Chapter Six): correlational analysis between the two reading-into-writing test scores and the real-life academic performance.

7.2.1 Context validity of EAP reading-into-writing test tasks

RQ1: What are the contextual characteristics of the academic writing tasks that students would normally encounter in real life? To what extent do the reading-into-writing test tasks resemble these contextual features under test conditions?

To answer RQ1, the study built on the procedures used in previous studies (e.g. Green et al, 2010; Green et al, 2012; Weir, 2012; Wu 2012) to analyse the overall task setting and the input text features of two real-life academic writing tasks and two reading-into-writing test tasks by expert judgment and automated textual analysis. Context validity evidence collected from expert judgment and automated textual analysis showed that the two reading-into-writing tasks resembled the contextual features of the real-life academic writing tasks in many important ways. The results, on the other hand, reveal some areas that may require improvement. Recommendations on task design are provided in Section 7.4.2.

Overall task setting

The results of the small-scale task survey in this study, as presented in Section 3.2.1, showed that essays and reports are commonly assigned to students in the

academic context of this study. Therefore, a *report* task and an *essay* task were selected from two different modules to represent the predominant real-life academic writing tasks in this study. Both reading-into-writing test tasks (Test Task A and Test Task B) required the production of an essay. While essay tasks are commonly used in standardised academic writing tests, typical writing-only essay test tasks may represent the target genre but neglect some essential contextual features of real-life academic essay tasks, especially in terms of the use of external reading sources and language functions (Moore & Morton, 2005). The results of this study showed that both reading-into-writing test tasks resembled the contextual features of the real-life academic writing tasks in many significant ways, which are described below.

In terms of **topic domains**, the *academic* and *professional* domains were dominant in the selected real-life tasks. Test Task A was considered to be in the *academic* and *social* domains while Test Task B fell into the *professional* and *social* domains. Regarding the **cognitive demands** imposed on the writer, the real-life tasks were knowledge-transforming tasks which required high-level processes. Both reading-into-writing tasks required the test takers to transform the ideas by selecting, organising and summarising relevant ideas from the input sources as well as evaluating different points of view. According to the judges, the test tasks did not require test takers to interpret, evaluate, and apply ideas in context to the extent that the real-life tasks did. However, they considered the level of cognitive demands of the test-tasks to be appropriate under the test conditions. Both real-life task required students to perform a range of **language functions**. Core language functions, those that were judged by 2 or more pairs of the judges, required by the two real-life tasks included *describing, defining, reasoning, citing sources, evaluating, synthesising* and *expressing personal views.* In addition, the report task required the functions of *illustrating visuals, predicting* and *recommending,* whereas the essay task required the functions of *persuading* and *summarising*. A majority of these language functions identified as necessary for the completion of the real-life academic writing tasks were also required by the two reading-into-writing test tasks. As judged by 2 or more pairs

of the judges, Test Task A required test takers to perform *expressing personal views*, *summarising*, *citing sources*, *evaluating*, *recommending*, *reasoning*, *synthesising* and *describing*. Test Task B required test takers to perform, mostly necessarily, *reasoning*, *summarising*, *expressing personal viewpoints*, *evaluating*, *recommending*, *synthesising* and *illustrating visuals.*

Regarding the **clarity of purpose**, **clarity of audience** and **clarity of marking criteria**, the real-life tasks were rated towards the positive end of a five-point scale (1=unclear; 5=clear), though with better ratings on the clarity of purpose and clarity of audience. Shaw & Weir (2007: 71) argued strongly for the importance of providing clear and unambiguous information on the task purpose, intended audience and marking criteria in any valid writing test. It is important to consider the fact that students in real-life academic contexts are given plenty of opportunities to clarify any unclear information whereas test takers do not usually have the same opportunities under test conditions. The results showed that the two reading-into-writing test tasks functioned well with regards to clarity of task purpose, intended audience and marking scheme. Both Test Task A and Test Task B received higher ratings for clarity of task purpose and clarity of audience than the real-life tasks. The judges considered that the real-life academic writing tasks provided very detailed criteria (the report task scoring 4.5 and the essay task scoring 4.6 out of 5.0). The rating of the clarity of marking criteria for Test Task A (3.8) was slightly lower than the real-life tasks' ratings, whereas Test Task B scored 4.6.

Input text features

Previous studies showed that most academic writing tasks involve integration of external reading materials (Bridgeman & Carlson, 1983; Carson, 2001; Grabe, 2003; Horowitz, 1986a, 1986b; Johns, 1993; Weir, 1983). Results of this study shared the same finding that the two real-life tasks required students to write based upon multiple reading resources of verbal and non-verbal materials, though integration of non-verbal materials seems to be more essential for the completion of the report task than the essay task. The results also revealed that the real-life

input texts contained a variety of genres, such as news / magazine articles, journal articles and book chapters, and different non-verbal materials, such as graphs, tables, pictures and diagrams.

One of the most heavily criticised aspects of the use of writing-only tasks to assess academic writing ability is that such task types do not engage test takers with the use of reading materials and the corresponding cognitive processes necessary to integrate external resources into written production (Carson, 2001; Hamp-Lyons & Kroll, 1997a; Johns, 1993; Johns & Mayes, 1990; Plakans, 2008, 2010; Weigle, 2002). The results of this study showed that Test Task A resembled the real-life academic performance conditions by requiring the test takers to write based upon two essays, whereas Test Task B resembled the real-life performance conditions by requiring the test takers to write based upon one report and one news article, each passage containing a diagram.

The difficulty level of a writing task can also be determined by the **discourse mode** of the input texts (Parodi, 2007). For example, when compared to an argumentative text, a narrative text tends to contain more factual information, and hence is usually less demanding to read. The results showed that the real-life input texts contained a combination of *expository* and *argumentative/evaluative* texts. The reading-into-writing test tasks contained either *expository* or *argumentative/evaluative* texts. Test Task A required test takers to process argumentative texts while Test Task B required test takers to process expository texts.

In addition, based on the ratings of a five-point scale on the **concreteness of ideas** (1=abstract; 5=concrete), the **explicitness of textual organisation** (1=inexplicit; 5=explicit), and the **cultural specificity** of the content (1=culturally neutral; 5=culturally specific), content in the reading-into-writing test input texts was regarded to be noticeably more concrete and more explicitly organised than the content of the real-life input texts. Test Task A input texts were more culturally specific than the others, but this was considered by the judges as appropriate since

Test Task A is primarily targeted at a test population with a homogenous cultural background.

Linguistic complexity of input texts

This study built on the automated textual analysis procedures employed in recent studies (e.g. Green et al, 2010; Green et al, 2012; Weir, 2012; Weir et al, 2013, Appendix 2; Wu, 2012) in the language testing literature to analyse the lexical complexity, syntactic complexity and degree of cohesion of the real-life input texts and the reading-into-writing test tasks.

The results, as reported in Section 4.3.2.1, revealed very specific information about the level of lexical complexity, syntactic complexity and degree of cohesion of the real-life input texts in terms of a set of carefully selected automated textual analysis indices (See Section 3.3.2.2 for the procedures). These indices on the real-life input texts were compared descriptively with those obtained in Green et al's (2010) study on the textual features of undergraduate reading texts to better understand the level of the real-life input texts. Generally speaking, the difficulty level of the input texts undergraduates used to complete their writing assignments was close to the undergraduate course book texts. However, the results suggested three areas of discrepancy in the linguistic features between the real-life input texts and undergraduate course book texts:

(1) The real-life input texts contained more frequent words (the first 1000 and the first 2000) and less low frequency words than the undergraduate texts.

(2) The real-life input texts contained slightly more *words before the main verbs of the main clauses* than the undergraduate texts, and a slightly lower density of connectives (*logical operator incidence score*) than the undergraduate texts.

(3) Sentences in the real-life input texts seemed to be less conceptually similar to the next sentence than those in the undergraduate course book texts.

Based on descriptive statistics, the findings suggested some distinctive features of the texts which were used for general study purposes and those for writing

purposes in real-life academic contexts. Real-life input texts for academic writing were slightly more challenging in syntactic features but less challenging in word frequency than the undergraduate course book texts. While further evidence is needed to confirm the results, the findings might provide insights for test developers when they develop input texts in reading-into-writing tests and those in reading comprehension tests.

The study then compared the textual indices of the real-life input texts with those of the two reading-into-writing test task input texts.

The difficulty level of sampled Test Task A input texts was comparable to the level of the real-life input texts, in terms of lexical complexity (as indicated by the proportion of first 1000 frequency words, proportion of first 2000 frequency words, proportion of academic words, frequency level of content words, average syllables per word and type-token ratio of all content words), syntactic complexity (as indicated by the average sentence length, mean number of words before the main verb in verb phrases and proportion of logical operators), and degree of text cohesion (as indicated by the percentage of adjacent sentences with one or more repeated arguments, word stems, content words, proportion of adjacent anaphor references, and adjacent semantic similarity). Still a few indices suggested that Test Task A input texts were more challenging than the real-life input texts in terms of lexical complexity due to a higher proportion of low frequency words, but less challenging than the real-life input texts in terms of syntactic complexity due to a higher similarity of sentence structures and a lower average number of modifiers per noun phrase in the texts. Recommendations are further discussed in Section 7.4.2.

Due to the small number of testlets available for Test Task B, only descriptive statistical analysis was performed on the textual indices between Test Task B and real-life input texts. Results suggested that the Test Task B input texts appeared to be more challenging than the real-life input texts in terms of most lexical features and syntactic features, but were more cohesive than the real-life input texts In terms of the lexical complexity, even though Test Task B input texts contained a

slightly higher proportion of high frequency words than the real-life input texts, Test Task B input texts had a higher proportion of academic words and low frequency words and a higher proportion of unique content words (type-token ratio). In terms of syntactic complexity, it seemed to be more demanding to process the noun-phrases and to sort out the logical connections between ideas in the Test Task B input texts than the real-life input texts. Nevertheless, the results of the text cohesion indices indicated that it would be easier to process the main themes in Test Task B input texts than in the real-life input texts.

Many researchers argued that, as far as academic writing is concerned, the integrated reading-into-writing test type is perhaps the most valid task type to simulate real-life writing conditions (Hamp-Lyons & Kroll, 1996; L Plakans, 2008; Weigle, 2004; Weir, et al., 2013). The results of RQ1 reported in this study showed that the two reading-into-writing test tasks had good contextual validity as they largely resembled the real-life academic writing tasks in terms of the overall task setting and input text features.

7.2.2 Cognitive validity of EAP reading-into-writing test tasks

RQ2: What are the cognitive processes that students would normally employ to complete the real-life academic writing tasks? To what extent do the reading-into-writing test tasks elicit these cognitive processes from test takers?

To answer RQ2, this study investigated the cognitive processes employed by 219 participants on two real-life academic writing tasks and two reading-into-writing test tasks through a carefully developed and validated Writing Process Questionnaire (See Section 3.4.1 for the procedures of developing the questionnaire). A total of 443 questionnaires were collected regarding real-life and test conditions in the study - 70 questionnaires on the real-life essay task, 73 on the real-life report task, 160 on the reading-into-writing Test Task A (essay task with multiple verbal inputs), and 140 on the reading-into-writing Test Task B (essay task with multiple verbal and non-verbal inputs).

Results collected from the two real-life academic writing tasks provided empirical evidence of the target cognitive processes to be measured in an academic writing test in terms of explicit cognitive parameters. Results collected from the two reading-into-writing test tasks revealed the extent to which the integrated task type activated these cognitive processes in the same manner as they were activated by the real-life academic writing tasks. Overall, both reading-into-writing test tasks demonstrated good cognitive validity. The results are summarised below.

The underlying structure of the cognitive processes of the five academic writing phases elicited by the real-life tasks

Based upon the literature review, writers are likely to go through several cognitive phases when they write from external sources, though the phases can be overlapping or looping back. The study considers the following five phases to be most relevant to the discussion of academic writing tests: (1) *conceptualisation*, (2) *meaning and discourse construction*, (3) *organising*, (4) *low-level monitoring and revising* and (5) *high-level monitoring and revising* (Field, 2004, 2008, 2011, 2013; Kellogg, 1994; Shaw & Weir, 2007). Considering the constraints of the questionnaire, this study focused broadly on the phases which are more metacognitive (i.e. easier to be self-reported) and did not investigate phases such as execution and micro-planning. Although there is a rich body of research on the cognitive processes involved in each of these five cognitive phases (e.g. Flower et al, 1990; Hayes, 1996; Hayes & Flower, 1983; Kellogg, 1994, 1996; Khalifa & Weir, 2009; Shaw & Weir, 2007; Spivey, 1990, 1991, 1997; Spivey & King, 1989), we still lack a comprehensive model which accounts for the processes involved in writing with the use of reading sources, especially in the L2 contexts (Hirvela, 2004). Building upon these studies, this study investigated the

underlying structure of the cognitive processes activated within each cognitive phase by two real-life academic writing tasks by exploratory factor analysis.

The results showed that the hypothesised academic writing phases arising from the literature review were largely supported by the statistical analysis of the questionnaire data collected in this study. Each academic writing phase activated under the real-life academic conditions involved two or more distinct yet correlated underlying cognitive processes. The **conceptualisation** phase involved the (1) task representation and macro-planning processes to conceptualise an understanding of the writing task and establish their macro plans. In addition, the process of (2) revising macro plan was particularly important in real-life academic writing. The **meaning and discourse construction** phase involved three underlying processes: (3) connecting and generating, (4) selecting relevant ideas and (5) global careful reading. The **organising** phase involved the processes of (6) organising ideas in relation to the input texts as well as (7) organising ideas in relation to the writer's own text. For the **low-level monitoring and revising** phase, there was a clear distinction between the (8) low-level editing processes employed while writing and the (9) low-level editing employed after the first draft has been completed. Similarly, the **high-level monitoring and revising** phase involved (10) while writing high-level editing and (11) after writing high-level editing. This study considers that these eleven cognitive processes involved within the five academic writing phases would be appropriate as the target cognitive processes for a valid academic writing test.

Comparisons of the processes employed by the high-achieving and low-achieving participants

Having identified the target cognitive processes, the study investigated whether the high-achieving and low-achieving participants employed these processes differently on the real-life academic writing tasks. The findings would indicate if the target cognitive parameters identified could potentially distinguish the performances of stronger writers from those of weaker writers. Shaw & Weir (2007) argued that when identifying the cognitive parameters to be examined in a

test, it is important to demonstrate 'how writers at different levels would employ these cognitive processes with 'educationally significant differences' (p.142). The results showed that the high achieving participants reported employing eight of the eleven cognitive processes (i.e. *task representation and macro-planning, careful global reading, selecting relevant ideas, connect and generate, organising ideas in relation to source texts, organising in relation to new text, low-level editing while writing* and *high-level editing while writing*) more than the low achieving groups. Apart from *task representation and macro-planning* and *careful global reading,* all differences were significant. Interestingly, three process parameters including *revising macro plan, low-level editing after writing* and *high-level editing after writing* did not distinguish the high-achieving and low-achieving groups. Generally speaking, such findings that the high-achieving participants employed most of these processes more than the low-achieving participants on the real-life tasks add further support to the case for considering these eleven cognitive processes involved within the five academic writing phases as the target cognitive process parameters for a valid academic writing test.

The study then investigated whether these cognitive parameters could also distinguish processes employed by the stronger test takers from those by the weaker test takers on the two reading-into-writing test tasks. The results showed that, on Test Task A, the high-achieving participants reported employing six of the processes (which included *revising macro plan, selecting relevant ideas, organising ideas in relation to own text, low-level editing while writing, low-level editing after writing,* and *high-level editing after writing*) more than the low-achieving group. However, all differences obtained were non significant ($p>0.05$). Regarding Test Task B, the high-achieving participants reported employing nine of the cognitive processes more than the low-achieving group (See Table 5.20). The differences of seven processes (which included *task representation and macro-planning, selecting relevant ideas, organising ideas in relation to source texts, low-level editing while writing, low-level editing after writing, high-level editing while writing* and *high-level editing after writing*) between the two groups were significant ($p<0.05$).

The results showed that the majority of the eleven cognitive parameters were able to distinguish the processes reported by the high-achieving and low-achieving writers on the real-life tasks and Test Task B. However, high-achieving and low-achieving writers did not report using these eleven cognitive parameters with significant differences. It should be noted that, as mentioned in Chapter Three, Test Task A was designed to indicate whether the test takers have reached a particular level, i.e. C1 in this context. Most (87.2%) of the participants in this study were at B2 level, as indicated by their IELTS scores. The profile of the participants might have limited the effectiveness of these eleven cognitive parameters distinguishing the processes reported by high-achieving and low-achieving writers on Test Task A.

Comparisons of the cognitive processes elicited under test and real-life conditions

Any writing test tasks, whether independent or integrated, which are cognitively valid should elicit from test takers the cognitive processes which they would normally employ in non-test conditions. In addition, Shaw & Weir (2007: 142) emphasised that valid cognitive parameters identified should also be able to demonstrate how writers at different levels would employ them differently. Therefore, this study examined the extent to which the eleven proposed cognitive parameters for reading-into-writing test tasks for academic purposes sufficiently resembled the cognitive processes which the test takers (as a whole group as well as in groups of high-, medium, or low- achievement) in non-test conditions.

The participants as a whole group reported employing all the eleven cognitive processes more on a scale from 1 (*definitely disagree*) to 4 (*definitely agree*) on the real-life tasks than on Test Task A. The differences reported in six processes were significant ($p<0.05$). Regarding Test Task B and the real-life tasks, the participants reported employing eight cognitive processes more in the real-life

conditions than the test conditions. The differences reported in six processes were significant ($p<0.05$). The results indicated that the participants employed most processes more on the real-life tasks than on the test tasks. While interpreting these findings, it is important to consider the fact that in real life contexts, students have considerably more time to utilise the cognitive processes and iteratively revisit their work. In contrast, few tests give dedicated time for either planning or monitoring. Most test tasks are performed under great time pressures. In addition, the real-life tasks were mainly produced by the computer whereas both test tasks used in this study were paper-based. The results implied that the fundamental discrepancies between the real-life and test conditions, such as time allowance and mode of writing, seem to have an impact on how writers employed these eleven processes.

The analysis which took level of performance into consideration showed that both reading-into-writing test tasks were able to elicit from high-achieving and low-achieving participants most of the cognitive processes to a similar extent as participants employed the processes on the real-life tasks. However, the middle group showed greater discrepancy in how they employed the processes under the test and real-life conditions. They tended to employ some processes more in the real-life conditions than the test conditions. They employed the processes of *conceptualisation, low-level monitoring and revising, and high-level monitoring and revising phases significantly less* on Test Task A than on real-life tasks. They employed the processes of *connecting and generating*, *low-level editing after writing* and *high-level editing after writing* significantly less on Test Task B than on real-life tasks. When compared to the high-achieving group, the middle-achieving group might not be able to employ all processes with full automaticity as they were at the transitional stage of developing their academic writing ability. Due to limited cognitive capacity, they may need more time to complete the processes. It would be essential in future research to investigate why, under test conditions, the middle-achieving participants employed some of the real life processes but not others on these two types of reading-into-writing tasks. Recommendations for further research will be discussed in detail in Section 7.3.

The underlying structure of the cognitive processes of the five academic writing phases elicited by the two reading-into-writing test tasks

The results showed that both Test Task A and Test Task B were largely able to elicit from the participants the same underlying cognitive processes as the real-life tasks did. The underlying structures of four out of the five phases of academic writing, which included **conceptualisation**, **organising**, **low-level organising and revising**, and **high-level monitoring and revising**, elicited on Test Task A and the real-life tasks were identical. And seven factors within these phases, which included *task representation and macro-planning, revising macro plan, selecting relevant ideas, organising ideas in relation to input texts, low-level editing after writing, low-level editing during writing* and *high-level editing after writing,* elicited by Test Task A contained the same individual questionnaire items as the corresponding factors identified by the real-life tasks.

On the other hand, the underlying structures of the cognitive processes of four phases of academic writing, which included **discourse and meaning construction**, **organising**, **low-level organising and revising**, and **high-level monitoring and revising**, elicited on Test Task B and the real-life tasks were identical, though the order of the factors of the discourse and meaning construction was different between Test Task B and the real-life tasks. Eight factors, which included *revising macro plan, selecting relevant ideas, careful global reading, organising ideas in relation to input texts, low-level editing after writing, low-level editing during writing, high-level editing after writing* and *high-level editing during writing,* elicited by Test Task B contained the same individual questionnaire items as the corresponding factors identified by the real-life tasks.

In short, the results of RQ2 provided strong evidence supporting the cognitive validity of the two integrated reading-into-writing test tasks. This particular integrated test type was able to elicit certain real-life processes, such as *task representation*, *selecting relevant ideas, careful global reading,* and *organising ideas in relation to source texts*, which might be elicited rather differently, if at all, in a standard writing only test.

7.2.3 Criterion-related validity of EAP reading-into-writing test tasks

RQ3: To what extent can performances on the reading-into-writing tests predict test takers' ability to perform on real-life academic writing tasks?

To answer RQ3, the study investigated the correlations between Test Task A (essay task with multiple verbal inputs) scores and real-life academic performance, and between Test Task B (essay task with multiple verbal and non-verbal inputs) scores and real-life academic performance. In this study, academic performance was measured by 2 writing tasks, 1 in-class question-and-answer test and 1 end-of-term case study examination. Indicated by the participants' IELTS reading and writing bands, 12.8% of them were at CEFR C1 level (IELTS: 6.5-7.0) and 87.2% were at B2 level (IELTS: 5.0-6.0).

This sub-section summarises the results of 1) the participants' performance on Test Task A, Test Task B, and the real-life tasks, 2) correlations between test scores and real-life academic performance, and the pattern of the correlations.

Participants' performance

160 participants completed Test Task A, which is part of a criterion-referenced test at the C1 level. 10.5% (n=17) passed the task. The low pass rate on Test Task A was expected because, as indicated by IELTS reading and writing bands, only about 12 % of the participants were at the C1 level. Regarding individual marking criteria, more participants achieved Band 3, i.e. the pass band on the categories *relevance and adequacy* (30%) and *coherence and organisation* (31.2%) than on the categories of *lexical use* (19%) and *grammatical use* (15%).

140 participants completed Test Task B, which is part of a University diagnostic test of academic English ability at the B2 level. 8.6% scored a total of 8 or 9 out of 9, which indicates no need for intervention in their academic writing ability. 32.8% of the participants scored 6 or 7, which indicates a need for low-level intervention whereas 58.6% of the participants scored 5 or below, which indicates a need for high-level intervention. In terms of individual criteria, 8.6%, 12.9%

and 10.7% of participants scored band 3 (which indicates an adequate or above adequate performance) on the categories of *content*, *organisation* and *language* respectively. Most participants were scored band 2 (which indicates a less than adequate performance) on *content* and *organisation* but most participants were scored band 1 (which indicates a significantly weak performance) on *language.*

Four points of reference (i.e. 2 writing tasks, 1 in-class question-and-answer test and 1 end-of-term case study examination) were selected for the analysis of the criterion-related validity in this study. The 219 participants, as a whole group, performed best on the report task (mean: 9.72 out of 16) while scoring lowest on the exam (mean: 6.03 out of 16). 0.5% of the participants in this study got an average Grade A on these four tasks. 21% of the participants got an average Grade B and about 30% got Grade C. Most participants (45.2%) got an average Grade D and 3.7% got an average Grade E (Fail).

Correlations between reading-into-writing test scores and real-life writing task scores

Test Task A scores correlated significantly with the mean real-life scores at a moderate level of $r=0.31$ ($p<0.001$), explaining 9.36% variance of the real-life performances. Regarding Test Task A scores and individual real-life scores, the correlations between the two ranged from 0.126 to 0.343. Test Task A scores correlated moderately with report scores at $r=.343$ ($p<0.001$), weakly with question-and-answer test scores at $r=0.212$ ($p=0.027$) and essay scores at $r=0.187$ ($p=0.050$). Test Task A scores did not correlate significantly with the end-of-term examination scores.

Test Task B scores correlated significantly with their mean real-life scores at a moderate level of $r=0.38$ ($p<0.001$), explaining 14.36% variance of the real-life performances. The correlations between Test Task B and individual real-life writing tasks ranged from 0.082 to 0.438. Test Task B scores correlated moderately with question-and-answer test scores ($r=0.438$, $p<0.001$) and with essay scores ($r=0.386$, $p<0.001$), and weakly with report scores at $r=0.283$

(p=0.005). Test Task B scores correlated weakly with case study examination at r=0.082 but the correlation was non significant.

The correlations between Test Task A scores and academic performance, and between Test Task B scores and academic performance reported in this study are apparently better than most previously reported figures in the literature. Davies & Criper (1988) reported a correlation of 0.3 between overall language test scores and academic outcomes. They concluded that language proficiency can explain about 10% of the variance in academic outcomes, which is frequently quoted as a benchmark level of predictive power of test scores. Regarding the predictive power of academic writing test scores, Cotton & Conrow (1998) and Kerstjen & Nery (2000) found no significant correlation between the participants' IELTS writing scores and their academic achievement. Kerstjen & Nery (2000) reported a correlation of 0.25 between their participants' IELTS writing test and academic scores. This indicates that reading-into-writing test tasks have promising predictive validity.

The patterns of the correlations between the two reading-into-writing test scores and academic performance were very similar. The two reading-into-writing test scores seemed to be able to 'predict' performance in the target context better at high and low levels than at the medium-level. There were clear correlations between Test Task A score 12 and real-life Grade B, and between Test Task A scores 7 to 9 and real-life the plots at the real-life Grade D. On the other hand, there were clear correlations between Test Task B score 7 to 8 and real-life Grade B, and between Test Task B score 3 and the real-life Grade D. The results showed that participants who achieved comparatively well (Grade B) on real-life tasks tended to have scored comparatively high on the reading-into-writing test scores. For participants who achieved comparatively poorly (Grade D) on real-life tasks tended to have scored comparatively poorly on the reading-into-writing test scores. However, for participants who achieved at the medium-level (Grade C), their scores on the reading-into-writing test tasks ranged widely.

In addition to the comparatively high correlations between reading-into-writing test scores and academic outcome, the reading-into-writing test scores seemed to be able to predict academic outcome at the comparatively high and low levels, i.e. Grade B and Grade D in this study. Nevertheless, the results showed that academic writing ability might have limited impact on medium-level academic performance.

7.3 Limitations of the study and areas for future research

The generalisability of the results obtained in this study on 1) the target parameters of the real-life academic writing tasks, and 2) the validation evidence of the two types of reading-into-writing test tasks is limited. There are several limitations which need to be considered when interpreting the results of this study. These limitations should be addressed in future research.

7.3.1 Sampling

Participants of the investigation of cognitive validity

A total of 219 participants participated in this study. As indicated by the participants' IELTS reading and writing bands, 12.8% of the participants were at C1 level whereas the majority of the participants (87.2%) were at B2 level. While the proficiency level of the participants was considered appropriate for the context of this study, future studies are advised to include more participants at C1 and C2 levels.

In addition, test takers' background was homogenous in the sense that all test takers were from the Business School in the UK academic context. As argued in Section 3.4.1, a comparatively homogenous profile of background knowledge was suitable for the cognitive investigation in this study. Nevertheless, future studies are advised to investigate the cognitive processes and test results of participants from a variety of disciplinary backgrounds and/or from different national contexts.

Real-life tasks and reading-into-writing test tasks

This study used a small sample of real-life academic writing tasks and reading-into-writing test tasks. Two real-life academic writing tasks and two reading-into-writing test tasks were selected based on a range of carefully chosen criteria (see Section 3.2). Due to the limited scope of the study, for the cognitive and contextual investigation, only one version of each real-life task and one testlet of each test task were analysed. For the investigation of the predictive validity, two additional points of reference, i.e. in-class test and end-of-term examination, were added to the real-life tasks. For future studies, an analysis of multiple testlets would be recommended, especially when only only component of the validity is concerned in the study. As explained in Chapter Three, the test scores and real-life academic writing scores reported in this study were achieved by standardised scoring procedures administrated by the test providers and the University. Therefore, the results of the predictive validity reported in Chapter Six need to be interpreted in light of the appropriateness of the scoring procedures used.

For the investigation of the input text features, the ten most cited input texts were selected for each real-life academic writing task for analysis (See Section 3.2.1). Twenty input texts from ten testlets of Test Task A were analysed. However, at the time of the study, limited test papers of Test Task B were available for analysis. Only two input texts from Test Task B were analysed. Therefore, the results regarding the contextual features of Test Task B should be interpreted as indicative.

In addition, this study investigated the validity of two types of reading-into-writing test tasks of academic purposes (i.e. *essay with multiple verbal inputs* and *essay with multiple verbal and non-verbal inputs*). Future studies are advised to investigate other reading-into-writing task types, giving priority to those which have not received much attention in the literature, for instance, tasks which require test takers to process multiple reading texts for both the reading comprehension section and writing section.

7.3.2 Research instruments

Expert judgement and automatic textual analysis

As discussed by different researchers in the literature, the methods of expert judgement and automatic textual analysis to investigate task features have their own limitations (See for example Green, et al, 2012; Weir, 2012; Wu, 2012). Therefore, it is not advisable to use either one of them alone.

In this study, the procedures for expert judgement of the contextual features of the reading-into-writing tasks were refined based on the experience of the two pilots (see Section 3.3.2.1). First of all, it is necessary to clarify which part of the tasks the analytic categories in the proforma should be applied to, e.g. the prompt, the input texts, and/or the output. Besides, it appeared that some categories, e.g. cognitive demands, topic domains, concreteness of ideas, required subjective judgment. In order to minimise irregularity, this study required the judges to complete the judgment individually and then discuss their results in pairs (for the procedures, see Section 3.3.3). Results arising from the pair discussions instead of individual judgments were reported. In addition, in the context of this study where a set of input texts needed to be analysed for each task, it was more effective to analyse the overall task setting and the features of input texts in two separate expert judgement meetings. The analysis of the linguistic complexity of the input texts, especially for lexical and syntactic complexity, should be supplemented by automated textual analysis tools. As demonstrated in the study, tools such as Coh-Metrix and VocabProfile could usefully supplement the expert judgment in comparing features of the test task input texts with those of real-life input texts in a systematic and quantitative manner. Nevertheless, as explained in detail in Section 3.3.2.2, some indices were difficult to interrupt, repetitive of each other, or not useful or effective in determining the complexity of a text. A careful selection of the indices was essential. Besides, these automatic tools are upgraded regularly. For example, this study used CohMetrix version 2.0. The tool was subsequently upgraded to version 3.0 in which substantial changes were made. The number of indices was increased to 108. The abbreviation and

numbering of many indices were amended. These changes unfortunately make it difficult to compare the results of this study and previous studies with those of future studies.

Writing process questionnaire

As presented in Section 3.4.2, think-aloud protocol has been regarded as the most persuasive way of demonstrating the processes employed (for example see Hayes & Flower, 1983; Spivey, 1997; Plakans, 2010). However, as think-aloud is a very time-consuming method, it is usually used in studies with a small number of participants. The think-aloud method has also been criticised because of the reactivity and disruption imposed on the actual cognitive processes, especially with L2 participants (Smagorinsky, 1994; Stratman & Hamp-Lyons, 1994). Such a method was, therefore, not suitable for the context of this study which involved a large number of L2 participants in both real-life academic and test conditions. This study investigated the writing processes participants employed while completing the reading-into-writing test tasks and the real-life academic writing tasks. Following the recommendations of Purpura (1991) on the use of cognitive process questionnaires, the construct of the questionnaire was developed based upon human information processing theory (See Section 3.4.2.1). In addition, the psychometric characteristics of the questionnaire and the underlying construct validity of the questionnaire were verified by a series of statistical analyses (see Section 3.4.2.1). The questionnaire in this study was constructed paying particular attention to all the processes which seemed to be most relevant to our discussion of the cognitive validity of academic writing tests. The questionnaire, however well developed, can only seek evidence of the participants' perceptions of what they did. One should not rely upon these perceptions as evidence of actual performance. Future studies should attempt to triangulate the questionnaire data by another instrument, such as post-test interview or key-stroke logging.

Besides, as this study aimed to investigate the cognitive processes elicited by authentic operationalised reading-into-writing test tasks and real-life academic writing tasks, the mode of writing was not investigated as a variable of the writing

process. Both test tasks were paper-based whereas the real-life test tasks were expected to be majorly computer-based (even though some participants might have completed the real-life tasks on paper and then typed out the script). Future studies should investigate into how the mode of writing influences the employment of the cognitive processes. For example, paper-based writing might require more rigid organisational planning, which stronger test takers engage in but the weaker ones often do not. It would also require the rapid creation of graphic handwritten forms. By contrast, computer-based writing would allow for massive recursive editing and would rely upon a very different process of execution (Severinson Eklundh & Kollberg, 2003; Weir, O'Sullivan, Jin, & Bax, 2007). Grammatical errors such as spelling can be corrected retrospectively, and therefore immediate accuracy would be far less of an issue.

7.4 Implications of the findings and the contributions of this study

To the best of the researcher's knowledge, this study is the first study to validate two types of operationalised reading-into-writing test tasks by 1) comparing the contextual features of the test tasks with those of the real-life academic writing tasks, 2) comparing the cognitive processes which they elicit from test takers with the cognitive processes elicited by the real-life academic writing tasks, and 3) investigating the relationships between the test scores and academic outcomes. Acknowledging the limitations mentioned above, the results of this study have numerous important implications for the development and validation of reading-into-writing test tasks to assess academic writing ability.

7.4.1 The application of the socio-cognitive framework extended to integrated reading-into-writing tests

This study aimed to extend the application of Weir's (2005) socio-cognitive validation framework, of which current application is limited to tests of independent language skills, to integrated reading-into-writing tests. The framework is regarded to have 'direct relevance and value to an operational language testing/assessment context' and 'to be both theoretically sound and

practically useful' in relation to test development and validation (Taylor, 2011:2). However, while the framework has been widely used in test validation research (e.g. Geranpayeh and Taylor (eds) (2013) - Examining Listening; Khalifa & Weir (2009) - Examining Reading; Taylor (ed) (2011) - Examining Speaking; Shaw & Weir (2007) - Examining Writing), its current application is limited to independent language tests. This study went beyond the scope of the earlier studies to extend three components of the socio-cognitive framework for integrated reading-into-writing test tasks.

Table 7.1 presents the framework with explicit contextual and cognitive parameters and a reference of the predictive power of reading-into-writing test tasks for academic writing purposes. The parameters proposed here were driven from the literature as well as the results of this study which investigated the contextual features of real-life academic writing tasks, the processes students employed to complete these real-life tasks, and the correlations between the reading-into-writing test scores and real-life performance. Therefore, the framework proposed by this study has good theoretical and practical value for test development and validation of reading-into-writing tests for academic purposes. The framework aims to assist test developers and further researchers who intend to develop valid reading-into-writing test tasks for assessing academic writing ability and to conduct validity studies in such integrated task types.

The instruments developed in this study to investigate the validity of reading-into-writing test tasks including the **contextual parameter proforma,** the selection of the **17 automated textual analysis indices**, and the **writing process questionnaire** would be useful resources for test development and validation of integrated reading-into-writing test tasks for academic purposes.

Table 7.1 A framework for reading-into-writing tests for academic purposes

Contextual validity parameters of academic writing tasks with integration of reading materials
Overall task setting
• Clarity of purpose
• Topic domain

○ Academic ○ Professional	
• Genre (e.g. essay, report)	
• Cognitive demands	
• Language functions to perform (e.g. describing, defining, reasoning, illustrating visuals, citing sources, evaluating, predicting, recommending, synthesising, expressing personal views, summarising, reasoning)	
• Clarity of audience	
• Clarity of making criteria	
Input text features	
• Input format	
• Verbal input genre	
• Non-verbal genre	
• Discourse mode ○ Argumentative ○ Expository	
• Concreteness of ideas	
• Explicitness of textual organisation	
• Cultural specificity	
• Lexical complexity ○ High frequency words (K1) ○ High frequency words (K1+K2) ○ Academic words ○ Low frequency words (Offlist) ○ Log frequency content words ○ Average syllables per word ○ Type-taken ratio (content words)	
• Syntactic complexity ○ Average words per sentence ○ Logical operator incidence score ○ Mean number of modifiers per noun-phrase ○ Mean number words before the main clause in sentences ○ Sentence syntax similarity	
• Degree of cohesion ○ Adjacent overlap argument ○ Adjacent overlap stem ○ Adjacent overlap content word ○ Proportion of adjacent anaphor references ○ Adjacent semantic similarity (LSA)	
Cognitive validity parameters in academic writing with integration of reading materials	
Cognitive phases	Cognitive processes
Conceptualisation	• Task representation and macro-planning

	• Revising macro plans
Meaning and discourse construction	• Careful global reading • Selecting relevant ideas • Connecting and generating
Translation and micro-planning[15]	• Translating ideas into linguistic forms • Micro-planning
Organising	• Organising ideas in relation to input texts • Organising ideas in relation to own texts
Low-level monitoring and revising	• Low-level editing while writing • Low-level editing after writing
High-level monitoring and revising	• High-level editing while writing • High-level editing after writing
Predictive validity of reading-into-writing test tasks	
Correlations between Test Task A (essay with multiple verbal inputs) scores and academic outcome	r=0.31 ($p<0.001$)
Correlations between Test Task B (essay with multiple non-verbal inputs) scores and academic outcome	r=0.38 ($p<0.001$)

In addition to the recommendations for future studies to address the limitations of the present study provided in Section 7.3, the following areas would also be important for future studies to further extend the validation framework for integrated tests:

a) further investigation into the cognitive validity parameters, including *task representation, careful global reading, organising ideas in relation to own text* and *high-level editing while writing*, which showed some discrepancy in the underlying structure between the test and real-life condition, preferably by more different research instruments such as think-aloud and keystroke logging.

b) equivalent evidence demonstrating how these target cognitive processes are addressed by other operationalised integrated reading-into-writing test task types to build a thorough understanding of the cognitive validity of the integrated test task type;

c) the remaining two validity components of the socio-cognitive framework - scoring validity and consequential validity of the reading-into-writing test tasks for academic purposes which were not covered in this study; and

[15] The processes of translation and micro-planning were not investigated in this study (for reasons, see Chapter Five)

d) the possibility of extending the application of the socio-cognitive framework to different types of integrated task, e.g. listening-into-writing and listening-to-speaking;

7.4.2 A more complete construct definition of reading-into-writing test tasks for academic purposes

Based upon the framework proposed above, this study has built a more complete construct definition of reading-into-writing test tasks for academic purposes. The results of this study demonstrated that reading-into-writing test tasks can successfully operationalise the target contextual and cognitive parameters, and possess a promising predictive power.

The a priori context validity and cognitive validity are arguably the most important components to shape the construct of a test during the test development and validation. Khalifa & Weir (2009: 81) argued that 'the contextual parameters operationalised in a test should mirror the criterial features of the target situation activity as far as possible'. These findings of this study indicated that reading-into-writing test tasks were able to reflect real-life writing performance conditions in terms of overall task setting and input text features satisfactorily.

Evidence of context validity on its own is insufficient. Any valid tests have to demonstrate the extent to which they elicit from test takers cognitive processes that correspond to the processes that are elicited by real-life tasks in the target language context (Glaser, 1991; Shaw & Weir, 2007). A major threat to the cognitive validity is that the tasks might tap into a skill which is solely used under test conditions and demonstrate little relation to the real-life processes (Field, 2013; Shaw & Weir, 2007, Chapter Three; Weir et al, 2013, Chapter Three). However, it is a challenging task because a coherent model of reading-into-writing was lacking in the literature (Hirvela, 2004). Although the integrated reading-into-writing task type is generally perceived to have good cognitive validity (e.g. Hamp-Lyons & Kroll, 1996; Plakans, 2010; Weigle, 2002, 2004; Weir, et al, 2013), empirical evidence supporting the cognitive validity of such task types was apparently insufficient (the number of studies on reading-into-

writing processing is considerably smaller than that on writing-only or reading-only processing) and incomprehensive (most studies focusing on particular processes rather than the entire reading-into-writing processing) in the literature. The results of this study revealed that the two types of reading-into-writing test tasks were largely able to map on to the eleven target cognitive processes that are employed in the real-life academic context. The results of the confirmatory factor analyses also showed that both Test Task A and Test Task B were able to elicit from the participants the same underlying structure of most of these cognitive processes yielded by the real-life data. The findings hence provided strong evidence for the cognitive validity of reading-into-writing test tasks as a tool to assess academic writing ability.

The construct of reading-into-writing test tasks for academic purposes defined in this study possessed a unique set of contextual and cognitive parameters which was not the same as those defined in the independent reading tests (e.g. Khalifa & Weir, 2009) or independent writing tests (e.g. Shaw & Weir, 2007). Therefore, the results of this study identified the significance for test developers, university admissions officers and other stakeholders to consider the role of integrated reading-into-writing tasks as against independent reading tasks and independent writing tasks in academic language assessments, if they would like to be more certain of their students' ability to cope with academic writing.

7.4.3 The use of reading-into-writing test tasks in the pedagogical setting for academic purposes

While the use of reading-into-writing tasks in EAP classrooms was not the focus of this study, the results of this study also have important implications for the use of such integrated task type for teaching and learning academic writing.

A clear explication of the cognitive and contextual demands is critical for language testing as well as teaching for academic purposes. Urquhart & Weir (1998: 172) argued that in both testing and teaching contexts, appropriate texts need to be selected for readers to perform the target reading activities developed for them. Their notion is also applicable for our discussion of reading-into-writing

here. EAP teaching institutions and teachers need to be aware of the demands of the cognitive processes and the nature of the reading and writing texts their students will encounter in their academic studies, so as to equip them to cope with the demands. Echoing with the literature (e.g. Grabe, 2003; Johns, 1981, 1993; Lenski & Johns, 1997), the findings indicated clearly that reading-into-writing tasks would be a valid tool to prepare students for the demands of academic writing in real life.

In addition, this study unpacked the essential contextual and cognitive parameters of reading-into-writing tasks for academic purposes for EAP teachers to help them select, modify or develop appropriate course materials. Similarly, the results indicated clearly which cognitive processes are essential for successful completion of real-life academic writing tasks which involve integration of reading materials. The results showed that the high-achieving participants employed most of these cognitive processes more frequently than the low-achieving participants. This indicated the need for students to practise these reading-into-writing skills in EAP classrooms.

7.4.4 Implications for test writers to develop more valid reading-into-writing test tasks for academic purposes

Through a carefully demonstrated link between the test and the real-life conditions, the results of this study strongly suggested that the integrated reading-into-writing task type is a valid tool to assess academic writing ability in terms of the context validity, cognitive validity and criterion-related validity (see Chapters Four, Five and Six respectively). To assist test writers to develop more valid reading-into-writing test tasks for academic purposes, recommendations of overall test setting and input text manipulation are provided below.

Overall task setting

(1) *Incorporating other common academic writing genres, such as report*.

(2) *Avoiding the use of topics in the social domain.* Based on the judges' response, topic domains identified in the real-life tasks were academic and professional.

However, Test Task A was *academic* and *social* while Test Task B fell into the *professional* and *social* domains. The judges felt that both test tasks' input texts contained rather general content, which was usually connected to the social domain. This is, however, not a straightforward issue. As argued in Section 4.2.3, test developers would have to consider the facts that 1) topics in the social domain, which is usually adopted in language tests of general proficiency, are not entirely appropriate for the context of EAP writing tests, and 2) topics used in EAP writing tests should not include topics which involve content at a high level of specific knowledge.

(3) *Incorporating more language functions, e.g. defining, illustrating visuals, recommending.* The two test tasks required fewer language functions than the real-life tasks did. It is important to cover these language functions which have been identified in real-life academic writing in the test specification, even though not all functions need to be tested in every single testlet.

Input text manipulation

Apart from overall task setting, test writers need to be aware of the possible effects of their manipulating input texts for test purposes.

(4) *Incorporating more input genres and a combination of argumentative texts and expository texts*. Real-life tasks incorporated a range of input genres and a combination of argumentative texts and expository texts. However, due to the need of standardisation, the range of the input genres and types of texts of test tasks are often limited. All input texts on Test Task A (from ten testlets) were regarded as belonging to a simplified version of the argumentative essay genre whereas Test Task B (from one testlet) contained texts belonging to simplified versions of the expository report and news/magazine article genre. It is important for test developers to monitor if a sufficient range of the input genres and text types, as stated in the specification, is represented across the testlets.

(5) *Reducing the lexical complexity.* The lexical complexity of the test task input texts was seemingly more demanding than the real-life input texts. This is likely to be the results of input text manipulation of incorporating sufficient 'idea units' into the input texts with a usually tight word limit. Test Task A input texts had a greater density of *low frequency words*, mostly proper nouns, than the real-life input texts whereas Test Task B input texts had a higher proportion of academic words and low frequency words and a higher type-token ratio of all content words than the real-life input texts. While standardising the text length of the input texts, test developers have to monitor if lexical complexity of the input texts was increased unnecessarily due to the process of text manipulation.

(6) *Maintaining the lower syntactic complexity and higher degree of cohesion.* Generally speaking, the results suggested that it was less demanding to build the textual representation of the test task input texts than that of the real-life-input texts because the test task input texts were more explicitly organised and more cohesive than the real-life input texts. However, it seems appropriate for test task input texts to maintain less demanding syntactic complexity and a higher degree of cohesion than the real-life input texts, so that test takers would be able to perform the processes under more demanding conditions, e.g. greater time restrictions.

7.4.5 Implications for the significance and meaningfulness of correlations between test scores and real-life scores

Both Test Task A and Test Task B scores correlated significantly to academic outcomes, the former at a level of $r=0.31$ ($p<0.001$) and the latter at a level of $r=0.38$ ($p<0.001$). As summarised in Section 7.2.3, when compared to the figures obtained in the literature, the results of this study provide evidence that reading-into-writing test tasks have comparatively good predictive validity. Both reading-into-writing test task types (*essay with multiple verbal inputs* and *essay with multiple verbal and non-verbal inputs*) were proved to be able to significantly predict academic outcome, defined in this study in terms of performance on

individual real-life academic writing tasks. This implies that such integrated reading-into-writing task type is a valid tool to assess test takers' ability in academic writing.

In addition, the results of the study have brought an important insight that the two reading-into-writing test scores were able to predict academic performance better at high and low levels than at the medium-level in the context of this study. This implies that academic writing ability, as measured by a language test, might have limited impact on the medium-level academic achievement in the context of this study. Therefore, any high-stakes decisions for these medium-level test takers need to be made with extra caution, and supported by other forms of evidence.

The results of the correlations between test scores and real-life scores not only reviewed a general pattern concerning the validity of the two reading-into-writing test tasks in predicting academic performance, but also indicated how score interpretation can be improved to fulfil the test purposes.

As argued in Section 6.3.3, the latest UKBA regulations require students who wish to apply to study at the level of bachelor's degree in the UK under Tier 4 (General) of points-based system to have a minimum of level B2 of the CEFR (UKBA, 2013). Results on Test Task A (i.e. the GEPT advanced test) using a pass/fail dichotomy might no longer be entirely appropriate for university entry purposes. It would seem more appropriate to adjust the score reporting method on the GEPT Advanced to indicate the test takers who have reached the minimum threshold of English requirement for higher education in the UK, i.e. B2 of CEFR. The results of this study showed that academic performance, especially at the levels of Grades B and D, could be predicted by different score ranges on Test Task A.

On the other hand, Test Task B (i.e. UoB reading-into-writing test) was designed to indicate the test takers' need for academic writing support. As mentioned previously, a score of 8 and 9 out of 9 indicates no need for academic writing support, a score of 6 and 7 indicates a low-level need for academic writing

intervention, whereas a score of 5 or below indicates a high-level need for academic writing intervention. Nevertheless, the results of this study seemed to suggest that these cut-off scores of Test Task B which indicate different levels of academic writing support required might not be the most effective. Based on the correlation pattern between Test Task B and real-life performance (See Section 6.3.3), most participants who scored 7 or 8 tended to achieve an overall Grade B on the real-life writing tasks. Therefore, a score of 7 or above seemed to indicate no need for academic writing support. In addition, most participants who got a total score of 2.5 to 3.5 on Test Task B achieved an overall Grade D on the selected real-life tasks. Therefore, a score of 4 or below seemed to indicate a high-level need for academic writing intervention. Evidence in future research is required to confirm these recommendations.

References

Ackerman, J. M. (1990). Reading, writing and knowing: The role of disciplinary knowledge in comprehension and composing. *Technical report*, 40. Berkley: Center for the Study of Writing at the University of California. Berkerley and Carnegie Mellon University.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, NY: Continuum.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kujper, H., Nold, G., & Takala, S. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly*, *31*(1), 3–30.

Allwright, J., & Banerjee, J. (1997). *Investigating the accuracy of admissions criteria: A case study in a British university*. Lancaster: Centre for Research in Language Education, Lancaster University.

ALTE. (1998). *ALTE Handbook of European Language Examinations and Examination Systems*. Cambridge: UCLES.

ALTE. (2011). The CEFR Grid for Writing Tasks v2.1. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/CEFRWritingGridv2_1_presentation.doc

Archibald, A. (2001). Managing L2 writing proficiencies: Areas of change in students' writing over time. *International Journal of English Studies*, *1*(2), 153–174.

Armbruster, B. B., Anderson, T. H., & Ostertag, J. (1987). Does text structure / summarization instruction facilitate learning from expository text? *Reading Research Quarterly*, *22*(3), 331–346.

Ascención Delaney, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*, 140–150.

Asencion, Y. (2004). Validation of reading-to-write assessment tasks performed by second language learners. Unpublished PhD dissertation: Northern Arizona University.

Avdi, E. (2011). IELTS as a predictor of academic achievement in a Master's Program. *EA Journal*, *26*(2), 42–49.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language, Studies in Language Testing 1*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Barlett, F. C. (1932). *Remembering*. Cambridge: Cambridge University Press.

Beck, I. L., McKeown, M. G., Sinatra, M. G., & Loxterman, J. A. (1991). Revising social studies text from a text processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly, 27*, 251–276.

Belcher, D., & Hirvela, A. (2001). *Linking literacies: Perspectives on L2 reading-writing connections.* Ann Arbor: University of Michigan Press.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of Written Composition*. Hillside, NJ: Lawrence Erbaum Associates.

Bohm, A., Follari, M., Hewett, A., Jones, S., Kemp, N., Meares, D., Pearce, D., et al. (2004). *Vision 2020. Forecasting international student mobility a UK perspective.* Retrieved from http://www.britishcouncil.org/eumd_-_vision_2020.pdf

Brewer, W. F. (1980). Literary theory, rhetoric, and stylistics: Implications for psychology. In R. J. Spiro, B. C. Bruce, & Brewer, W. F (Eds.), *Theoretical models and processes of reading comprehension* (pp. 221–239). Hillsdale, NJ: Erlbaum.

Bridgeman, B., & Carlson, S. (1983). *Survey of Academic writing Tasks Required of Graduate and Undergraduate Foreign Students.* Princeton, NJ: Educational Testing Service.

Bridges, G. (2010). Demonstrating cognitive validity of IELTS Academic Writing Task 1. *Research Notes, 42*, 24-33.

Britt, M. A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology, 25*(4), 313–339.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Written Communication, 8*, 533–556.

Burtis, P. J., Bereiter, C., Scardamalia, M., & Tetroe, J. (1983). The development of planning in writing. In B. M. Kroll & G. Wells (Eds.), *Explorations in the development of writing* (pp. 153–174). New York: Wiley.

Campbell, C. (1990). Writing with others' words: using background reading text in academic compositions. In B. Kroll (Ed.), (pp. 211–230). *Second language writing: Research insights for the classroom* (pp. 211–230). Cambridge: Cambridge University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London: Longman.

Carson, J. (2001). A task analysis of reading and writing in academic contexts. In A. Hirvela (Ed.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 48-83). Ann Arbor, MI: University of Michigan Press.

Carson, J., & Leki, I. (Eds.) (1993). *Reading in the compositon classroom. Second language perspectives*. Boston: Heinle and Heinle.

Charge, N., & Taylor, L. B. (1997). Recent developments in IELTS. *ELT Journal, 51*(4), 374–380.

Cherkes-Julkowski, M., Sharp, S., & Stolzenberg, J. (1997). *Rethinking Attention Deficit Disorders*. Cambridge, MA: Brookline Books.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421–442.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background on reading comprehension, Studies in Language testing 4*. *Studies in Language Testing*. Cambridge: UCLES/Cambridge Uiversity Press.

Cobb, T. (2003). VocabProfile, The Compleat Lexical Tutor. Retrieved from http://www.lextotor.ca

Cohen, A. (1984). On taking language tests: What the students report. *Language Testing*, *1*, 70–81.

Cooper, A., & Bikowski, D. (2007). Writing at the graduate level: What tasks do professors actually require? *Journal of English for Academic Purposes*, 6, 206–221.

Cotton, F., & Conrow, F. (1998). An Investigation of the Predictive Validity of IELTS amongst a Group of International Students studying at the University of Tasmania. *IELTS Research Reports*, *1*(4), 72–115.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Strasbourg: Council of Europe.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

CRELLA. (forthcoming). *The construct document of UoB reading-into-writing test*. Unpublished test development report.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

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

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984–989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, *35*(2), 115–135.

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

Cumming, A, Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, *21*(2), 107–145.

Cumming, A, Kantor, R., Eedosy, U., Eouanzoui, K., James, M., & Erdosy, U. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*, 5–43.

Cumming, A. (1997). The testin of second-language writing. In C. Clapham (Ed.), *The encyclopedia of language and education: Volume 7. Language assessment* (pp. 51–63). Dordrecht, The Netherlands: Kluwer.

Davies, A. (2008). Assessing Academic English: Testing English Proficiency, 1950-1989 – the IELTS solution. Cambridge: Cambridge University Press.

Davies, A., & Criper, C. (1988). *English Language Testing Service Research Report 1. ELTS Validation Project Report*. British Council/UCLES.

Dooey, P. (1999). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. In K. Martin, N. Stanley and N. Davison (Eds), *Teaching in the Disciplines/Learning in Context* (pp. 114–118). Proceedings of the 8th Annual Teaching Learning Forum, The University of Western Australia, February 1999. Perth: UWA.

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS Test. *Prospect, 17*(1), 36–54.

Douglas, D. (2000). *Assessing language for specific purposes: theory and practice*. Cambridge: Cambridge University Press.

Eiken. (2013). Characteristics of TEAP. Retrieved from http://www.eiken.or.jp/teap/merit.html

English Language Testing. (2013). Password English Tests. Retrieved from https://www.englishlanguagetesting.co.uk/Password-English-Tests/

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper. TOEFL Monograph Series MS-17*. ETS.

Esmaeili, H. (2002). Reading-to-write reading and writing tasks and ESL students' reading and writing performance in an English langauge test. *The Canadian Modern Language Review, 58*, 599–622.

ETS. (2013). *TOEFL iBT Test Content*. Retrieved from http://www.ets.org/toefl/ibt/about/content/

Eysenck, M. W., & Keane, M. (2005). *Cognitive Psychology* (5th ed.). Hove: Psychology Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Pschological methods, 4*(3) 272-299.

Feak, C., & Dobson, B. (1996). Building on the impromptu: A source-based writing assessment. *College ESL, 6*(1), 73–84.

Feast, V. (2002). The impact of IELTS scores on performance at university. *International Education Journal, 3*(4), 70–89.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington DC: National Academy Press.

Field, J. (2004). *Psycholinguistics: the key concepts*. London: Routledge.

Field, J. (2008). *Listening in the Language Classroom.* Cambridge: Cambridge University Press.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking, Studies in Language Testing 30* (pp. 65–111). Cambridge: UCLES/Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining Listening: Research and Practice in Assessing Second Language Listening, Studies in Language Testing 35* (pp. 77–151). Cambridge: UCLES/Cambridge University Press..

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research*, *57*, 481–506.

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, *35*(1), 39–50.

Flower, L. (1990). Reading-to-write: Understanding the task. In L. Flower, V. Stein, J. Ackerman, M. J. Kantz, K. McCormick, & W. C. Peck (Eds.), *Reading to write. Exploring a cognitive & soical process* (pp. 35–73). New York: Oxford University Press.

Flower, L., & Hayes, J. R. (1980). The dynamic of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 31–50). Hillsdale, New Jersey: Erlbaum.

Flower, L., Stein, V., Ackerman, J., Kantz, M. J., McCormick, K., & Peck, W. C. (1990). *Reading-to-write: Exploring a Cognitive and Social Process*. New York: Oxford University Press.

Fulcher, G. (2010). *Practical Language Testing.* London: Hodder Education.

Galbraith, D., & Torrance, M. (1999). *From Problem Solving to Text Production*. *Studies in Writing*. Amsterdam: Amsterdam University Press.

Geranpayeh, A., & Taylor, L. (Eds). (2013). *Examining Listening: Research and practice in assessing second language listening, Studies in Language Testing 35*. Cambridge: UCLES/Cambridge University Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building* . Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrick & E. L. Baker (Eds.), *Testing and Cognition* (pp. 17–30). Englewood Cliffs: Prentice Hall.

Goldman, S. R., & Rakestraw, J. A. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research Vol. III* (pp. 311–335). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Gomulicki, B. R. (1956). Recall as an abstractive process. *Acta Psychologica*, *12*, 77–94.

Grabe, W. (2001). Reading-writing relations: Theoretical perspectives and instructional practices. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 15–47). Ann Arbor: The University of Michigan Press.

Grabe, W. (2003). Reading and writing relations: Second language perspectives on research and practice. In B Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 243–259). Cambridge: Cambridge University Press.

Grabe, W. (2009). *Reading in a Second Language: Moving from Theory to Practice*. New York: Cambridge University Press.
Grabe, W., & Kaplan, F. L. (1996). *Theory and Practice of Writing: An applied linguistic perspective*. London: Longman.
Grabe, W., & Stoller, F. L. (2002). *Teaching and Researching Reading*. London: Longman.
Graesser, A. C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A Web facility that helps survey methodologists improve the comprehensibility of questions. *Public Opinion Quarterly*, *70*, 3–22.
Graesser, A. C., McNamara, D., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.
Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, *40*(5), 223–234.
Graham, S. (2006). Writing. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (pp. 457–477). Mahwah, New Jersey: Lawrence Erlbaum Associates.
Graham, S., & Harris, K. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist*, *35*, 3–12.
Graham, S., & Harris, K. R. (1996). Self-regulation and strategy instruction for students who find writing and learning challenging. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications.* (pp. 347–360). Mahwah, New Jersey: Lawrence Erbaum Associates.
Green, A. (2005). EAP study recommendations and score gains on the IELTS academic writing test. *Assessing Writing*, *10*, 44–60.
Green, A. (2012a). The password test – design, development, and reliability. *English Language Testing*. Retrieved from http://www.englishlanguagetesting.co.uk/uploads/The-Password-Test-Design-Development-and-Reliability_5.pdf
Green, A. (2012b). *Theoretical and empirical bases for language construct definition across the ability range*. Cambridge: Cambridge University Press.
Green, A. B. (2007). *IELTS Washback in Context: Preparation for academic writing in higher education, Studies in Language Testing 25*. Cambridge: UCLES/Cambridge University Press.
Green, A., Unaldi, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, *27*(2), 191–211.
Green, A., Weir, C. J., Chan, S. H. C., Field, J., Taylor, L., Bax, S., & Nakatsuhara, F. (2012). *Report to Cambridge ESOL UK on the contextual parameters of CAE Examinations*. Unpublished research report.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree program. TOEFL Research Reports, RR-95-44*, Princeton, NJ: Educational Testing Service.
Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment. *College ESL*, *6*(1), 52–72.
Hamp-Lyons, L., & Kroll, B. (1997a). *TOEFL 2000 -- Writing: Composition, community, and assessment, TOEFL Monograph 5*. Princeton, NJ: Educational Testing Service.
Hamp-Lyons, L., & Kroll, B. (1997b). *TOEFL 2000 - Writing: Composition, community and assessment*. *TOEFL Monograph Report*. Princeton, NJ: Educational Testing Service.
Hartmann. (1995). Eight readers reading: The intertextual links of proficient readers reading multiple passages. *Reading Research Quarterly*, *30*, 520–561.
Hayes, J. R. (1996). A new Framework for understanding cognition and affect in writing. In S. Ransdell (Ed.), (pp. 1–28). Mahwah, New Jersey: Lawrence Erbaum Associates.
Hayes, J. R., & Flower, L. (1980). Identifying the organisation of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, N. J.: Lawrence Erbaum Associates.
Hayes, J. R., & Flower, L. (1983). Uncovering Cognitive Processes in Writing: an Introduction to Protocol Analysis. In P. Mosenthal, L. Tamor, & S. Walmsley (Eds.), *Research on Writing: Principles and Methods* (pp. 207–220). New York: Longman.
Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. *IELTS Research Reports*, *2*(52-63).
Hirvela, A. (2004). *Connecting reading and writing in second language writing instruction*. Ann Arbor, MI: The University of Micgigan Press.
Horowitz, D. M. (1986a). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly*, *20*, 445–460.
Horowitz, D. M. (1986b). Essay examination prompts and the teaching of acdemic teaching. *English for Specific Purposes*, *5*, 107–120.
Horowitz, D. M. (1991). ESL writing assessments: Contradictions and resolutions. In Liz Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 71–86). Norwood, NJ: Ablex.
Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge: Cambridge University Press.
Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A., Michael, R., & Walkinshaw, I. (2012). Tracking international students ' English proficiency. *IELTS Research Report Onine Series*, *1*, 1–41.
Hutchison, K. A. (2003). It is semantic priming due to association strength or to feature overlap? A micro-analytic review. *Psychonomic Bulletin and Review*, *10*, 785–813.
Hyland, K. (2002). *Teaching and Researching Writing, Applied Linguistics in Action Series*. London: Longman.

Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance, Part 1. In P. McGovern & S. Walsh (Eds.), *IELTS Research Report Volume 7* (pp. 137–204). IELTS Australia and British Council.

Johns, A. M. (1981). Necessay English: A faculty survey. *TESOL Quarterly*, *15*, 51–57.

Johns, A. M. (1993). Reading and writing tasks in English for academic purposes classes: Products, processes and resources. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 274–289). New York: Newbury House.

Johns, A. M., Bawarshi, A., Richard, M., Hyland, K., Paltridge, B., Reiff, M. J., & Tardy, C. (2006). Crossing the boundaries of genre studies: Commentaries by experts. *Journal of second language writing*, *15*, 234–249.

Johns, A. M., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics*, *11*(3), 253–271.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17.

Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory and Cognition*, *15*(3), 256–266.

Kellogg, R. T. (1994). *The psychology of writing.* New York: Oxford University Press.

Kellogg, R. T. (1996). A model working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications.* (pp. 57–71). Mahwah, New Jersey: Lawrence Erbaum Associates, Publishers.

Kellogg, R. T. (1999). Components of working memory in writing. In M. Torrance & G. Jeffery (Eds.), *THe cognitive deamds of writing: Processing capacity and working memory effects in text production* (pp. 43–61). Amsterdam: Amsterdam University Press.

Kellogg, R. T. (2001). Competition for working memory among writing processes. *American Journal of Psychology*, *114*(2), 175–192.

Kennedy, C., & Thorp, D. (2007). A corpus-based investigation of linguistic responses to an IELTS Academic Writing task. In L. Taylor (Ed.), *IELTS collected papers; research in speaking and writing assessment* (pp. 316–377). Cambridge: Cambridge University Press.

Kerstjens, M., & Nery, C. (2000). Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance. *IELTS Research Reports, 3*, 85-108.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading, Studies in Language Testing 29*. Cambridge: UCLES/Cambridge University Press.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Lawrence Erlbaum.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394.

Lado, R. (1961). *Language testing: the construction and use of foreign language tests*. London: Longman.
Ledoux, K., Traxler, M. J., & Saab, T. Y. (2007). Syntactic priming in comprehension: evidence from event-related potentials. *Psychological science, 18*, 135–143.
Lee, W., Kantor, R., & Mollaun, P. (2002). Score reliability as an essential prerequisite for validating new writing and speaking tasks for TOEFL. *Paper presented at the Annual TESOL Convention, Salt Lake City.*
Leki, I. (1993). Reciprocal themes in ESL reading and writing. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 221–233). Boston: Heinle and Heinle.
Leki, I., & Carson, J. G. (1994). Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly, 28*, 81–101.
Leki, I., & Carson, J. G. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31*, 39–69.
Lenski, S. D., & Johns, J. L. (1997). Patterns of reading-to-write. *Reading Research and Instruction, 37*, 15–38.
Levelt, W. J. M. (1989). *Speaking. From intention on articulation*. Cambridge, MA-London: ACL-MIT Press.
Lewkowicz, J. (1997). The intergrated testing of a second language. In C. Clapham (Ed.), *Encyclopedia of language and education, Vol. 7: Language testing and assessment* (Vol. 7, pp. 121–130). Dordrecht, The Netherlands: Kluwer academic publishers.
Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.
Lowenthal, D. (1980). Mixing levels of revision. *Visible Language, 14*, 383–387.
LTTC. (2012). About GEPT. Retrieved from http://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT.htm
Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.
MacArther, C. A., Graham, S., & Harris, K. R. (2004). Insights from instructural research on revision with struggling writers. In L. Allal, L. Chanquoy, & P. Largy (Eds.), *Revision: Cognitive and instructional processes* (pp. 125–137). Amsterdam, Netherlands: Kluwer Academic Press.
Mathison, M. A., & Spivey, N. N. (1993). Writing from academic sources: Authorship in writing the critique. Berkeley, CA.: National Center for the Study of Writing and Literacy.
McCarthy Young, K., & Leinhardt, G. (1998). Writing from primary documents: A way of knowing in history. *Written Communication, 15*, 25–68.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd edition)* (13–103). London, NY: McMillan.
Mislevy, R. J. (1992). Foundations of a new test theory. In R. J. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory of a new generation of tests* (pp. 19–39). Hillsdale, NJ: Erbaum.

Moore, T., & Morton, J. (1999). Authenticity in the IELTS academic module writing test: A comparative study of task 2 items and university assignments. *IELTS Research Report, 2, paper 4.* Canberra, IDP: IELTS Australia.

Moore, T., & Morton, J. (2005). Dimensinos of difference: a comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, *4*, 43-66.

Moore, T., Morton, J., & Price, S. (2010). *Construct validity in the IELTS academic reading test: A comparison of reading requirements in IELTS test items and in university study*. Unpublished IELTS Research Report. British Council/IDP Australia.

Murray, D. M. (1978). Internal revision. A process of discovery. In C. R. Cooper & L. Odell (Eds.), *Research on comosing.* (pp. 113–139). Urbana, IL: National Council of Teachers of English.

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing*. Unpublished PhD thesis, the University of Reading.

Odell, L. (1980). Business writing: Observations and implications for teaching composition. *Theory into Practice*, *19*, 225–232.

Oller, J. W. (1979). *Language Tests at School*. London: Longman.

Parodi, G. (2007). Reading – writing connections: Discourse-oriented research. *Reading and Writing*, 225–250.

Pearson. (2010). *PTE Academic Tutorial*. Pearson company. Retrieved from http://www.pearsonpte.com/SiteCollectionDocuments/PTEA_Tutorial_2Jul10_v2.pdf

Pearson, P. D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension and implicit information. *Journal of Reading Behaviour*, *11*(3), 201–209.

Perfetti, C. A. (1997). Sentences, individual differences and multiple texts: Three issues in text comprehension. *Discourse Processes*, *23*, 337–355.

Perfetti, C. A., Britt, M. A., & Georgi, M. C. (1995). *Text-based learning and reasoning : Studies in history*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of document representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). London: Lawrence Erlbaum Associates.

Perl, S. (1979). The composing processes of unskilled college writers. *Research in the Teaching of English*, *13*(4), 317–336.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*, 633–651.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*, 111–129.

Plakans, L. (2009a). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes*, 8, 252-266.

Plakans, L. (2009b). Discourse synthesis in integrated second language writing asessment. *Language Testing*, *26*(4), 561–587.

Plakans, L. (2010). Independent vs . Integrated Writing Tasks: A Comparison of Task Representation. *Tesol Quarterly*, *44*(1), 185–194.
Pollitt, A., & Taylor, L. (2006). Cognitive psychology and reading assessment. In M. Sainsbury, C. Harrison, & A. Watts (Eds.), *Assessing reading: from theories to classrooms* (pp. 38–49). Cambridge: National Foundation for Educational Assessment.
Purpura, J E. (1998). The development and construct validation of an instrument designed to investigate selected cognitive background characteristics of test-takers. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 111–140). Mahwah, New Jersey: Lawrence Erbaum Associates.
Purpura, J. E. (1997). An Analysis of the Relationships Between Test Takers' Cognitive and Metacognitive Strategy Use and Second Language Test Performance. *Language Learning*, *47*(2), 289–325.
Purves, A. C., Soter, A., Takala, S., & Vahapassi, A. (1984). Towards a domain-referenced system for classifying assignments. *Research in the Teaching of English*, *18*(4), 385–416.
Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Engewood Cliffs, NJ: Prentice-Hall.
Read, J, & Hayes, B. (2003). The impact of the IELTS test on preparation for academic study in New Zealand. *IELTS Research Reports*, *4*, 153–206.
Read, John. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes*, *9*(2), 109–121.
Ruiz-Funes, M. (2001). Task representation in foreign language reading-to-write. *Foreign Language Annals,*, *34*(1), 226–234.
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.
Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing*, *17*, 85–114.
Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. (Ed. . Rosenberg (Ed.), *Advances in Applied Psycholinguistics, Volume 2: Reading, writing and language learning*. Cambridge: Cambridge University Press.
Scardamalia, M., & Bereiter, C. (1991). Literate expertise. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 172–194). Cambridge: Cambridge University Press.
Scardamalia, M., & Bereiter, C. (1996). Adaptation and understanding: A case for new cultures of schooling. In S. Vosniadou, E. DeCorte, R. Glaser, & H. Mandl (Eds.), *International perspectives on the design of technology-supported learning environments* (pp. 149–163). Mahwah, NJ: Erlbaum.
Scardamalia, M., & Paris, P. (1985). The function of explicit discourse knowledge in the development of text representations and composing strategies. *Cognition and Instruction*, *2*, 1-39.
Segev-Miller, R. (2007). Cognitive processes in discourse synthesis: The case of intertextual processing strategies. In G. Rijlaarsdam, M. Torrance, L. Van

Waes, & D. Galbraith (Eds.), *Writing and Cognition: Research and Applications* (pp. 231–250). Amsterdam: Elsevier.
Seifert, C. M., Robertson, S. P., & Black, J. B. (1985). Types of inferences generated during reading. *Journal of Memory and Language*, *24*, 405–422.
Severinson Eklundh, K., & Kollberg, P. (2003). Emerging discourse structure: computer-assisted episode analysis as a window to global revision in university students' writing. *Journal of Pragmatics*, *35*(6), 869–891.
Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*, *Studies in Language Testing 26*. Cambridge: UCLES/Cambridge University Press.
Shi, L. (2004). Textual borrowing in second language writing. *Written Communication*, *21*(2), 171–200.
Smagorinsky, P. (1994). Thnk-aloud protocol analysis: Beyond the black box. In P. Smagorinsky (Ed.), *Speaking about writing* (pp. 3–19). USA: Sage Publications, Inc.
Spivey, N. N. (1984). *Discourse synthesis: Constructing texts in reading and writing. Outstanding Dissertation Monograpg*. Newark, DE: International Reading Association.
Spivey, N. N. (1990). Transforming texts: Constructive processes in reading and writing. *Written Communication*, *7*(2), 256–287.
Spivey, N. N. (1991). The shaping of meaning: options in writing the comparison. *Research in the Teaching of English*, *25*, 390–418.
Spivey, N. N. (1997). *The constructivist metaphor: Reading, writing and the making of meaning*. San Diego, CA: Academic Press.
Spivey, N. N. (2001). Discourse synthesis: Process and product. *Discourse synthesis: Studies in historical and contemporary social epistemology* (pp. 379–396). Westport, CT: Praeger.
Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly*, *24*(1), 7–26.
Stratman, J., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for Research. In P. Smagorinsky (Ed.), (pp. 89–112). USA: Sage Publications, Inc.
Taylor, L. (Ed). (2011). *Examining Speaking: Research and practice in assessing second language speaking*, *Studies in Language Testing 30*. Cambridge: UCLES/Cambridge University Press.
The British National Corpus. (2007). BNC. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/
Tierney, R. J., & Shanahan, T. (1991). Research on the reading-writing relationship: Interactions, transactions, and outcomes. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research Vol. II* (pp. 246–280). Mahwah, New Jersey: Longman.
Trinity College London. (2009). *Integrated skills in English (ISE) examinations Syllabus - From 1 February 2010*. Trinity College London: London.
Trinity College London. (2012). *Integrated Skills in English. Theoretical background and research*. Internal test development documents.

Trinity College London. (2013). *Exam Information: Integrated Skills in English (ISE).* Trinity College London: London,
UKBA. (2013). Eglish Language Ability. Retrieved from http://www.ukba.homeoffice.gov.uk/visas-immigration/studying/adult-students/can-you-apply/english-language/
UKCISA. (2012). Annual report 2011-2012. Retrieved from http://www.ukcisa.org.uk/files/pdf/about/annual_review.pdf
University of Bedfordshire. (2013). Business School About Us. Retrieved from http://www.beds.ac.uk/howtoapply/departments/businessschool/about-us
University of Reading. (2013). *About the Test of English for Educational urposes (TEEP).* Retrieved from https://www.reading.ac.uk/ISLI/english-language-courses/english-language-tests/islc-teep-about-teep.aspx
Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Proocess, product and practice*. Harlow: Longman.
Ushioda, E., & Harsch, C. (2011). Addressing the needs of international students with academic writing difficulties: Pilot project 2010/11, Strand 2: Examining the predictive validity of IELTS scores. Retrieved from http://www2.warwick.ac.uk/fac/soc/al/research/groups/ellta/projects/strand_2_project_report_public.pdf
Van Dijk, T A, & Kintsch, W. (1983). The notion of macrostructure. In T A van Dijk & W. Kintsch (Eds.), *Strategies of discourse comprehension* (pp. 189–223). New York: Academic Press.
Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.
Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, The baseline study. ETS TOEFL Monograph Series*, *34*(June), Princeton, NJ: Educational Testing Service.
Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2 , Coping with change. TOEFL iBT Research Report*, *05*(July), Princeton, NJ: Educational Testing Service.
Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.
Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(9), 27–55.
Weir, C J. (1983). *Identifying the language problems of the overseas students in tertiary education in the United Kingdom*. Unpublished PhD dissertation: University of London.
Weir, C J. (1988). Construct validity. In A. Hughes, D. Ported, & C. Weir (Eds.), *ELT Validation Project: Proceeding of a Conference Held to Consider the ELTS Validation Project Report*. Cambridge: The British Council and the University of Cambridge Local Examination Syndicate.
Weir, C J. (1990). *Communicative language testing*. Englewood Cliffs, NJ: Prentice Hall.
Weir, C J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Weir, C J. (2013). Measured constructs: A history of Cambridge English language examinations 1913-2012. *IELTS Research Report*, *51*, 2–10.
Weir, C J, Chan, S. H. C., & Nakatsuhara, F. (2013). Examining the Criterion-realted Validity of the GEPT Advanced Reading and Writing Tests: Comparing GEPT with IELTS and Real-life Academic Performance. *LTTC-GEPT Research Reports*, *RG-01*, 1–43.
Weir, C J, Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. *IELTS Research Report*, *9*(3), 97–156.
Weir, C J, O'Sullivan, B., Jin, Y., & Bax, S. (2007). Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS Writing component: effects and impact. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports Volume 7* (pp. 311–347). IELTS Australia and British Council.
Weir, C J, Vidakovic, I., & Galaczi, E. (2013). *Measured Constructs: A history of the constructs underlying Cambridge English language (ESOL) examinations 1913-2012*. Cambridge: Cambridge University Press.
Weir, C J, & Wu, J. R. W. (2006). Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167–197.
Weir, C J, Yang, H., & Jin, Y. (2000). *An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes, Studies in Language Testing 12*. Cambridge: UCLES/Cambridge University Press.
Weir, C. J. (2012). *TEAP Writing Project: A review of the washback concept in language education with particular implications for Japan and the EIKEN foundation of Japan organisation.* Unpublished Research Report.
Wengelin, A., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior research methods*, *41*(2), 337–51.
Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, *91*(2), 301–311.
Wu, R. Y. F. (2012). *Establishing the validity of the General English Proficiency Test Reading Component through a critical evaluation on alignment with the Common European Framework of Reference*. Unpublished PhD dissertation: University of Bedfordshire.
Yen, D., & Kuzma, J. (2009). Higher IELTS score, higher academic performance? The validity of IELTS in predicting the academic performance of Chinese students. *Worcester Journal of Learning and Teaching*, *3*, 1–7.
Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing*, *25*(4), 521–551.
Yu, G. and Lin, S.-W. (forthcoming). A comparablility study on the cognitive processes of taking GEPT (Advanced) and IELTS (Academic) Writing tasks using graph prompts. *LTTC-GEPT Research Reports*.

Appendix 2.1 Examples of reading-into-writing test tasks

2.1.1 Pearson PTE Academic Part 1 - Summarize Written Text

The copyright of Pearson PTE Academic Part 1 belongs to Person. The sample is included in hard copies of the thesis for examining purposes and do not appear in the electronic version of the thesis.

(taken from Pearson, 2010)

2.1.2 Trinity College London's ISE III exam Task 1

The copyright of Integrated Skills of English III Examination belongs to Trinity College London. The sample is included in hard copies of the thesis for examining purposes and do not appear in the electronic version of the thesis.

(taken from Trinity, 2013)

Appendix 3.1 Real-life tasks and reading-into-writing test tasks

3.1.1 Real-life Task A – Essay task

The copyright of Real-life Task A belong to University of Bedfordshire. The sample is included in hard copies of the thesis for examining purposes and do not appear in the electronic version of the thesis.

3.1.2 Real-life Task B – Report task

The copyright of Real-life Task B belong to University of Bedfordshire. The sample is included in hard copies of the thesis for examining purposes and do not appear in the electronic version of the thesis.

3.1.3 Reading-into-writing Test Task A

The copyright of GEPT Advanced Writing Past Paper belong to the Language Training and Testing Center (LTTC). Official permission to use of the test has been granted by the LTTC. The test item is included in hard copies of the thesis for examining purposes and do not appear in the electronic version of the thesis.

3.1.4 Reading-into-writing Test Task B

Appendix 3.2 Glossary of Contextual Parameters Proforma

Part 1 – Overall task setting

1. Purpose - Is the communicative purpose for completing the task clearly presented?
2. Topic Domain - What is the domain of the topic?
    - Personal - relates to personal lives (e.g. family, relatives, friends, etc.).
    - Social - relates to issues concerning the members of the public.
    - Professional - relates to expert and specialised knowledge of a profession.
    - Academic - relates to a particular discipline or field of study (which may have no practical purpose or use).
3. Genre - What is the genre of the text to be produced?
    - Essay is a piece of writing which is often written from an author's personal point of view.
    - Report is an informational piece of work made with the specific intention of relaying information or recounting certain events.
    - Case study is an intensive analysis of a person, group, or event in a specific context.
    - Summary is a short document that summarises a longer report or proposal or a group of related reports, in such a way that readers can rapidly become acquainted with a large body of material without having to read it all.
4. Cognitive demands - Which level of cognitive demands does the task impose on the candidates/students? (Think of the minimum requirement to complete the task).
    - Telling/retelling content: the text production is primarily guided by a direct retrieval of content from long-term memory or a direct copy from the input texts.
    - Organising/reorganising content: the text production requires writers to purposefully organise the content they retrieved from long-term memory and/or selected from the input texts in order to solve the rhetorical problems of the writing task.
    - Transforming content: the text production requires writers to establish a high awareness of the rhetorical situation of the writing task. Writers are required to strategically organise as well as transform (e.g. synthesise, interpret, evaluate) the content they retrieved from long-term memory and/or selected from the input texts to fulfil the writing goals.
5. Language functions - What language functions do the candidates/students have to demonstrate?
6. Intended reader - Is the intended reader clearly presented?
7. Knowledge of criteria - Are the marking criteria clearly presented?

Part 2 – Input text features

8. Input format – What is the format of the input?
9. Verbal input genre – What is the genre of the input text?
10. Non-verbal input – What is the non-verbal input provided in the input text?
11. Discourse mode – What is the primary discourse mode of the input text?
    - Narrative texts recount an event or a series of related events.
    - Descriptive texts describe a person, place or thing using sensory details.
    - Expository texts give information about or an explanation of an issue, subject, method or idea.
    - Argumentative texts typically involve a course of reasoning.
12. Concreteness of ideas – How concrete or abstract is the content of the input text?
13. Explicitness of textual organisation – How explicit or inexplicit is the textual organisation of the input text?
14. Cultural specificity – How culturally neutral or specific is the content of the input text?

Appendix 3.3 Expert Judgement Feedback Questionnaire

Based on the experience applying the Contextual Parameter Proforma, how confident do you feel when you choose your response? Please tick 1, 2, 3 or 4 to indicate how confident you were. If your answer is 2 or 1, please specify the reason.
4 = very confident
3 = confident
2 = not confident
1 = not confident at all

	4	3	2	1	Reasons
Part 1 - Overall task setting					
1. Purpose					
2. Topic domain					
3. Genre					
4. Cognitive demands					
5. Language functions					
6. Intended reader					
7. Knowledge of criteria					

	4	3	2	1	Reasons
Part 2 - Input text features					
8. Input format					
9. Verbal input genre					
10. Non-verbal input					
11. Discourse mode					
12. Concreteness of ideas					
13. Explicitness of textual organisation					
14. Cultural specificity					

Appendix 3.4 Writing Process Questionnaire (The pilot version – 54 items)

No.	Questionnaire items	No. in the main study
Reading task prompt		
1	I read the task prompt (i.e. instructions) carefully to understand each word in it.	1.1
2	I thought of what I might need to write to make my essay relevant and adequate to the task.	1.2
3	I thought of how my essay would suit the expectations of the intended reader.	1.3
4	I was able to understand the instructions for this writing test very well.	1.4
5	I thought about the purpose of the task.	1.5
Reading source texts		
6	I read through the whole of each source text carefully.	2.1
7	I read the whole of each source text more than once.	2.2
8	I used my knowledge of how texts like these are organised to find parts to focus on.	2.3
9	I searched quickly for part(s) of the texts which might answer the question.	2.4
10	I read some relevant part(s) of the texts carefully.	2.5
11	I used my knowledge of the topic to help me to understand the texts.	Deleted
12	I read the task prompt again while reading the source texts.	2.6
13	I took notes on or underlined the important ideas in the source texts.	2.7
14	I linked the important ideas in the source texts to what I know already.	2.9
15	I worked out how the main ideas across the source texts relate to each other.	2.11
16	I developed new ideas or a better understanding of existing knowledge.	2.12
Before writing		
17	I organised the ideas I plan to include in my essay.	3.1
18	I recombined or reordered the ideas to fit the structure of my essay.	3.2
19	I prioritised the important ideas in the source texts in my mind.	2.8
20	I removed some ideas I planned to write.	3.3
21	I tried to use the same organizational structure as in one of the source texts.	3.4
While writing		
22	I sometimes paused to organize my ideas.	4.1
23	I developed new ideas while I was writing.	4.2
24	I made further connections across the source texts.	4.3

25	I re-read the task prompt.	4.4
26	I selectively reread the source texts.	4.5
27	I monitored and edited the content development of my text.	deleted
28	I checked that the content was relevant.	4.7
29	I checked that I included all appropriate main ideas from all the source texts.	4.10
30	I checked that I included my own viewpoint on the topic.	4.11
31	I checked that the essay was well-organised	4.8
32	I checked that the essay was coherent, e.g. appropriate use of topic sentences, connectives and signals of changes in ideas etc.	4.9
33	I checked that the quotations were properly made, e.g. the quotes were relevant, the quotes were integrated grammatically into the essay, etc	4.12
34	I checked that I had put the ideas of the source texts into my own words.	4.13
35	I checked the possible effect of my writing on the intended reader.	4.14
36	I monitored and edited the linguistic aspect of my text.	deleted
37	I checked the accuracy of the sentence structures.	4.15
38	I checked if the range of sentence structures was adequate.	
39	I checked the appropriateness of vocabulary.	4.16
40	I checked the range of vocabulary.	
41	I monitored and edited the content development of my text.	deleted
After writing the first draft		
42	I checked that the content was relevant.	5.7
43	I checked that I included all appropriate main ideas from all the source texts.	5.10
44	I checked that I included my own viewpoint on the topic.	5.11
45	I checked that the essay was well-organised	5.8
46	I checked that the essay was coherent, e.g. appropriate use of topic sentences, connectives and signals of changes in ideas etc.	5.9
47	I checked that the quotations were properly made, e.g. the quotes were relevant, the quotes were integrated grammatically into the essay, etc	5.12
48	I checked that I had put the ideas of the source texts into my own words.	5.13
49	I checked the possible effect of my writing on the intended reader.	5.14
50	I monitored and edited the linguistic aspect of my text.	deleted
51	I checked the accuracy of the sentence structures.	5.15
52	I checked if the range of sentence structures was adequate.	
53	I checked the appropriateness of vocabulary.	5.16
54	I checked the range of vocabulary.	

Appendix 3.5 Writing Process Questionnaire (The main study version – 48 items)

Reading task prompt	
1	1.1: I read the whole task prompt (i.e. instructions) carefully.
2	1.2: I thought of what I might need to write to make my text relevant and adequate to the task.
3	1.3: I thought of how my text would suit the expectations of the intended reader.
4	1.4: I was able to understand the instructions for this writing task very well.
5	1.5: After reading the prompt, I thought about the purpose of the task. .
Reading source texts	
6	2.1: I read through the whole of each source text carefully.
7	2.2: I read the whole of each source text more than once.
8	2.3: I used my knowledge of how texts like these are organised to find parts to focus on.
9	2.4: I searched quickly for part(s) of the texts which might help complete the task.
10	2.5: I read some relevant part(s) of the texts carefully.
11	2.6: I went back to read the task prompt again.
12	2.7: I took notes on or underlined the important ideas in the source texts.
13	2.8: I prioritised important ideas in the source texts in my mind.
14	2.9: I linked the important ideas in the source texts to what I know already.
15	2.10: I worked out how the main ideas in each source text relate to each other.
16	2.11: I worked out how the main ideas across the source texts relate to each other.
17	2.12: I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.
18	2.13: I changed my writing plan while reading the source texts.
Before writing	
19	3.1: I organised the ideas for my text before starting to write.
20	3.2: I recombined or reordered the ideas to fit the structure of my essay.
21	3.3: I removed some ideas I planned to write.
22	3.4: I tried to use the same organizational structure as in the source texts.
While writing	
23	4.1: While I was writing I sometimes paused to organize my ideas.
24	4.2: I developed new ideas
25	4.3: I made further connections across the source texts.
26	4.4: I re-read the task prompt.
27	4.5: I selectively re-read the source texts.
28	4.6: I changed my writing plan (e.g. structure, content etc)
29	4.7: I checked that the content was relevant.
30	4.8: I checked that my text was well-organised.
31	4.9: I checked that my text was coherent.

32	4.10: I checked that I included all appropriate main ideas from all the source texts.
33	4.11: I checked that I included my own viewpoint on the topic.
34	4.12: I checked that the quotations were properly made.
35	4.13: I checked that I had put the ideas of the source texts into my own words.
36	4.14: I checked the possible effect of my writing on the intended reader.
37	4.15: I checked the accuracy and range of the sentence structures.
38	4.16: I checked the appropriateness and range of vocabulary.
After writing the first draft	
39	5.7: I checked that the content was relevant.
40	5.8: I checked that my text was well-organised.
41	5.9: I checked that my text was coherent.
42	5.10: I checked that I included all appropriate main ideas from all the source texts.
43	5.11: I checked that I included my own viewpoint on the topic.
44	5.12: I checked that the quotations were properly made
45	5.13: I checked that I had put the ideas of the source texts into my own words.
46	5.14: I checked the possible effect of my writing on the intended reader.
47	5.15: I checked the accuracy and range of the sentence structures.
48	5.16: I checked the appropriateness and range of vocabulary.

Appendix 3.6 Writing Process Questionnaire – Student version

WRITING PROCESS QUESTIONNAIRE

Section 1: Personal data

1. **Name:** _____(*Given name*)_____(*Surname*)
2. **Reference number:** _____
3. **Gender:** male / female *(Please circle)*
4. **IELTS results** *(if any)*: Overall band:_____ Reading: _____ Writing: _____

Section 2: Writing processes

In this section, there are some statements about how you might complete the test you have just taken. Please answer all the questions, thinking about what you did

- *while reading the task prompt* 閱讀試題時
- *while reading the source texts* 閱讀兩篇文章時
- *before starting to write* 開始寫作前
- *while writing the first draft* 寫作首稿時
- *after writing the first draft* 完成首稿後

in the test taking experience you have just had.

Please circle the extent of your agreement or disagreement to each statement below, using the following 4-point scale:

4 Definitely agree 完全同意
3 Mostly agree 基本上同意
2 Mostly disagree 基本上不同意
1 Definitely disagree 完全不同意

	1. **Reading task prompt for the test** (i.e. instructions) 閱讀試題時 Please think about what you did while you were reading the task prompt.	Definitely Agree / Mostly Agree / Mostly Disagree / Definitely Disagree
1.1	我仔細地閱讀試題去明白當中每一個字。 I read the task prompt (i.e. instructions) **carefully** to understand each word in it.	4 3 2 1
1.2	我考慮爲要寫什麼內容令文章貼題充足。 I thought of **what I might need to write** to make my essay relevant and adequate to the task.	4 3 2 1
1.3	我考慮我的文章怎樣符合讀者的期望。 I thought of how my essay would suit **the expectations of the intended reader**.	4 3 2 1
1.4	我能夠完全明白試題。 I was able to **understand** the instructions for this writing test **very well**.	4 3 2 1
1.5	我考慮這篇文章要達到的目的。 I thought about **the purpose** of the task.	4 3 2 1
閱讀兩篇試題文章前，你還做了些什麼？ What else did you do while reading the prompt?		

2. **Reading the source texts** (i.e. the two articles) 閱讀兩篇文章時 Please think about what you did while you were reading the source texts.		Definitely Agree	Mostly Agree	Mostly Disagree	Definitely Disagree
2.1	我仔細地閱讀兩篇文章的全部。 I read through **the whole of** each source text **carefully**.	4	3	2	1
2.2	我閱讀兩篇文章的全部多於一次。 I read the whole of each source text **more than once**.	4	3	2	1
2.3	我用對這類文章結構的認識去尋找的要注意的部份。 I **used my knowledge of how texts like these are organized** to find parts to focus on.	4	3	2	1
2.4	我迅速地尋找對答題有幫助的部份。 I **searched quickly** for part(s) of the texts which might answer the question.	4	3	2	1
2.5	我仔細地閱讀對答題有幫助的部份。 I read some relevant part(s) of the texts **carefully**.	4	3	2	1
2.6	我再次閱讀試題。 I **read the task prompt again**.	4	3	2	1
2.7	我在兩篇文章的重點做筆記或畫底線。 I **took notes on or underlined** the important ideas in the source texts.	4	3	2	1
2.8	我在腦海為兩篇文章的重點按重要性排序。 I **prioritised the important ideas** in the source texts in my mind.	4	3	2	1
2.9	我把兩篇文章的重點聯繫到我已有的知識。 I **linked the important ideas in the source texts to what I know already**.	4	3	2	1
2.10	我把一篇文章的重點聯繫起來。 I worked out how the **main ideas in each source text** relate to each other.	4	3	2	1
2.11	我把兩篇文章的重點聯繫起來。 I worked out how the **main ideas across the source texts** relate to each other.	4	3	2	1
2.12	我產生新的觀點或更加明白已有的知識。 I **developed new ideas or a better understanding** of existing knowledge.	4	3	2	1
2.13	我修改寫文章的計畫（如內容或結構）。 I **changed my writing plan** (e.g. structure, content etc)	4	3	2	1
閱讀兩篇試題文章時您還做了些什麼？ What else did you do while reading the source texts?					

3. **Before starting to write** 開始寫作前 Please think about what you did before starting to write your essay.		Definitely Agree	Mostly Agree	Mostly Disagree	Definitely Disagree
3.1	我整理計畫要寫的論點。 I **organized the ideas** I plan to include in my essay.	4	3	2	1
3.2	為了配合自己文章的結構，我從新聯合或排序計畫要寫的論點。 I **recombined or reordered the ideas to fit the structure** of my essay.	4	3	2	1
3.3	整理論點時，我曾刪除某些論點。 I **removed** some ideas I planned to write.	4	3	2	1
3.4	我嘗試採用其中一篇試題文章的結構。 I **tried to use the same organizational structure** as in one of the source texts.	4	3	2	1
寫作之前您還做了些什麼？ What else did you do before writing?					

4. **While writing the 1st draft and after writing the 1st draft** 寫作首稿時及完成首稿後 Please think about what you did while you were writing the first draft of your essay and after you had finished the first draft.		While writing the 1st draft	
		4 = Always / 3 = Often / 2 = Rarely / 1 = Never	
4.1	我間中暫停寫作去整理文章的論點。 While I was writing, I sometimes paused to **organize** my ideas.	4 3 2 1	
4.2	寫作時我產生新的觀點。 I **developed new ideas** while I was writing.	4 3 2 1	
4.3	寫作時我看到更多兩篇試題文章的關聯。 I **made further connections** across the source texts.	4 3 2 1	
4.4	我再次閱讀試題。 I **re-read the task prompt**.	4 3 2 1	After writing the 1st draft
4.5	我再次有選擇地閱讀兩篇試題文章。 I **selectively reread** the source texts.	4 3 2 1	4 = Always / 3 = Often / 2 = Rarely / 1 = Never
4.6	寫作時我修改寫文章的計畫(如內容或結構)。 I **changed** my **writing plan** (e.g. structure, content etc).	4 3 2 1	
4.7	我檢查了文章內容是否貼題。 I checked that the content was **relevant**.	4 3 2 1	4 3 2 1
4.8	我檢查了文章是否結構嚴謹。 I checked that the essay was **well-organized**	4 3 2 1	4 3 2 1
4.9	我檢查了文章是否連貫通順(如起首句子和連接詞的運用)。 I checked that the essay was **coherent**, e.g. appropriate use of topic sentences, connectives etc.	4 3 2 1	4 3 2 1
4.10	我檢查了文章已經包括兩篇試題文章的主要論點。 I checked that I included **all appropriate main ideas** from all the source texts.	4 3 2 1	4 3 2 1
4.11	我檢查了文章已經包括個人觀點。 I checked that I included my **own viewpoint** on the topic.	4 3 2 1	4 3 2 1
4.12	我檢查了有否適當地引用其他作者的意見或話語(如合乎文法和配合文章論點)。 I checked that the **quotations** were properly made, e.g. the quotes were relevant, the quotes were integrated grammatically into the essay, etc	4 3 2 1	4 3 2 1
4.13	我檢查了有否用自己的文字表達原文的論點。 I checked that I had put the ideas of the source texts into **my own words**.	4 3 2 1	4 3 2 1
4.14	我評估了文章對目標讀者可能產生的影響。 I checked the **possible effect** of my writing **on the intended reader**.	4 3 2 1	4 3 2 1
4.15	我檢查了句子的結構是否準確、種類是否足夠。 I checked the **grammatical accuracy and range** of the sentence structures.	4 3 2 1	4 3 2 1
4.16	我檢查了詞彙的拼寫和用法是否準確、種類是否足夠。 I checked the **spelling, usage and range** of the vocabulary.	4 3 2 1	4 3 2 1
寫作首稿時您還做了些什麼? What else did you do while writing the first draft? 完成首稿後您還做了些什麼? What else did you do after writing the first draft?			

The end

Appendix 5.1 Comparisons of the cognitive processes elicited by the two real-life tests

		Report		Essay					
		Mean	Std Dev	Mean	Std Dev	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
1.1	I read the whole task prompt (i.e. instructions) carefully.	3.41	.779	3.43	.527	2408.000	4893.000	-.669	.504
1.2	I thought of what I might need to write to make my text relevant and adequate to the task.	3.51	.669	3.34	.634	2158.000	4643.000	-1.796	.072
1.3	I thought of how my text would suit the expectations of the intended reader.	3.18	.770	3.04	.751	2282.000	4767.000	-1.195	.232
1.4	I understood the instructions for this writing task very well.	3.25	.703	3.03	.636	2091.500	4576.500	-2.105	.035
1.5	After reading the prompt, I thought about the purpose of the task. .	3.10	.869	3.14	.728	2550.000	5251.000	-.022	.982
2.1	I read through the whole of each source text slowly and carefully.	2.95	.797	2.93	.729	2488.000	4973.000	-.299	.765
2.2	I read the whole of each source text more than once.	3.01	.858	2.89	.877	2369.000	4854.000	-.795	.427
2.3	I used my knowledge of how texts like these are organised to find parts to focus on.	3.16	.727	3.01	.648	2228.500	4713.500	-1.459	.145
2.4	I searched quickly for part(s) of the texts which might help complete the task.	3.14	.787	2.97	.680	2195.000	4680.000	-1.581	.114
2.5	I read some relevant part(s) of the texts carefully.	3.45	.688	3.27	.658	2133.500	4618.500	-1.894	.058
2.6	I went back to read the task prompt again.	3.23	.842	3.43	.579	2320.500	5021.500	-1.044	.297
2.7	I took notes on or underlined the important ideas in the source texts.	3.27	.786	3.30	.768	2522.500	5223.500	-.144	.886
2.8	I prioritised important ideas in the source texts in my mind.	2.85	.908	2.96	.669	2445.000	5146.000	-.479	.632
2.9	I linked the important ideas in the source texts to what I know already.	3.18	.887	3.07	.666	2224.500	4709.500	-1.448	.147
2.10	I worked out how the main ideas in each source text relate to each other.	3.07	.805	3.06	.759	2513.500	4998.500	-.180	.857
2.11	I worked out how the main ideas across the source texts relate to each other.	3.05	.880	3.06	.740	2487.000	4972.000	-.293	.769
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.	3.19	.811	3.11	.733	2358.500	4843.500	-.861	.389
2.13	I changed my writing plan while reading the source texts.	2.86	.855	2.83	.900	2523.000	5008.000	-.137	.891
3.1	I organised the ideas for my text before starting to write.	3.41	.723	3.30	.709	2312.500	4797.500	-1.079	.281
3.2	I recombined or reordered the ideas to fit the structure of my essay.	3.16	.782	3.17	.659	2518.000	5003.000	-.163	.871
3.3	I removed some ideas I planned to write.	2.99	.858	3.07	.767	2467.500	5168.500	-.382	.702
3.4	I tried to use the same organizational structure as in the source texts.	2.86	.887	2.87	.833	2553.500	5254.500	-.006	.995
4.1	While I was writing I sometimes paused to organize my ideas.	2.86	.751	3.16	.673	2040.500	4741.500	-2.292	.022
4.2	I developed new ideas	3.25	.703	3.19	.804	2483.500	4968.500	-.313	.754

4.3	I made further connections across the source texts.	2.90	.802	2.93	.840	2483.000	5184.000	-.314	.753
4.4	I re-read the task prompt.	3.22	.804	3.23	.820	2528.000	5229.000	-.118	.906
4.5	I selectively re-read the source texts.	3.05	.797	3.20	.773	2307.000	5008.000	-1.070	.285
4.6	I changed my writing plan (e.g. structure, content etc)	2.79	.781	2.60	.907	2271.500	4756.500	-1.222	.222
4.7	I checked that the content was relevant.	3.34	.692	3.30	.622	2419.000	4904.000	-.611	.541
4.8	I checked that my text was well-organised.	3.23	.773	3.09	.737	2248.500	4733.500	-1.344	.179
4.9	I checked that my text was coherent.	3.32	.762	3.13	.721	2169.000	4654.000	-1.686	.092
4.10	I checked that I included all appropriate main ideas from all the source texts.	3.19	.776	3.06	.849	2342.000	4827.000	-.934	.350
4.11	I checked that I included my own viewpoint on the topic.	3.22	.750	3.30	.622	2471.500	5172.500	-.377	.706
4.12	I checked that the quotations were properly made.	3.18	.788	3.24	.859	2385.500	5086.500	-.742	.458
4.13	I checked that I had put the ideas of the source texts into my own words.	3.26	.800	3.40	.646	2368.000	5069.000	-.832	.405
4.14	I checked the possible effect of my writing on the intended reader.	2.75	.846	3.06	.814	2059.500	4760.500	-2.162	.031
4.15	I checked the accuracy and range of the sentence structures.	2.86	.855	3.06	.740	2263.000	4964.000	-1.290	.197
4.16	I checked the appropriateness and range of vocabulary.	2.99	.842	3.06	.814	2453.500	5154.500	-.439	.661
5.7	After I had finished the first draft, I checked that the content was relevant.	2.86	.976	2.99	1.136	2274.500	4975.500	-1.195	.232
5.8	After I had finished the first draft, I checked that my text was well-organised.	2.81	1.036	2.86	1.067	2468.000	5169.000	-.371	.711
5.9	After I had finished the first draft, I checked that my text was coherent.	2.86	1.045	2.81	1.094	2512.000	4997.000	-.184	.854
5.10	After I had finished the first draft, I checked that I included all appropriate main ideas from all the source texts.	2.74	.958	2.89	1.015	2304.000	5005.000	-1.065	.287
5.11	After I had finished the first draft, I checked that I included my own viewpoint on the topic.	2.73	1.004	2.90	1.079	2263.000	4964.000	-1.244	.214
5.12	After I had finished the first draft, I checked that the quotations were properly made	2.74	.958	2.87	1.006	2327.000	5028.000	-.968	.333
5.13	After I had finished the first draft, I checked that I had put the ideas of the source texts into my own words.	2.86	1.045	3.00	1.116	2306.000	5007.000	-1.060	.289
5.14	After I had finished the first draft, I checked the possible effect of my writing on the intended reader.	2.48	1.015	2.70	1.026	2246.000	4947.000	-1.320	.187
5.15	After I had finished the first draft, I checked the accuracy and range of the sentence structures.	2.74	.928	2.79	.991	2450.500	5151.500	-.458	.647
5.16	After I had finished the first draft, I checked the appropriateness and range of vocabulary.	2.79	.957	2.76	1.042	2530.500	5015.500	-.104	.917

Appendix 5.2 Results of KMO and Bartlett's tests (real-life data)

Data of the conceptualisation phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.748
Bartlett's Test of Sphericity	Approx. Chi-Square	225.559
	df	28
	Sig.	.000
Data of the discourse and meaning construction phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.737
Bartlett's Test of Sphericity	Approx. Chi-Square	347.377
	df	45
	Sig.	.000
Data of the organising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.761
Bartlett's Test of Sphericity	Approx. Chi-Square	447.182
	df	55
	Sig.	.000
Data of the low-level monitoring and revising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.725
Bartlett's Test of Sphericity	Approx. Chi-Square	730.375
	df	28
	Sig.	.000
Data of the high-level monitoring and revising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.802
Bartlett's Test of Sphericity	Approx. Chi-Square	1248.146
	df	66
	Sig.	.000

Appendix 5.3 Rejected factor solutions (Real-life)

Table 1 Meaning and discourse construction phase (real-life): two-factor solution (rejected)

		F1	F2
2.5	I read some relevant part(s) of the texts carefully.	.767	
2.4	I searched quickly for part(s) of the texts which might help complete the task.	.688	
2.7	I took notes on or underlined the important ideas in the source texts.	.679	
4.2	I developed new ideas while I was writing.	.629	
2.9	I linked the important ideas in the source texts to what I know already.	.565	
4.5	I selectively re-read the source texts while writing.	.456	.397
4.3	I made further connections across the source texts while I was writing.	.343	.338
2.1	I read through the whole of each source text slowly and carefully.		.892
2.2	I read the whole of each source text more than once.		.803
1.1	I read the whole task prompt (i.e. instructions) carefully		.518
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.	.324	.803

Table 2 Organisation phase (real-life): three-factor solution (rejected)

		F1	F2	F3
2.10	I worked out how the **main ideas** in each source text relate to each other.	.876		
2.11	I worked out how the **main ideas** across the source texts relate to each other.	.851		
2.8	I prioritised **important ideas** in the source texts in my mind.	.599		.312
3.1	I organised the ideas for my text before starting to write.		.859	
3.2	I recombined or reordered the ideas to fit the **structure** of my text.		.775	
2.3	I used my knowledge of how texts like these are organised to find parts to focus on.	.368	.651	
3.3	I removed some ideas I planned to write.			.788
4.1	While I was writing I sometimes paused to organize my ideas.			.786
3.4	I tried to use the same organizational structure as in the source texts.			.347

Table 3 Low-level monitoring and revising phase (real-life): four-factor solution (rejected)

		F1	F2	F3	F4
5.15	After I had finished the first draft, I checked the accuracy and range of the sentence structures.	.922			
5.16	After I had finished the first draft, I checked the appropriateness and range of vocabulary.	.902			
5.12	After I had finished the first draft, I checked that the quotations were properly made.	.860		.364	
5.13	After I had finished the first draft, I checked that I had put the ideas of the source texts into my own words.	.817			.359
4.15	I checked the accuracy and range of the sentence structures.		.843		
4.16	I checked the appropriateness and range of vocabulary.		.759		
4.12	I checked that the quotations were properly made.			.632	
4.13	I checked that I had put the ideas of the source texts into my own words.				.563

Appendix 5.4 KMO and Bartlett's tests (Test Task A data)

Data of the task representation phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.711
Bartlett's Test of Sphericity	Approx. Chi-Square	249.075
	df	28
	Sig.	.000
Data of the meaning and discourse construction phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.800
Bartlett's Test of Sphericity	Approx. Chi-Square	321.165
	df	36
	Sig.	.000
Data of the organizing phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.771
Bartlett's Test of Sphericity	Approx. Chi-Square	261.735
	df	28
	Sig.	.000
Data of the low-level monitoring and revising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.722
Bartlett's Test of Sphericity	Approx. Chi-Square	722.215
	df	28
	Sig.	.000
Data of the high-level monitoring and revising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.850
Bartlett's Test of Sphericity	Approx. Chi-Square	1306.474
	df	66
	Sig.	.000

Appendix 5.5 KMO and Bartlett's tests (Test Task B data)

Data of the task representation phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.749
Bartlett's Test of Sphericity	Approx. Chi-Square	227.366
	df	28
	Sig.	.000
Data of the meaning and discourse construction phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.747
Bartlett's Test of Sphericity	Approx. Chi-Square	156.755
	df	36
	Sig.	.000
Data of the organising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.795
Bartlett's Test of Sphericity	Approx. Chi-Square	206.881
	df	28
	Sig.	.000
Data of the low-level monitoring and revising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.806
Bartlett's Test of Sphericity	Approx. Chi-Square	811.074
	df	28
	Sig.	.000
Data of the high-level monitoring and revising phase		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.858
Bartlett's Test of Sphericity	Approx. Chi-Square	1251.441
	df	66
	Sig.	.000

Appendix 5.6 Rejected factor solutions (Test Task A)

Table 1 Meaning and discourse construction phase (Test Task A): three-factor solution (rejected)

		F1	F2	F3
2.5	I read some relevant part(s) of the texts carefully.	.922		
2.4	I searched quickly for part(s) of the texts which might help complete the task.	.752		
2.7	I took notes on or underlined the important ideas in the source texts.	.588		
4.3	I made further connections across the source texts while I was writing.		.621	
2.2	I read the whole of each source text more than once.		.553	
2.9	I linked the important ideas in the source texts to what I know already.		.423	
2.1	I read through the whole of each source text slowly and carefully.		.380	
4.2	I developed new ideas while I was writing.			
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.			.786

Table 2: Organisation phase (Test Task A): three-factor solution (rejected)

		1	2	3
2.11	I worked out how the main ideas across the source texts	.776		
2.10	I worked out how the main ideas in each source text relate to each other.	.624		
2.3	I used my knowledge of how texts like these are organised to find parts to focus on.	.614		
3.1	I organised the ideas for my text before starting to write.	.542	.422	
2.8	I prioritised important ideas in the source texts in my mind.	.510		
3.2	I recombined or reordered the ideas to fit the structure of my text.	.506		
4.1	While I was writing I sometimes paused to organize my ideas.		.623	
3.3	I removed some ideas I planned to write.			.645

Table 3 High-level monitoring and revising phase (Test Task A): three-factor solution (rejected)

		F1	F2	F3
5.15	After I had finished the first draft, I checked the accuracy and range of the sentence structures.	.868		
5.16	After I had finished the first draft, I checked the appropriateness and range of vocabulary.	.849		
5.12	After I had finished the first draft, I checked that the quotations were properly made.	.834		
5.13	After I had finished the first draft, I checked that I had put the ideas of the source texts into my own words.	.784		
4.15	I checked the accuracy and range of the sentence structures.		.872	
4.16	I checked the appropriateness and range of vocabulary.		.780	
4.12	I checked that the quotations were properly made.			.726
4.13	I checked that I had put the ideas of the source texts into my own words.			.707

Table 4 High-level monitoring and revising phase (Test Task A): three-factor solution (rejected)

		1	2	3
5.9	After I had finished the first draft, I checked that my text was coherent.	.908		
5.7	After I had finished the first draft, I checked that the content was relevant.	.899		
5.10	After I had finished the first draft, I checked that I included all appropriate main ideas from all the source texts.	.895		
5.8	After I had finished the first draft, I checked that my text was well-organised.	.887		
5.11	After I had finished the first draft, I checked that I included my own viewpoint on the topic.	.878		
5.14	After I had finished the first draft, I checked the possible effect of my writing on the intended reader.	.670		.440
4.10	I checked that I included all appropriate main ideas from all the source texts.		.798	
4.7	I checked that the content was relevant.		.759	
4.11	I checked that I included my own viewpoint on the topic.		.700	
4.8	I checked that my text was well-organised.		.665	
4.9	I checked that my text was coherent.		.650	
4.14	I checked the possible effect of my writing on the intended reader.			.833

Table 5 High-level monitoring and revising phase (Test Task A): four-factor solution (rejected)

		1	2	3	4
5.8	After I had finished the first draft, I checked that my text was well-organised.	.905			
5.7	After I had finished the first draft, I checked that the content was relevant.	.904			
5.9	After I had finished the first draft, I checked that my text was coherent.	.904			
5.10	After I had finished the first draft, I checked that I included all appropriate main ideas from all the source texts.	.882			
5.11	After I had finished the first draft, I checked that I included my own viewpoint on the topic.	.862			
5.14	After I had finished the first draft, I checked the possible effect of my writing on the intended reader.	.707			
4.7	I checked that the content was relevant.		.793		
4.8	I checked that my text was well-organised.		.778		
4.10	I checked that I included all appropriate main ideas from all the source texts.		.734		.331
4.11	I checked that I included my own viewpoint on the topic.		.634		.534
4.9	I checked that my text was coherent.		.629		
4.14	I checked the possible effect of my writing on the intended reader.			.897	

Appendix 5.7 Rejected factor solutions (Test Task B)

Table 1 Conceptualisation phase (Test Task B): two-factor solution (rejected)

		F1	F2
1.5	After reading the prompt, I thought about the purpose of the task.	.802	
1.2	I thought of what I might need to write to make my text relevant and adequate to the task.	.738	
1.3	I thought of how my text would suit the expectations of the intended reader.	.737	
4.6	I changed my writing plan (e.g. structure, content etc) while I was writing.	.440	
1.4	I understood the instructions for this writing task very well.	.392	
2.13	I changed my writing plan while reading the source texts.		
2.13	I changed my writing plan while reading the source texts.		.908
2.6	I went back to read the task prompt again while I was reading the source texts.		.393

Table 2: Meaning and discourse construction phase (Test Task B): four-factor solution (rejected)

		F1	F2	F3	F4
2.5	I read some relevant part(s) of the texts carefully.	.889			
2.4	I searched quickly for part(s) of the texts which might help complete the task.	.866			
2.7	I took notes on or underlined the important ideas in the source texts.	.832			
2.1	I read through the whole of each source text slowly and carefully.		.949		
2.9	I linked the important ideas in the source texts to what I know already.		.675		
2.2	I read the whole of each source text more than once.			.780	
4.3	I made further connections across the source texts while I was writing.			.695	
4.2	I developed new ideas while I was writing.				.949
2.12	I developed new ideas or a better understanding of existing knowledge while I was reading the source texts.		.309		.564

Table 3: Low-level monitoring and revising phase (Test Task B): three-factor solution (rejected)

		F1	F2	F3
5.15	After I had finished the first draft, I checked the accuracy and range of the sentence structures.	.868		
5.16	After I had finished the first draft, I checked the appropriateness and range of vocabulary.	.849		
4.12	I checked that the quotations were properly made.	.834		
4.13	I checked that I had put the ideas of the source texts into my own words.	.784		
4.15	I checked the accuracy and range of the sentence structures.		.872	
4.16	I checked the appropriateness and range of vocabulary.		.780	
4.12	I checked that the quotations were properly made.			.726
4.13	I checked that I had put the ideas of the source texts into my own words.			.707

Appendix 6.1 Marking Scheme of Test Task A

The copyright of GEPT Advanced Writing Marking Scheme belong to the Language Training and Testing Center (LTTC). Official permission to use of the marking scheme has been granted by the LTTC. The marking scheme is included in hard copies of the thesis for examining purposes and do not appear in the electronic version of the thesis.

Appendix 6.2 Marking Scheme of Test Task B

The copyright of UoB Reading-into-Writing Diagnostic Test Marking Scheme belong to Centre for Research in English Language Learning and Assessment (CRELLA). Official permission to use of the marking scheme has been granted by CRELLA. The marking scheme is included in hard copies of the thesis for examining purposes and do not appear in the electronic version of the thesis.

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate: Sathena Hiu Chong Chan

Signature: 

Date: 09 August 2013